Quantitative Biology 2021, 9(1): 8–22
https://doi.org/10.15302/J-QB-021-0245

FEATURE

A personal journey on cracking the genomic codes

Michael Q. Zhang[1,2,*]

[1] Department of Biological Sciences, Center for Systems Biology, the University of Texas at Dallas, Richardson, TX 75080-3021, USA
[2] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and System Biology, Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
* Correspondence: michael.zhang@utdallas.edu, michaelzhang@mail.tsinghua.edu.cn

Received February 23, 2021

In February 2001, the IHGSC (International Human Genome Sequencing Consortium) [1] and Celera Genomics [2] each reported draft sequences providing a first overall view of the human genome. On the occasion of the 20th anniversary celebration, I would like to provide some of my personal account on how the HGP (Human Genome Project) has impacted my academic research interests and to offer some candid advice for the new comers.[1]

FROM A MATHEMATICAL PHYSICIST TO A COMPUTATIONAL BIOLOGIST

In 1977, I entered USTC (University of Science and Technology of China) as Mechanical Engineering student among the first generation of college admission after the "Cultural Revolution" and had a dream to become a "rocket scientist" designing spacecrafts. In 1981, CUSPIA (China-U.S. Physics Examination and Application), organized by Prof. T. D. Lee of Columbia University (U.), gave me the opportunity to study non-equilibrium statistical physics under Prof. Joel Lebowitz at Rutgers University. Even though I was mainly interested in driven diffusive systems and supercomputer simulation of non-equilibrium phase-transitions (where vectorized FORTRAN and bitwise manipulation were essential for computing speed and memory-efficiency) [3–6], I did have the freedom of visiting Prof. Author Jaffe at Harvard to learn SUSY (Supersymmetry) field theory (Quantum Field Theory and Statistical Mechanics are equivalent under Feynman-Kac with Wick-rotation) so that I could show fluctuation-dissipation relation to be the consequence of a non-equilibrium Ward-Takahashi identity resulting from the SUSY [7] (which turned to be also related to Jarzynski's equality [8]). Since I realized that undergraduates were evaluated by knowledge intake but graduates by knowledge production, I had to disciplined myself so that I could publish at least one paper per year. Three unforgettable memories during PhD studies were (1) Visiting Fermilab: after Qualifying Exam, I knocked Prof. Devlin's door asking a summer job with his international high-energy experimental team at the Collider Detector of Fermilab so that I had chance to visit my USTC classmate Chao Tang at Kadanoff lab in U. Chicago. Together we also visited the world First Reactor (Chicago Pile-1), Sears Tower (the world tallest building then) and watched Bizet's Carmen at Lyric Opera of Chicago; (2) Joel's characteristic Stat Mech meetings: held twice a year and most were 3-minute talks (if not able to explain a result in 3 minutes, one would not have understood thoroughly oneself!); (3) Princeton-Rutgers mathematical physics annual holiday retreats: people spent happy hours drinking and eating in a professor's house (Elliott Lieb, Michael Aizenman, *et al.*), students really enjoyed discussing with not only local faculty (Goldstein, Speer, *et al.*) but also talking to foreign guests: Herbert Spohn, David Ruelle, Giovanni Gallavotti, George Parisi, Bernard Derrida, Brézin, Zinn-Justin, Yakov Sinai, *et al*. In 1987, I started postdoctoral

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

research under Prof. Jerry Percus on density or entropy functional for nonuniform fluid models [9–12] and later with Prof. Peter Lax on integrable systems [13] at Courant Institute of NYU. Courant was such a wonderful place which allowed to interacting with many top applied mathematicians in person: Penske, McKean, Varadhan, Nirenberg, Papanicolaou, Deift, Schwarz, Spencer, Mishra, Mallet, *et al*. Most vivid memory of Courant event was when the legendary "country-less and homeless" Hungarian mathematician Paul Erdős ("the Euler of our time" according to Ernst Straus) delivered his lecture in a packed and steamy lecture hall, then offering prizes to young audients who claimed to have solved some "Erdős problems"! Four years of postdoctoral research at Courant is my most prolific period, resulting 13 publications among which 8 was single authored. There I learned what problem to pick is much more important than how to solve it.

As many other physicists, I was also inspired by Schrödinger's book "what is life"? Apparent "Anti-2nd-Law-of-Thermadynamics" certainly makes life system extremely fascinating, where non-equilibrium is the rule rather than the exception. When I saw the two articles by Charles DeLisi "The Human Genome Project" in *American Scientist* and "Computers in molecular biology: current applications and emerging trends" in *Science* in 1988, I got excited reading "As the first 'Big Science' project in biology, mapping and deciphering the complete sequence of human DNA will stimulate research in fields ranging from computer technology to theoretical chemistry" [14]. But not until Jim Watson, president of Cold Spring Harbor Laboratory (CSHL) became the first director of Office for Human Genome Research (OHGR, later changed to NCHGR, then NHGRI) of National Institutes of Health (NIH) in 1989 and Tom Marr was recruited one year later from Los Alamos National Laboratory (LANL, T10 Group) GenBank to build the first Bioinformatics group at CSHL, William Chang (former PhD student of Eugene Lawler from Berkeley CS Department (Dept.)) and I joined the group in 1991 as Computational Genome Research Fellows. So it was the HGP (1990−2005) that transformed my academic career. I was not only inspired by Watson's promise [15], but also attracted by the beauty of Cold Spring Harbor Laboratory's unique and historical environment, where Max Delbrück–another physicist-turned-biologist started the famous Phage Group (with Salvador Luria and Alfred Hershey) in 1941 and Jim was then among the students in his initial Phage course. Indeed, one would never know what outside world might be waiting after graduation, as the Roman says: "Luck is what happens when preparation meets opportunity".

In 1992, I won one of the two first NCHGR/NIH K01 awards (Leonid Kruglyak in Eric Lander Lab got the other) helping me transition into Genome Research. With such independent fund, I was able to take Life Technology Training Center courses on Recombinant DNA Techniques I and II. I also took Rich Gibbs's Advanced Automated/Diagnostic DNA Sequencing Workshop at Baylor College of Medicine in the second year. Most importantly I could "rent" a bench in David Beach's wet-lab to learn some interesting "recipes" from CSHL molecular biology protocol "cookbook": Molecular Cloning: A Laboratory Manual" by Sambrook & Maniatis. I learned the hard way that one was supposed to label everything with one's name or they would be gone very quickly (*e.g.,* my new set of pipettes, clean autoclaved flasks/beakers, even restriction enzymes in the freezer)! Even worst, a new comer tended to be blamed when some bad things happened (*e.g.,* contamination of PCR machine). I remember to compete for limited sequencing lanes with the bench mate Greg Hannon (one of the 20+ postdocs in Beach lab, David was very busy piloting a small airplane between CSH, Long Island and his company in Boston every day and hardly had time to talk to most of them except Greg!) and he would never let me washing the glass plate (expensive and hard to replace) on the ABI Sequencer. I recall I was trying very hard with different dye-nucleotides incorporation using Pharmacia sequencer in order to replace Liang & Pardee's radioactive labels of differential display method with fluorescent labels, after many failures I found out only dye-terminators would work. Despite all these, I could enjoy sheer pleasure when experiments did work out successfully, often after many failures! And I began to better understand biological papers and seminars (*e.g.,* what are "gels", how to "cut a band to sequence") as well as to appreciate bench works. Immerged in such a culture background, CSHL meetings and courses were the best places in the world to learn advances in molecular biology and genomics. I even took a full CSHL intensive hands-on course on "Molecular Genetics, Cell Biology and Cell Cycle of Fission Yeast Course" (Fig. 1, I was on the far left) in the fall of 1993, mapping genes by tetrad dissection and understanding Mendelian genetics and Morgan-Stuyvesant linkage through experimentation. I learned when a bioinformatics student receives initial raw data, first task is not to dive in deep analysis, but to (1) know how the data was produced and what could be the main source of errors; (2) check if the quality and the quantity are sufficient, and feedback such QC diagnostics immediately. Any computational biologist must have true love to biology and sincere appreciation to bench work in order to win the trust of the experimental counter-part. As part of CSHL effort to construct a physical map of the fission yeast genome (Beach-Yanagida-McCombie were competing with Hans Lehrach group at Max Planck in Berlin at the time. Many years later when I lectured at a

summer Bioinformatics Workshop organized by Martin Vingron – former postdoc of Mike Waterman, I was asking Hans how he was managed to run a lab of ~100 people. He said easy, using evolutionary principle: partition people in groups, let them competing and the fittest will survive!) I proposed a theoretical model to show our non-random anchoring approach is far superior. Using density functional theory, the exact solution (discrete extension and generalization of Landers-Waterman model [16]) could explain the simulated and the experimental data well [17,18].

FROM GENE-FINDING TO FUNCTIONAL GENOMICS

What could be more interesting than cracking the code of life and understanding our own DNA instruction book? It was Rich Roberts, one of the Nobel laureates for discovering RNA splicing (the other is Phil Sharp) and who's office was in the next door to mine in Hershey, he suggested me to study exon-intron recognition problem which turns out to be the key for split gene structure prediction. Inspired by works of Steve Mount (Splice Donor/Acceptor motifs), Jim Fickett (CDS Markov Property), Gary Stormo (Protein-DNA binding specificity) and other pioneers, I started with developing methods for fission yeast gene structure prediction [19] to supporting Beach-Yanagida-McCombie fission yeast genome mapping and sequencing project. Later Ting (Tim) Chen, Steven Skiena's talented PhD student from CS Dept. of SUNY at Stoney Brook, helped eventually implementing the algorithm *POMBE* [20] (he then went on to George Church lab for postdoc. 5 years later when Wing Wong invited me to talk at Harvard in 2003, George told me about his space life detection project and his former student Martha Bulyk showed her PBMs). In 1994 Patrick J O'Hara invited me to talk at ZymoGenetics in Seattle, I took the opportunity to visit Steve Henikoff in the next door FHCR and also to Lee Hood's new and almost finished MBT building in Washington University, chatting with Philip Green on some computational problems, who had just moved from Washington University at St Louis and was biking to work every day. Since I started my own lab in 1996 and won my first R01 award, I began to heavily involve in human gene structure prediction (this discrimination method was exon-centric and an extension of Solovyev's LDA approach to non-linear QDA, thanks to Bruce Stillman who suggested to call it "MZEF" for "Michael Zhang Exon Finder"!) [21] and systematic classification of all human exons into 16 mutually exclusive classes [22]. Exon-intron discrimination studies also helped me to establish a close and long-term collaboration with Adrian Krainer's lab on RNA splicing regulation, especially using SELEX to refine various RBP binding motifs [23]. In October 1997, Stephen Burley invited me to talk on gene-finding at Rockefeller University (this is a very unique place only having biology and physics, much later Caltech also combined biology with engineering), where I had extensive discussions with Stephen on his co-crystal structure TBP/TATA-box and with Nat Heintz (cloned first cell cycle gene histone) and Bob Roeder (cloned PolII, purified all the TAFs and reconstituted in vitro transcription system) in addition to visited old physics friends Eric Siggia, E.G.D. Cohen and Mitchell Feigenbaum, it was a pity (no kidding) that few biologists would know about Feigenbaum number, Gallavotti-Cohen Fluctuation Theorem or Siggia's intermittency in

Figure 1. A snapshot of taking a full CSHL intensive hands-on course in 1993. Reprinted from Annual Report 1993 in Cold Spring Harbor Laboratory (CSHL).

turbulence at Kolmogorov's singularity unfortunately!

Finding genes was only the first goal of HGP, immediate next was to quantifying mRNA transcription (and gene expression in general) from each gene and microarray transcriptome ignited the functional genomics. In 1997, I joined a team of Stanford and CSHL four labs, led by David Botstein together with Bruce Futcher and Pat Brown, carried out a comprehensive project to identify ~800 cell cycle regulated genes [24], it was a lot of fun working with young enthusiastic postdocs (Paul Spellman, Gavin Sherlock, Vishy Iyer, Mike Eisen, *et al.*) and our dataset has since been used in many Systems Biology classes till this day! I was asked to give a talk in Berkeley in summer of 1999 by Gerry Rubin who was the director of Berkeley Drosophila Genome Project and told me he proved my Merck Research Institute grant (1997−1999). I took the opportunity to visit Bob Tjian, discussing transcriptional regulation while enjoying his beautiful gold fishes in a big tank in his office. In the fall, Sam Karlin invited me to visit him at Math Dept. of Stanford for 3 weeks. I had a great respect to Sam, not only he was a famous mathematician, but had also made important contributions to genomics: *e.g.,* Karlin-Alschul *BLAST* statistics [25], Burge-Karlin *GENSCAN* [26]). He explained GI/G/1 queuing problem to me, which led to the exact solution to *BLAST* statistics historically. After giving talks in Stanford, I visited Ron Davis (former postdoc of J. Watson) in Biochemistry and his Genome Technology Center, Ron jointly developed DNA microarray and published the first yeast expression profiling with Pat Brown [27,28]. I learned the history of microarray development and how he helped founding Affymetrix and many other biotech companies in the Silicon Valley, *etc*. I never forget Ron's heroic story on Sep. 28, 2001: Due to the terrorist attack on September 11, several international and west-coast people were not able to attend the 2nd Workshop on Computational Biology that I organized with Aravinda Chakravarti (Fig. 2), but Ron showed up giving his keynote and then had to drove all the way back New York to California! (Recently I invited Ron to talk at the University of Texas at Dallas (UTD) about his biggest challenge to cure his own son! I discussed my experience working with people came from his lab: Rick Young, Mike Snyder, Joe Eckert, *et al.*). Taking advantage of being in Palo Alto then, I also visited David Cox (to learn Radiation Two-Hybrid Mapping method) who co-directed the UCSF-Stanford Genome Center with Rick Myers (former postdoc of Maniatis) in addition to the two other Genome Centers: Davis's in Biochemistry and Brown-Botstein Center in Genetics and Pediatrics. Since bioinformatics is driven by technologies, being aware of new data generation and analysis technologies is absolutely essential.

In May, 1998, Chao Tang invited me to talk at his

Figure 2. Michael Zhang and Aravinda Chakravarti. Reprinted from Annual Report 2001 in Cold Spring Harbor Laboratory (CSHL).

Princeton Summer Lecture Series on Biophysics at NEC Research, Arnie Levine was not able to attend but sent Leibler's postdoc Uri Alon (former physics PhD from Weizmann Institute physics PhD) to discuss their cancer microarray data analysis. Since I visited Chao Tang (he was doing postdoc with Per Bak and Kurt Wiesenfeld) at the Brookhaven Lab in 1990 and I introduced sequence alignment phase-transition problem to Terry Hwa, I had been tried to convert Chao, Terry and Hao Li, Michael Lässig, *et al.* to biologists, in addition to train many junior physics students/postdocs, and eventually succeeded. I remember I wrote a letter to Chris Sander, a bioinformatics pioneer originated from physics, trying to convince him moving from EMBL to CSHL, but got a short reply: "Thanks Mike! I'm well taken care of..." Soon after I saw the news "Bioinformatics Pioneer Chris Sander Accepts Position with Millennium" and the rest is history. In early 1999, I became as the director of Bioinformatics course for the first class of 6 PhD students: Elizabeth Thomas, Amy Caudy, *et al.* of the Watson School of Biological Sciences, joined with the two new faculty recruits: Andy Neuwald and Lincoln Stein, meanwhile served as a consultant to Tularik Genomics (later acquired by Amgen). In the summer, Edward Domany's student Gaddy Getz from Physics Dept. of Weizmann Institute visited my lab, he developed an alternative hierarchical adaptive clustering algorithm *SPC* (super-paramagnetic clustering) and reanalyzed our yeast

cell cycle data [29]. Xinyuan Fu (former Jim Darnell's postdoc at Rockefeller) came to visit me discussing his ambitious plan to establishing Tsinghua Institute of Genome Research and invited me to give a talk at Yale (which was finally realized in 2002 when I had an opportunity to discuss science with Tian Xu, Xingwang Deng, Richard Lifton, Junhyong Kim, Mark Gerstein, Mike Snyder, Sherman Weissman). In the fall of 1999, I organized the 1st CSHL Workshop (Fig. 3) on Computational Biology: Bridging the Gap between Sequence and Function with Eugene Koonin and Edward Uberbacher, Lee Hood (he co-founded the world first Institute of Systems Biology with Alan Aderem and Ruedi Aebersold in 2000, when I was on the site-visit at Alan's Seattle Biomed with NIAID/NIH SAB team in 2015, I got the chance to visit ISB personally and chatted with Lee and Sui Huang.) and Terry Sejnowski (former John Hopfield's student, I invited a computational neural biologist because I believe machine-learning and learning-machine may eventually converge. When John decided to move back to the East coast, we had an interesting chat about Hopfield-network and Hopfield-proofreading in my office in Hershey, CSHL) gave the keynotes. During the workshop, I was totally stunned when I saw Wally Gilbert (Nobel laureate for inventing chemical DNA sequencer), 67 year old Harvard professor and another physicist-turned-biologist, was standing outside the Grace Auditorium by his poster stand, patiently explaining his intron-early theory to a foreign student; especially in contrast to his former student Laura Landweber, who started her assistant professor in Princeton in 1994 at the age of 26 and was an invited speaker on evolution of the genetic codes. Wally told me how he abandoned all the experimental instruments and turned his wet-lab into a dry-lab with just computers, his courage certainly compels universal admiration! It reminded me of his prediction in 1991: "The new paradigm now emerging, is that all the genes will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical" [30].

Figure 3. Michael Zhang and B. Stillman. Reprinted from Annual Report 1999 in Cold Spring Harbor Laboratory (CSHL).

When I was invited to talk at EBI/EMBL on Genome Based Gene Structure Determination in Cambridge, UK in summer of 2000, I visited Rich Durbin and Tim Hubbard at Sanger Institute, and had long chat with Ewan Birney at EBI. Ewan was an Eton intern student with Jim Watson and Adrian Krainer (former Maniatis student) at CSHL before he continued his PhD with Durbin at Sanger. I had failed to recruit him as postdoc at CSHL, he preferred to stay at St John's College and later joined EBI. At the time of my visit, he was under a tremendous pressure to save the EBI by his Ensembl as EMBL was considering to consolidate EBI into Heidelberg! But on the bright side, he and David Stewart started in May, 2001 a dramatic GenesWeep – worldwide competition on human gene number prediction with the initial estimate at ~60,000 and the final winner at ~21,000 when complete human genome was announced at the CSHL Genome meeting in the summer of 2003 [31]. Right before the Genome meeting, Stefan Wiemann invited me to talk on Genomics and Cancer conference at KDFZ, during the trip I visited Peer Bork at EMBL and chatted with Ron Shamir in a small restaurant on top of the hill overlooking the Rhine in Heidelberg (unfortunately it would had been too far south from St. Goar and the Lorelei statue!). Six months later, Roderic Guigó (former postdoc of Temple Smith and Jim Fickett) invited Ewan Birney, Michael Brent, Termitaries E, Lior Pachter, Hugues Roest Crollius, Victor Solovyev and me in Barcelona (sponsored by Fundacio La Caixa) to discuss the major challenges: hypothesis driven experiments and biologically realistic mathematical models that need to be overcome in order to further complete the human gene catalog [32].

FROM PROMOTER AND *CIS*-REGULATORY FINDING TO REGULATORY GENOMICS

Early on I realized, to understand how gene expressions are regulated, identifying *cis*-regulatory elements in promoters and enhancers would be necessary. To facilitate the yeast cell cycle study, my student Jian Zhu built the first yeast promoter database *SCPD* [33] (later ported into Mike Cherry's YPD). By assuming co-regulated gene promoters may share common transcription factor binding sites (TFBSs), using motif search, I identified many known and putative TFBSs from co-expression clusters of the cell cycle data [34]: bi-directional histone promoter UAS1/UAS2, Swi5/Ace2 in M/G1 cluster, MCB and SCB in G1/S cluster. Remarkably, in G2/M cluster, a motif pairs (Mcm1-X) regulating Swi5 and Clb2, and the putative co-factor "X" matched to an unknown factor SFF [35] (this was a summary of an invited talk at Kyoto

conference "Holistic Views of biology" in April 1998, sponsored by Otsuka America Inc.) but the gene encoding SFF has not been identified then. Later by detailed functional study it was identified as Fkh1/Fkh2 [36]! Since the yeast has a compact genome (12 Mb, completed in 1996) with limited introns, noncoding intergenic regions were sufficient for promoter and *cis*-regulatory element discovery. Not only I proved the possibility of predicting novel TFBSs with gene-expression data, but also denmostrated the importance of controlling for GC-content and repeats [35]: *e.g.,* it is well known that homopolymeric dA:dT sequences are extremely abundant in most of the yeast promoters. They affect nucleosome formation *in vitro* and are required for wild-type levels of transcription *in vivo*. This ubiquitous promoter element stimulates transcription via its intrinsic DNA structure (Vishy Iyer's PhD thesis work under Kevin Struhl [37]). But they can create a lot of problems *in silico* during an alignment [35]. In August of 1998, Nikolay Kolchanov and Victor Solovyev invited me to talk at the first international conference on bioinformatics of Genome Regulation and Structure (BGRS) at Russian Academy of Sciences in Novosibirsk, giving me opportunity to interact with Russian scientists. During the meeting excursion, Mikhail Gelfand and I were sharing a cabin on the Altai Mountain when I was telling him his grandfather Israel Gelfand's representation theory was one of my favorites and my PhD advisor Joel Lebowitz was working very hard to get his grandfather and Yakov Sinai (both were Andrey Kolmogorov's students) together with other repressed Soviet Jewish scientists immigrating to the States.

Transcriptome could only give indirect regulatory information, but direct mapping TFBSs *in vivo* can provide direct physical (causal) targets. In a first ChIP-chip experiment, Blat and Kleckner demonstrated its revolutionary power by mapping cohensin binding sites along budding yeast chromosome III [38] in 1999. By printing all yeast intergenic sequence fragments on a single chip, genomewide CHIP-chip were accomplished simultaneously by Bing Ren of Rick Young lab on Gal4/Ste12 and by Vishy Iyer of Pat Brown lab on MBF/SBF (together with David Botstein and Mike Snyder labs) around the fall of 2000 [39,40]. Combining both expression-array and genomic array data, serial regulation of 9 cell cycle TFs in a Gene Regulation Network (GRN) was reported [41]. We (led by a postdoc Mamoru Kato, former student of Tatsuhiko Tsunoda and Toshihisa Takagi, the latter was the director of Human Genome Center. Takagi invited me to lecture at U. of Tokyo when I visited Satoru Miyano and Sumio Sugano labs in April 1998) further showed this TF chain can be extended to a chain of composite regulatory modules: different modules may share a common TF component in the same pathway or a TF component cross-talking to other pathways [42]. In celebrating new millennium, I told Russ Altman that I would organize the Pacific Symposium on Biocomputing (PSB) in January 2000 on "Identification of Coordinated Gene Expression and Regulatory Sequences" with Jean-Michel Claverie, Minoru Kanehisa, Dan Prestridge and Gary Stormo in Honolulu, Hawaii. On my way back from PSB'2000, I stopped over to talk at USC and meet with Mike Waterman, Simon Tavaré, Pavel Pevzner, *et al.* I also visited Eric Davison lab at Caltech, discussing his fascinating gene circuit for spatial control of transcription in sea urchin development and invited him to give a lecture at CSHL next year. Shortly after, I went to visit Warren Ewens when he invited me to give a talk at U. Penn. Their classical works on sequence comparison, evo-devo and populational genetics guided many generations of computational biologists and bioinformaticians.

In one of the CSHL meeting, I met Rich Ebright (Director at the Waksman Institute of Microbiology of Rutgers), he taught me bacteria promoter architecture and together with Danny Reinberg (former Roeder's postdoc), we discussed my finding about human core promoter architecture. Yi Zhang was doing postdoc in Danny's lab at that time. I found many outstanding Chinese scholars are hard-core biochemists: Xiaodong Wang, Xiangdong Wang, Yang Shi, Yi Zhang, Zhijian (James) Chen, *et al.* because they are hard workers. As UTSW people said, any slight activity could be purified by Xiaodong, because he never gave it up unlike other American colleagues! It turned out getting most human promoter sequences was not possible without the human draft genome. Even if long upstream genomic sequence is available, unlike in yeast where 500 bp upstream of AUG were generally regarded as "the promoter region", because of the long 5′ UTR, potential long first intron which could split 5′UTR or alternative transcriptional start sites (TSS), knowing AUG is insufficient to determine TSS. I decided to systematically study human promoter structures using machine learning by first characterizing nucleosome periodic signature [43] and CpG islands [44], both led by a postdoc Ilya Ioshikhes (I met his PhD advisor Edward Trifonov of Weizmann Inst. on Altai Mountain trip when he demonstrated his mushroom picking skills, which he learned during his family were exiled to Siberia by Starlin regime); then by CART classification of human 5′UTRs using Sugano's 5′RACE data [45] and finally by integrating both promoter and first intron donor site signals into *FirstEF* (First-Exon-Finder) to predict both promoters and first exons simultaneously in the whole draft genome as the sole input [46], both led by another excellent postdoc Ramana Davuluri. Such effort made it possible to predict putative E2F followed by ChIP validation in collaboration with Alexander Kel *et al.* [47] or to ChIP-clone/PCR E2F targets followed by

EMSA validation in collaboration with Peggy Farnham lab [48] (our predicted human promoters helped designing Nimbelgen oligo-array which had better performance than Affymetrix tiling arrays for ChIP-chip experiments, [private conversation with Roland Green and Peggy Farnham]). Shortly after, Bing Ren of Rick Young lab in collaboration with Brian Dynlacht lab did E2F CHIP-chip with ~1000 putative promoter sequences of known cell cycle genes [49]. Availability of the complete genome allowed the Affymetrix genomic tiling chips [50] or Nimbelgen oligo-promoter-array be used for genomewide ChIP-chip studies [51]. In March 2003, Philip Benfey, Stephen Small, Susan Celniker and I organized the first CSHL Systems Biology meeting: Genomic Approaches to Transcriptional Regulation [52], Stuart Kim gave the keynote on Global Analysis of Conserved Genetic Pathways and Molecular Machines. Mike Eisen, Gary Stormo, Eric Siggia and I chaired the sessions on Computational Approaches to Identifying CREs. Stanislas Leibler's synthetic biology approach to gene regulation network was in stark contrast to Rick Young/David Gifford *in vivo* ChIP-chip approach, so was Stormo's Protein-DNA interaction approach *vs*. Siggia's evolutionary one. I also arranged a job interview for Bing Ren and hoped to recruit him to CSHL, but failed (many CSHL colleagues were too conservative, not regarding Genomics as hard science!). Fortunately Ludwig Institute turned out to be much better for him. Anyway Rick invited me to talk at the Keystone Symposium: Biological Discovery Using Diverse High-throughput Data in the next Spring, together with David, Stuart and Sue Lindquist, we continued our discussion on future goals in GRN analysis at the dinner after skiing (Manolis Kellis, just finished his PhD thesis under Eric Lander, was very shy young student then and came to introduce himself at the base of the slope on the second day), of course both Rich and David rented separate airplane as usual to pilot to a small airport nearby.

Anticipated massive expression and ChIP data coming, I began to develop machine learning algorithms for large-scale systematic identification of tissue-specific *cis*-regulatory modules. I was fortunate to have three talented postdocs: Debo Das – a physicist, Andrew Smith – a computer scientist, and Pavel Sumazin – an applied mathematician joining my lab. Debo Das extended Bussemaker's *REDUCE* (linear regression) [53] to *MARS-Motif* (multivariant adaptive regression spline) [54–56]. Pavel developed *DWE* (Discriminate Word Enumeration) [57], Andrew extended it to *DME* (Discriminate Matrix Enumeration) [58]. *DWE*/*DME* allow to enumerating and ranking enriched all possible TFBS motifs in the target sequences (ChIP+ sequences or DEG promoters) *vs.* background controls (ChIP– sequences or non-regulated gene promoters), while *MARS-Motif* could subsequently regress all candidate motifs against target gene expression levels in order to extracting active and functional ones [59]. Using 56 different tissue transcriptomes and the draft genome, we were able to show DNA motifs in human and mouse proximal promoters predict tissue-specific expression [60] and to build TCat (The Catalog of Tissue-Specific Regulatory Motifs) [61]. Bioinformatics is a rapidly moving field, one needs to be able to aim and to shoot at the moving targets constantly.

FROM TRANSCRIPTIONAL TO POST-TRANSCRIPTIONAL REGULATION

Alternative exon usage has been a major mode of tissue-specific or conditional specific gene regulation. Following Steve Mount, Stefan Stamm and I collected and characterized the earliest neuron alternative exon database since 1994 [62,63] (later Bob Darnell, Roeder's former student, told me our database was instrumental for his study on Nova splicing. Doug Black also told me similar stories when I gave a talk at UCLA.). At EBI 2000 meeting in Cambridge, I demonstrated [64] how MZEF played important role in annotation of APP (a famous Alzheimer's disease gene encoding amyloid beta (A4) precursor protein,) which is about ~300 kb with ~20 exons residing in low-GC 21q21, which made gene-finding much more difficult (MZEF was used by Chr21 sequencing consortium led by the Japanese HGP). Another example is the striking multiple alternative first exon organization of neural cadherin gene clusters (*Pcdh*) in 5q31, Qiang Wu (postdoc of Maniatis Lab and former student of Adrian Krainer) were not fooled by automatic "repeat-marking" but manually and painstakingly aligning cDNAs in order to keep repetitive first exons (lesson should be learned by all bioinformaticians) [65]! When the mouse sequences became available from Rick Myers lab, my postdoc Theresa Zhang worked together with Qiang to demonstrate conservation of the rich array of CpG-islands as well as TFBS motifs [66]. Further experiments showed that alternative promoter choice determines first exon donor site selection. Recently, using CRISPR inversion, it was also shown that these clusters of enhancers/promoters are regulated topologically by CTCF in an orientation dependent manner [67].

In collaboration with David Spector lab, the discovery of CTN-RNA, a member of the nuclear regulatory RNA (nrRNA) family was also a surprise [68] emphasizing nuclear retention as another important mechanism for gene regulation. The 3′UTR of CTN-RNA contains elements for A-to-I editing, involved in its nuclear retention. Under stress, CTN-RNA is posttranscription-

ally cleaved to produce protein-coding mCAT2 mRNA. Another example is Malat1/Neat2, which is enriched in nuclear speckles only when RNA polymerase II-dependent transcription is active and it regulates synaptogenesis by modulating gene expression [69]. In collaboration with Josh Dubnau lab, through learning & memory training/testing, we identified synaptic targets of *Drosophila* RBP Pumilio [70], as local mRNA translation within dendrites of neurons is a mechanism for activity-dependent synaptic plasticity. In collaboration with Xiaowo Wang and Yanda Li labs, we developed *MIROR* which could identify cancer targets by cell-type specific microRNA occupancy rate changes instead of differential expression level changes [71]. We could also use synthetic circuits to quantitatively model microRNA-mediated regulation on competing endogenous RNAs with help from Zhen Xie lab [72]. In collaboration with Greg Hannon and Josh Huang Labs, we profiled cell-type specific miRNAs in mouse brain [73].

Understanding RNA splicing codes has been my long-term goal since 1993 when I extended Weight Matrix Method to Weight Array Method for scoring splicing signals [74] and my postdoc Jinhua Wang (former student of Zhijian Yao, together with Boqin Qiang and Yan Shen, they were running the Northern Genome Center of China in Beijing around 2000) developed *ESEfinder* for predicting exonic splicing enhancers [75] in collaboration with Krainer Lab. In 2004, a very talented student–Chaolin Zhang joined my lab at CSHL from SUNY Stony Brook genetics program and strengthened our close interaction with Krainer lab. He moved to Bob Darnell lab at Rockefeller in 2008 (I remember during one of Bob's PhD student defense, David Allis and I were the examination committee members. When David told me both of his two sons were fighting in Iraq War at that moment, I felt extremely sympathetic to this great scientist!) and then started his own lab at Columbia University in 2012. He discovered "zero-size exon" or "dual-specificity splice site" [76], investigated evolutionary impact of limited splicing fidelity in mammalian genes [77], constructed the regulatory network of the tissue-specific splicing factors Fox-1,2 [78] (which he further refined by HITS-CLIP and applied to autism in Darnell lab [79]). During his 2008 summer internship in Jason Johnson lab at Rosetta, he completed a computational analysis of differential expression of 24,426 human alternative splicing events and predicted *cis*-regulation in 48 tissues using custom built exon-chip [80]. When Chris Burge invited me to talk at Massachusetts Institute of Technology (MIT) in that year, I felt bad because Chaolin turned his offer down when choosing Bob's lab. But Chris told the good news that since I insisted Phil Sharp should recruit him into a Center member to beefing up Center's bioinformatics strength, Phil did. Phil was the Director of MIT the Cancer Center when I was at the onsite visit with the NCI 5 year evaluation team in 2004. (In 2011, I invited Phil to give a lecture at the Tsinghua University Centenary Celebration). When I went back for the 2009 onsite review, Tyler Jacks, who succeeded Phil as the new director, came to me to confirm Chris' important role in the Center. I remember on that day, as Susan Hockfield, president of MIT showed everyone around the new gassy Koch Institute for Integrative Cancer Research and I plauded her new vision of integrating cancer biology with quantitative science by, on each floor, physically housing biology labs on one-side and math/phys/eng labs on the other, so that they would have to bump into each other in the middle coffee/lunch area! I joked "it is amazing to see three woman Ivy League presidents!" (herself, Drew Gilpin Faust of Harvard and Shirley M. Tilghman of Princeton). She laughed winking "and two are biologists"!

Based on a Bayesian model to estimate inclusion ratios, a new alternative exon detection algorithm *SpliceTrap* was developed mainly by Jie Wu a PhD student [81], he also worked with Chaolin Zhang to develop another algorithm *OLego* for fast and sensitive mapping of spliced mRNA-Seq reads using small seeds, specifically designed for *de novo* mapping of spliced mRNA-Seq reads. *OLego* adopts a multiple-seed-and-extend scheme, and does not rely on a separate external aligner. It identified hundreds of novel micro-exons (< 30 nt) in the mouse transcriptome, many of which are phylogenetically conserved and can be validated experimentally *in vivo* by Olga Anczuków (postdoc of Adrian R Krainer) [82]. A protein-protein interaction (SR and hnRNP proteins to the core spliceosome) network was built and validated (work led by Martin Akerman, a joint postdoc) [83].

FROM EPIGENETICS TO 4D GENOMICS (2006–2021)

Early on in HGP, it was clear cell-specific regulatory codes would include epigenetic ones. We started in 2005 on human brain DNA methylation profiling in collaboration with Tim Bester lab (and Jingyu Ju's sequencing core) at Columbia U. [84]. and trained machine learning model for genomewide mCpG prediction [85]. We then carried out high definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing of cancer cells [86,87]. After joining the epigenome mapping consortium for NIH Roadmap project, we reported comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications [88,89]. When Jin Gu of Tsinghua faculty visited my new lab at UTD, we developed *FastDMA* – an Infinium Human Methylation 450K Beadchip analyzer [90] and using such chip, Xiaotu

Ma (postdoc) discover SCT promoter DNA methylation as a novel lung cancer biomarker and further validated experimentally in collaboration with John Minner and Adi Gazdar labs of UTSW [91,92]. In collaboration with Matteo Pellegrini lab of UCLA, Weilong Guo (a Tsinghua student) developed a new BS-Seeker2 pipeline for bisulfite sequencing or RRBS data analysis [93] and applied to study nonCpG methylation patterns in human cells [94].

Histone modifications is another type of epigenetic codes that are generally less stable than DNA methylations. In 2006, my student Dustin Schones graduate from Dept. of Physics of Stoney Brook U. and went to Keji Zhao lab at NHLBI(The National Heart, Lung, and Blood Institute)/NIH as a postdoc, helping to set up bioinformatics support for high throughput ChIP-seq experiments, we began to collaborate on systematically profiling (~40) histone modifications in T cells, a NYU PhD student Jeff Rosenfeld did bioinformatics analysis in my lab (I asked both Stephen Small and Andrea Califano to serve in Jeff's thesis exam committee.) [95]. As the epigenome mapping center (Bing Ren, James Thomson, Joe Eckert, Wei Wang and my labs) project, Wei Xie (postdoc in Ren lab) led a comprehensive epigenomic analysis of multi-lineage differentiation of human embryonic stem cells [96], a postdoc Pradipta Ray, former student of Erix Xing, coordinated bioinformatics work in my lab after we moved to UT Dallas in 2010 (Jim Watson autographed his book as a gift for me before I said goodbye and left his house). We also did integrative analysis of haplotype-resolved epigenomes across human tissues [97] and completed the entire consortium report "Integrative analysis of 111 reference human epigenomes" at the end of the NIH Roadmap project in 2015 [98]. Using epigenetic marks, we improved the accuracy of our boosting algorithm for transcriptional start site prediction [99,100].

Because ChIP-seq for epigenetic marks could have both sharp-peak (people often use *MACS* from Shirley Liu lab to analyze) and broad-peaks (people often use *SICER* from Keji Zhao lab), when Haipeng Xing of Math Dept. of Stony Brook U. told me about his Bayesian Change Point (*BCP*) algorithm for predicting stock market "jump" and "crash", I realized it would be the exact model genomics field needs! Since my friend Peter Gergen made me as an adjunct professor in Math/Phycics/Genetics of Stony Brook, I asked two of the math students: Yifan Mo and Will Liao(jointly supervised with Haipeng) to develop a novel and fast ChIP-seq analysis tool based on the exact *BCP* model [101,102] in 2011 (see independent benchmark of different tools [103]), *BCP* also performed well in revealing FoxO function in Treg cells [104] in collaboration with Ming Li Memorial Sloan-Kettering Cancer Center and others. Ming and I were part of the Starr Cancer Consortium (including Broad, CSHL, MSKCC, Rockefeller and Weill Cornell Medicine) during 2009−2012. FoxO family has always been one of my favorites, which is important for cancer and aging. Zhenyu Xuan (postdoc and later CSHL Bioinformatics Cancer Core manager) carried out an extensive comparative analysis of FoxO pathways which is conserved from worm to human [105], I reported at Functional Genomics of Ageing 2004 at Crete Island of Greece, where I have learned the genetic advances from Cynthia Kenyon, Gary Ruvkun, Leo Guarente and his student David Sinclair. When Anne Brunet was about to depart from HMS (Michael Greenberg lab) to start her own lab at Stanford, she contacted me about collaboration on studying relationship between pro-longevity FoxO3 and its relation to tumor suppressor p53 (both are deacetylated by the Sirt1), it took almost 7 years before the paper was finally published [106] proving FoxO3 is a direct target of p53.

Due to the advent of 3C based technology (originally developed by Job Dekker of Nancy Kleckner lab in 2002) and ChIA-PET from Yijun Ruan lab (2009), importance of higher-level chromatin topological regulatory codes was unveiled and NIH launched 4D Nucleosome project in 2015. Mapping chromatin loops not only can reveal higher-order chromosomal DNA architecture but also can link active and distal enhancers to their target genes (particularly important to annotate unknown GWAS SNPs in the vast non-coding regions). During the epigenome mapping project, we did ChIP-seq analysis for the two major DNA binding proteins: PolII (transcribing genes) [107] and CTCF (partitioning gene activity domains) [108]. Since ChIA-PET offers higher resolution and specific TF loci, we initially focused on improving data analysis model (MICC [109] led by student Chao He), pipeline (*ChIA-PET2* [110] led by student Quipeng Li), detecting co-factor complexes and differential long-range interaction by *stochastic process* techniques (*3CPET* [111] and *FIND* [112], both led by student Mohamed Nadhir Djekidel), and finely mapping chromatin domain boundaries (*HiCDB* [113] let by student Fengling Chen). In fact, long-range chromatin contact maps provided valuable training data for developing predictive models based on multiple 1D ChIP-seq datasets alone. First such prediction model for ER chromatin loops was developed and validated with 3C data and real large-scale regulatory interaction study in 2014 [113–115]. New pipeline *ChIA-PET3* was also developed by Yong Chen (postdoc and mathematician) and applied to profile chromatin landscapes from multiple prostate cancer cell lines in collaboration mainly with Ram Mani lab in pathology Dept. of UTSW [116]. He also developed a new algorithm [117] for analyzing CAPTURE-3C-seq (dual to ChIA-PET, in the sense, one starts with a specific binding-site

instead of a binding factor but allowing to extract both chromatin loops as well as localized protein complexes simultaneously [118,119]), which is a novel protocol invented by Jian Xu of UTSW and successfully applied to beta-globin Locus Control Regions (LCRs) and many other super-enhancers which could nucleate multi-loop condensations. Another complementary protocol, invented by Xiangdong Fu, to capture Chromatin-RNA Binding Protein (RBP) interactions has also been developed recently after Zhengyu Liang (a Tsinghua student who developed BL-Hi-C in Tsinghua in 2017 [120] and went to Fu lab as a postdoc in UCSD) [121]. In order to validating functional TF-binding and long-range chromatin interaction at single cell level, imaging method is also essential. In collaboration with various imaging experts, we have also developed new super-resolution methods [122,123], Tn5-FISH [124] (led by a PhD student Jing Niu) as well as GUI visualization tool Web3DMol [125] (led by a Master student Maoxiang Shi).

To access functional enhancers, detection of eRNAs is a complementary approach to mapping enhancer-promoter chromatin loops. Peng Xie, a talent student, initiated a wonderful collaboration with Fei (Xavier) Chen (Ali Shilatifard's student of Northwestern University) after they met at a CSHL meeting. By profiling eRNAs, they discovered a functional elongation enhancer which regulates promoter-proximal pause-release when bound by PAF1 [126]. He then worked with experimenters in Tae-Hoon Kim (Tae moved from Ren lab to Yale becoming a faculty and now is the head of our department at UTD) lab, we studied dynamics of eRNAs in a classical inducible system. One unexpected finding from our study is that the fate of an enhancer can be different from the promoter it regulates after their transient functional and physical association. In many independent cases, eRNA production continues while the enhancer becomes disengaged from its associated promoter and the target gene undergoes post-induction repression [127].

FROM CELLS/TISSUES TO ORGANS/ BODIES

Single cell omics allow people dissect tissue/spatial/temporal hyperacuity and seek functional meaning of biological variations. Led by Xiangyu Li in 2016, a joint student with Xuegong Zhang and Jin Gu, we used network embedding based representation learning (SCRL) which could efficiently implement scRNA-seq data-driven non-linear projection and incorporate prior biological knowledge (such as pathway information) to learn more meaningful low-dimensional representations in either cell- or gene-space [128]. Led by Zehua Liu, a Tsinghua student of Tim Chen and co-supervised by Jiang Rui (former postdoc of Tim at USC) and me, developed algorithm for reconstruction of cell cycle pseudo-time from scRNA-seq [129]. Led by the UTD student Peng Xie, *SuperCT*: a supervised-learning framework for enhanced characterization of tumor heterogeneity using multiple scRNA-seq data was developed [130]. Led by Tuanlin Xiong, Tsinghua postdoc of Qiangfeng (Cliff) Zhang lab, a scATAC-seq analysis via latent feature extraction (*SCALE*) method was developed, which combines a deep generative framework and a probabilistic Gaussian Mixture Model to learn latent features [131].

… … …

This is an unfinished section which I hope to come back in the future. It is a good time to pause and reflect: For structure genomics, one should ask (a) what determines the number of genes (number of the "independent generators" for population dynamics)? (b) what determine the order and inter-distance between gene loci (eigenvalue spacing for stochastic evolutionary dynamics like the "Dyson-Wigner Brownian particles")? For functional genomics and GRN, one should ask (a) what is the invariance of the cellular dynamics (could pathways be some "low-energy level" collective excitation modes)? (b) can one derive the "Green function" for given typical "boundary conditions" like the heat kernel, which could simplify or decompose the holly grail causality: Genotype → Phenotype)? For epigenomics and 4DG: one should ask (a) what is the structural property of this adaptor/mediator in microenvironment ← epigenome → genome (could genome be the natural "boundary")? (b) could one follow the lineage tree down from the "stem cell" (germ cell) and identify, at each branching point, what is the key topological changes *e.g.*, in terms of Betti number (like building the "Mendeleev's Periodic Table" describing how electrons are changed in the outer shell of an atom)? After all, Systems Biology is about connections from one level to the next in the hierarchy, how to connect molecular networks to cellular networks, and in turn to organ networks *etc.*, is the fundamental question which may hopefully be resolved with upcoming massive spatial single cell omics data. Then it is dynamics that must be tackled in order to solve causal mechanisms. At the theoretical level, what should be appropriate mathematical language to formulate biological problems more systematical (like the Category theory)? It (network-of-network) seems very much like Quantum Mechanics, semigroup of random matrix or matroid might be a part of the answer. I saw a new commentary "Truth and Beauty in physics and biology" made by Ben MacArthur [132] which is a bit superficial. When I talked to my physics friends how to think biology principle (a) favor "specificity" over "generality" (*e.g.*, think about how a single mutation could turn a strain from commensal to pathological, like COVID virus, "order-of-magnitude"

back-of-envelop-calculation" simply would not cut; (b) favor "innovation" over "conservation" at all cost even if risking life (extreme form of "use it or lose it"); (c) "anticipation" or "planning", not just "adaptation". It is still not clear if Deep learning (like AlphaFold) could really help to reverse engineering the "Equation of Motion" in Biology? In any case, HGP has led to a prospective path which no return, but it still has a long way ahead before we could see the true light!

Before closing, I would like to mention some interaction with the Chinese genomics community and training young generatons. On the 2003 CSHL Genome Meeting, I was very happy to see Huanming Yang's BGI (Beijing Genomics Institute) represented the Chinese contribution to the finishing of the human genome sequencing, Zhengyu Xuan, former Runsheng Chen's student (now a faculty of UTD), was leading BGI's initial bioinformatics team (Jun Wang succeeded him) and became my postdoc in 2000. We always welcomed CSHL Genome Meeting attendants from BGI or BIG each year with warm hospitality, even after Huanming Yang and Jun Yu (Maynard Olson's former postdoc) split with Jian Wang and Jun Wang to head BIG of CAS in 2003. I often recall many friends: Bailin Hao, Chuanting Zhang, Yanda Li, Zhirong Sun, Runsheng Chen, Da-fu Ding, Liaofu Luo, Yixue Li, Luhua Lai, *et al.* who were the bioinformatics pioneers in China. Since 2000, when Chao Tang and Luhua Lei started the Center for Theoretical Biology of PKU, I have served as a SAB member. In the same year, I met Zhu Chen (then the director of the Chinese Southern Genome Center in Shanghai. When he became the Minister of Health later, I brought Mike Waterman, David Lipman to discuss Open Source Policies with him. David Botstein had asthma and canceled his trip in the last minute.) at International Symposium on Bioinformatics in Pohang, Korea, he invited me to visit his Genome Center. After I chaired the International Symposium on Bioinformatics at International HUGO Conference HGM2002, he delegated Guoping Zhao to show me around genome research institutions, Lin He's genomic disease lab and visited Yixue Li's bioinformatics empire in Shanghai. In 2004, Zhu Chen and Gang Pei were trying to recruit me to become one of the directors for a new CAS-MPG Partner Institute for Computational Biology in SIBS (Andreas Dress, a mathematician was the founding director 2005–2010, Jing-Dong (Jackie) Han, former postdoc of Marc Vidal of Dana Farber, was the second 2010–2015), I insisted Li Jin (Former postdoc of Cavalli-Sforza), would be a better fit for the job as Li was already a faculty in Fudan. But I did serve as a member of the Scientific Advisory Board chaired by Mike Waterman. Around the same time, Zhirong Sun, Jing Cheng, Zihe Rao from Dept. of Biotechnology and Yanda Li, Xuegong Zhang from Dept. of Automation convinced me to serve as an adjunct professor of Tsinghua University. In 2005, Tieniu Tan and Tianzi Jiang asked me to serve as a visiting scholar of their Institute of Automation, CAS and in the same year, Xiaodong Wang and Xingwang Deng invited me to attend their commencement ceremony of NIBS. In 2007, we organized Tsinghua summer school for bioinformatics, training many students from all over the country (edited book was published in 2003 [133]). In 2009, Pak Sham, director of Genome Research Center of the University of Hong Kong (HKU) invited me to be a visiting professor to help building up their bioinformatics team and I had a pleasant chat with Lap-Chee Tsui, then Vice-Chancellor and President of HKU about his pioneering works with Francis Collins (now NIH director) on mapping/cloning/sequencing the CF gene which sparked the HGP. In 2010, Xiaoliang (Sunney) Xie asked me to serve as an initial SAB member and attended his commencement ceremony of Biomedical Pioneering Innovation Center (BIOPIC) of PKU. In 2014, Terry Hwa (not able to attend physically), Chao Tang and I organized the first Quantitative Biology meeting at CSH Asia in Suzhou where many physicist-turned-biologists shown up (*e.g.*, Bill Bialek, Johan Elf, Thierry Emonet, Daniel Fisher, Hao Li, Qi Ouyang, Jose Onuchic, Boris Shraiman, Sander Tans, Yuhai Tu, Lingchong You).

ACKNOWLEDGEMENT

MQZ is supported by the Cecil H. and Ida Green Distinguished Chair in Systems Biology Science.

REFERENCES

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860–921
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) The sequence of the human genome. Science, 291, 1304–1351
3. Zhang, M. Q. (1988) Nonequilibrium steady states of some stochastic lattice gas models. Nucl. Phys. B Proc. Suppl., 5, 278–283
4. Zhang, M. Q., Wang, J. S., Lebowitz, J. L. and Valles, J. L. (1988) Power law decay of correlations in stationary nonequilibrium lattice gases with conservative dynamics. J. Stat. Phys., 52, 1461
5. Krug, J., Lebowitz, J.L., Spohn, H., and Zhang, M. Q. (1986) The fast rate limit of diffusive system. J. Stat. Phys., 44, 535.308
6. Zhang, M. Q. (1989) A fast vectorized multispin coding algorithm for 3D monte carlo simulations using Kawasaki Spin-exchange dynamics. J. Stat. Phys., 56, 939–950
7. Zhang, M. Q. (1987) Supersymmetrical approach to critical dynamics of relaxational models. Phys. Rev. B Condens. Matter,

36, 3824–3829
8. Mallick, K., Moshe, M. and Orland, H. (2011) A field-theoretic approach to non-equilibrium work identities. J. Phys. A, 44, 095002
9. Zhang, M. Q. and Percus, J. K. (1989) Direct correlations of the capillary wave model and construction of free energy density functional for the liquid-vapor interface system. J. Chem. Phys., 90, 3795–3799
10. Zhang, M. Q. and Percus, J. K. (1989) Inhomogeneous Ising model on a multiconnected networks. J. Stat. Phys., 56, 695–708
11. Zhang, M. Q. and Percus, J. K. (1990) A Recursive density functional formalism of nonuniform fluids. J. Chem. Phys., 92, 6779–6785
12. Zhang, M. Q. (1991) Exact response functions of a 2D fermion fluid. J. Math. Phys., 32, 1344–1349
13. Zhang, M. Q. (1991) How to find the Lax Pairs from the Yang-Baxter equations. Commun. Math. Phys., 141, 523–531
14. DeLisi, C. (1988) Computers in molecular biology: current applications and emerging trends. Science, 240, 47–52
15. Watson, J. D. (1990) The human genome project: past, present, and future. Science, 248, 44–49
16. Lander, E. S. and Waterman, M. S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2, 231–239
17. Zhang, M. Q. and Marr, T. G. (1993) Genome mapping by nonrandom anchoring: a discrete theoretical analysis. Proc. Natl. Acad. Sci. USA, 90, 600–604
18. Zhang, M. Q. and Marr, T. G. (1994) Genome mapping with random anchored clones: A discrete theoretical analysis. J. Stat. Phys., 73, 611–623
19. Zhang, M. Q. and Marr, T. G. (1994) Fission yeast gene structure and recognition. Nucleic Acids Res., 22, 1750–1759
20. Chen, T. and Zhang, M. Q. (1998) Pombe: A gene-finding and exon-intron structure prediction system for fission yeast. Yeast, 14, 701–710
21. Zhang, M. Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl. Acad. Sci. USA, 94, 565–568
22. Zhang, M. Q. (1998) Statistical features of human exons and their flanking regions. Hum. Mol. Genet., 7, 919–932
23. Liu, H. X., Zhang, M. and Krainer, A. R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev., 12, 1998–2012
24. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell, 9, 3273–3297
25. Karlin, S. and Altschul,S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA, 87, 2264–2268
26. Burge, C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268, 78–94
27. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467–470
28. Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. and Davis, R. W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA, 94, 13057–13062
29. Getz, G., Levine, E., Domany, E. and Zhang, M. Q. (2000) Super-paramagnetic clustering of yeast gene expression profiles. Physica A, 279, 457–464
30. Gilbert, W. (1991) Towards a paradigm shift in biology. Nature, 349, 99
31. Pennisi, E. (2003) Human genome: A low number wins the genedweep pool. Science, 300, 1484.
32. Guigó, R., Birney,E., Brent, M., Dermitzakis, E., Pachter, L., Roest Crollius, H., Solovyev, V. and Zhang, M. Q. (2004) Needed for completion of the human genome: hypothesis driven experiments and biologically realistic mathematical models. arXiv, q-bio/0410008
33. Zhu, J. and Zhang, M. Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. Bioinformatics, 15, 607–611
34. Zhang, M. Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. Comput. Chem., 23, 233–250
35. Lydall, D., Ammerer, G. and Nasmyth, K. (1991) A new role for MCM1 in yeast: cell cycle regulation of SW15 transcription. Genes Dev., 5, 2405–2419
36. Zhu, G., Spellman, P.T., Volpe, T., Brown, P.O., Botstein, D., Davis, T.N. and Futcher, B. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. Nature 406, 90–94
37. Iyer, V. and Struhl,K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J., 14, 2570–2579
38. Blat, Y. and Kleckner, N. (1999) Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell, 98, 249–259
39. Ren, B., Robert, F., Wyrick, J.J., Aparico, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000) Science Genome-wide location and function of DNA binding proteins. Science, 290, 2306–2309
40. Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. and Brown, P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature, 409, 533–538
41. Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. Cell, 106, 697–708
42. Kato, M., Hata, N., Banerjee, N., Futcher, B. and Zhang, M. Q. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol., 5, R56
43. Ioshikhes, I., Trifonov, E. N. and Zhang, M. Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc. Natl. Acad. Sci. USA,

96, 2891–2895
44. Ioshikhes, I. P. and Zhang, M. Q. (2000) Large-scale human promoter mapping using CpG islands. Nat. Genet., 26, 61–63
45. Davuluri, R. V., Suzuki, Y., Sugano, S. and Zhang, M. Q. (2000) CART classification of human 5′ UTR sequences. Genome Res., 10, 1807–1816
46. Davuluri, R. V., Grosse, I. and Zhang, M. Q. (2001) Computational identification of promoters and first exons in the human genome. Nat. Genet., 29, 412–417
47. Kel, A. E., Kel-Margoulis, O. V., Farnham, P. J., Bartley, S. M., Wingender, E. and Zhang, M. Q. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J. Mol. Biol., 309, 99–120
48. Weinmann, A. S., Bartley, S. M., Zhang, T., Zhang, M. Q. and Farnham, P. J. (2001) Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol. Cell. Biol., 21, 6820–6832
49. Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R. A. and Dynlacht, B. D. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev., 16, 245–256
50. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko,V., Cheng, J., Williams, A. J., *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell, 116, 499–509
51. Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. and Ren, B. (2005) A high-resolution map of active promoters in the human genome. Nature, 436, 876–880
52. Alkema, W. and Wasserman,W. W. (2003) Understanding the language of gene regulation. Genome Biol., 4, 327
53. Bussemaker, H. J., Li, H. and Siggia, E. D. (2001) Regulatory element detection using correlation with expression. Nat. Genet., 27, 167–171
54. Das, D., Banerjee, N. and Zhang, M. Q. (2004) Interacting models of cooperative gene regulation. Proc. Natl. Acad. Sci. USA, 101, 16234–16239
55. Das, D., Nahle, Z. and Zhang, M. Q. (2006) Adaptively inferring human transcriptional subnetworks. Mol. Sys. Biol., 2:2006.0029
56. Das, D. and Zhang, M. Q. (2007) Predictive Models of Gene Regulation: Application of Regression Methods to Microarray Data, ed. Korenberg, M., 377: 95, In: Methods in Molecular Biology. Springer
57. Sumazin, P., Chen, G., Hata, N., Smith,A. D., Zhang, T. and Zhang, M. Q. (2005) DWE: discriminating word enumerator. Bioinformatics, 21, 31–38
58. Smith, A. D., Sumazin, P. and Zhang, M. Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proc. Natl. Acad. Sci. USA, 102, 1560–1565
59. Smith, A. D., Sumazin, P., Das, D. and Zhang, M. Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. Bioinformatics, 21, i403–i412
60. Smith, A. D., Sumazin, P., Xuan, Z. and Zhang, M. Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. Proc. Natl. Acad. Sci. USA, 103, 6275–6280
61. Smith, A. D., Sumazin, P. and Zhang, M. Q. (2007) Tissue-specific regulatory elements in mammalian promoters. Mol. Syst. Biol., 3, 73
62. Stamm, S., Zhang, M. Q., Marr, T. G. and Helfman, D. M. (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Res., 22, 1515–1526
63. Stamm, S., Zhu, J., Nakai, K., Stoilov,P., Stoss, O. and Zhang, M. Q. (2000) A database and statistical analysis of aternative exons. DNA Cell Biol., 19, 739–756
64. Zhang, M. Q. (2001) Discriminant analysis and its application in DNA sequence motif recognition. Brief. Bioinform., 1, 331–342
65. Wu, Q. and Maniatis, T. (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell, 97, 779–790
66. Wu, Q., Zhang, T., Cheng, J.-F., Kim, Y., Grimwood, J., Schmutz, J., Dickson, M., Noonan, J. P., Zhang, M. Q., Myers, R. M., *et al.* (2001) Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. Genome Res., 11, 389–404
67. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., Jung, I., Wu, H., Zhai,Y., Tang, Y., *et al.* (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell, 162, 900–910
68. Prasanth, K. V., Prasanth, S. G., Xuan, Z., Hearn, S., Freier, S. M., Bennett, C. F., Zhang, M. Q., Spector, D. L. and Spector, D. L. (2005) Regulating gene expression through RNA nuclear retention. Cell, 123, 249–263
69. Bernard, D., Prasanth, K. V., Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., Zhang, M. Q., Sedel, F., Jourdren, L., Coulpier, F., *et al.* (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J., 29, 3082–3093
70. Chen, G., Li, W., Zhang, Q.-S., Regulski, M., Sinha, N., Barditch, J., Tully, T., Krainer, A. R., Zhang, M. Q. and Dubnau, J. (2008) Identification of synaptic targets of *Drosophila pumilio*. PLOS Comput. Biol., 4, e1000026
71. Xie, P., Liu, Y., Li, Y., Zhang, M. Q. and Wang, X. (2014) MIROR: a method for cell-type specific microRNA occupancy rate prediction. Mol. Biosyst., 10, 1377–1384
72. Yuan, Y., Liu, B., Xie, P., Zhang, M. Q., Li, Y., Xie, Z. and Wang, X. (2015) Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. Proc. Natl. Acad. Sci. USA, 112, 3158–3163
73. He, M., Liu, Y., Wang, X., Zhang, M. Q., Hannon, G. J. and Huang, Z. J. (2012) Cell-type-based analysis of microRNA profiles in the mouse brain. Neuron, 73, 35–48
74. Zhang, M. Q. and Marr, T. G. (1993) A weight array method for splicing signal analysis. Comput. Appl. Biosci., 9, 499–509
75. Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. and Krainer, A. R. (2003) ESEfinder: A web resource to identify exonic splicing

enhancers. Nucleic Acids Res., 31, 3568–3571
76. Zhang, C., Hastings, M. L., Krainer, A. R. and Zhang, M. Q. (2007) Dual-specificity splice sites function alternatively as 5′ and 3′ splice sites. Proc. Natl. Acad. Sci. USA, 104, 15028–15033
77. Zhang, C., Krainer, A. R. and Zhang, M. Q. (2007) Evolutionary impact of limited splicing fidelity in mammalian genes. Trends Genet., 23, 484–488
78. Zhang, C., Zhang, Z., Castle, J., Sun, S., Johnson, J., Krainer, A. R. and Zhang, M. Q. (2008) Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. Genes Dev., 22, 2550–2563
79. Weyn-Vanhentenryck, S., Sun, S., Yan, Q., Mele, A., Farny, N., Silver, Z., Zhang,M. Q., Krainer, A.R., Darnell, R.B. and Zhang, C. (2014) HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep., 6, 1139–1152
80. Castle, J. C., Zhang, C., Shah, J. K., Kulkarni, A. V., Cooper, T. A. and Johnson, J. M. (2008) Differential expression of 24,426 human alternative splicing events and predicted *cis*-regulation in 48 tissues. Nat. Genet., 40, 1416
81. Wu, J., Akerman, M., Sun, S., McCombie, W. R., Krainer, A. R. and Zhang, M. Q. (2011) SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics, 27, 3010–3016
82. Wu, J., Anczuków,O., Krainer, A. R., Zhang, M. Q. and Zhang, C. (2013) OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res., 41, 5149–5163
83. Akerman, M., Fregoso, O. I., Das, S., Ruse, C., Jensen, M. A., Pappin, D. J., Zhang, M. Q. and Krainer, A. R. (2015) Differential connectivity of splicing activators and repressors to the human spliceosome. Genome Biol., 16, 119
84. Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J. and Bestor, T. H. (2006) Large-scale structure of genomic methylation patterns. Genome Res., 16, 157–163
85. Das, R., Dimitrova,N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H. and Zhang, M. Q. (2006) Computational prediction of methylation status in human genomic sequences. Proc. Natl. Acad. Sci. USA, 103, 10713–10716
86. Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L., *et al.* (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res., 19,1593–1605
87. Zhang, M. Q. and Smith, A. D. (2010) Challenges in understanding genome-wide DNA methylation. J. Comp. & Tech., 25, 26–34
88. Harris, R. A., Wang, T., Coarfa, C., Nagarajan,R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y., *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat. Biotechnol., 28, 1097–1105
89. Chung, W.-Y., Schmitz, R. J., Biorac, T., Ye, D., Dudas, M., Meredith, G. D., Adams, C. C., Ecker, J. R. and Zhang, M. Q. (2013) Constructing hepitypes: phasing local genotype and DNA methylation. JNSNE, 4, 335–346
90. Wu, D., Gu, J. and Zhang, M. Q. (2013) FastDMA: an infinium humanmethylation450 beadchip analyzer. PLoS One, 8, e74275
91. Ma, X., Wang, Y. W., Zhang, M. Q. and Gazdar, A. F. (2013) DNA methylation data analysis and its application to cancer research. Epigenomics, 5, 301–316
92. Zhang, Y.A., Ma, X., Sathe, A., Fujimoto, J., Wistuba, I.I., Lam, S., Yatabe, Y., Wang, Y.W., Stastny V, Gao, B., *et al.* (2016) Validation of SCT Methylation As a Hallmark Biomarker for Lung Cancers. J. Thorac. Oncol., 11, 346–360
93. Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M. Q., Chen, P. Y. and Pellegrini, M. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics, 14, 774
94. Guo, W., Chung, W. Y., Qian, M., Pellegrini, M. and Zhang, M. Q. (2014) Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. Nucleic Acids Res., 42, 3009–3016
95. Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Peng, W., Zhang, M. Q., *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet., 40, 897–903
96. Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell, 153, 1134–1148
97. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C. A., Lin, S., Lin, Y., Qiu,Y., *et al.* (2015) Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature, 518, 350–354
98. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330
99. Zhao, X., Xuan, Z. and Zhang, M. Q. (2007) Boosting with stumps for predicting transcription start sites. Genome Biol., 8, R17
100. Wang, X., Xuan, Z., Zhao, X., Li, Y. and Zhang, M. Q. (2009) High-resolution human core-promoter prediction with CoreBoost_HM. Genome Res., 19, 266–275
101. Xing, H., Liao, W., Mo, Y. and Zhang, M. Q. (2012) A novel Bayesian change-point algorithm for genome-wide analysis of diverse ChIPseq data types. J. Vis. Exp., 70, e4273
102. Xing, H., Mo, Y., Liao, W. and Zhang, M. Q. (2012) Genomewide localization of protein-DNA binding and histone modification by BCP with ChIP-seq data. PLoS. Comput. Biol., 8, e1002613
103. Thomas, R., Thomas, S., Holloway, A. K. and Pollard, K. S. (2017) Features that define the best ChIP-seq peak calling algorithms. Brief Bioinform., 18, 441–450
104. Ouyang, W., Liao, W., Luo, C. T., Yin,N., Huse, M., Kim, M. V., Peng, M., Chan, P., Ma,Q., Mo, Y., *et al.* (2012) Novel Foxo1-dependent transcriptional programs control T(reg) cell function. Nature, 491, 554–559

105. Xuan, Z. and Zhang, M. Q. (2005) From worm to human: bioinformatics approaches to identify FOXO target genes. Mech. Ageing Dev., 126, 209–215
106. Renault, V. M., Thekkat, P. U., Hoang, K. L., White, J. L., Brady, C. A., Kenzelmann Broz, D., Venturelli, O. S., Johnson, T. M., Oskoui, P. R., Xuan, Z., *et al.* (2011) The pro-longevity gene FoxO3 is a direct target of the p53 tumor suppressor. Oncogene, 30, 3207–3221
107. Kim, T. H., Abdullaev,Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov,V. V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell, 128, 1231–1245
108. Barrera, L. O., Li, Z., Smith, A. D., Arden, K. C., Cavenee, W. K., Zhang, M. Q., Green, R. D. and Ren, B. (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res., 18, 46–59
109. He, C., Zhang, M. Q. and Wang, X. (2015) MICC: an R package for identifying chromatin interactions from ChIA-PET data. Bioinformatics, 31, 3832–3834
110. Li, G., Chen, Y., Snyder, M. P. and Zhang, M. Q. (2017) ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. Nucleic Acids Res., 45, e4
111. Djekidel, M. N., Liang, Z., Wang, Q., Hu, Z., Li, G., Chen, Y. and Zhang, M. Q. (2015) 3CPET: finding co-factor complexes from ChIA-PET data using a hierarchical Dirichlet process. Genome Biol., 16, 288
112. Djekidel, M. N., Chen, Y. and Zhang, M. Q. (2018) FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. Genome Res., 28, 412–422
113. Chen, F., Li, G., Zhang, M. Q. and Chen, Y. (2018) HiCDB: a sensitive and robust method for detecting contact domain boundaries. Nucleic Acids Res., 46, 11239–11250
114. He, C., Wang, X., and Zhang, M. Q. (2014) Nucleosome eviction and multiple co-factor binding predict estrogen receoptor alpha associated long-range interactions. Nucleic Acids Res., 42, 6935–6944
115. Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q. and Snyder, M.P. (2014) Genome-wide map of regulatory interactions in the human genome. Genome Res., 24, 1905–1917
116. Chen, Y., Wang, Y., Xuan, Z., Chen, M. and Zhang, M. Q. (2016) *De novo* deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. Nucleic Acids Res., 44, e106
117. Ramanand, S. G., Chen, Y., Yuan, J., Daescu,K., Lambros, M. B., Houlahan, K. E., Carreira, S., Yuan, W., Baek, G., Sharp, A., *et al.* (2020) The landscape of RNA polymerase II-associated chromatin interactions in prostate cancer. J Clin Invest, 130, 3987–4005
118. Liu, X., Zhang, Y., Chen, Y., Li, M., Shao, Z., Zhang, M.Q. and Xu, J. (2018) CAPTURE: *In situ* analysis of chromatin composition of endogenous genomic loci by biotinylated dCas9. Curr. Protoc. Mol. Biol. 123, e64
119. Liu, X., Chen, Y., Zhang, Y., Liu,Y., Liu, N., Botten, G. A., Cao, H., Orkin, S. H., Zhang, M. Q. and Xu, J. (2020) Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9. Genome Biol., 21, 59
120. Liang, Z., Li, G., Wang, Z., Djekidel,M. N., Li, Y., Qian, M. P. Zhang, M.Q. and Chen, Y. (2017) BL-Hi-C: efficient and sensitive approach for structural and regulatory chromatin interactions. Nat. Commun., 8, 1622
121. Xiao, R., Chen, J. Y., Liang, Z., Luo, D., Chen, G., Lu, Z. J., Chen, Y., Zhou, B., Li,H., Du, X., *et al.* (2019) Pervasive chromatin-RNA binding protein interactions enable RNA-based regulation of transcription. Cell, 178, 107–121.e18
122. Chen, F. X., Xie, P., Collings, C. K., Cao, K., Aoi, Y., Marshall, S. A., Rendleman, E. J., Ugarenko,M., Ozark, P. A., Zhang, A., *et al.* (2017) PAF1 regulation of promoter-proximal pause release via enhancer activation. Science, 357, 1294–1298
123. Kim, Y. J., Xie, P., Cao, L., Zhang, M.Q., and Kim, T.H. (2018) Global transcriptional activity dynamics reveal functional enhancer RNAs. Genome Res., 28, 1799–1811
124. Ni, Y., Cao, B., Ma, T., Niu,G., Huo, Y., Huang, J., Chen, D., Liu, Y., Yu,B., Zhang, M. Q., *et al.* (2017) Super-resolution imaging of a 2.5 kb non-repetitive DNA *in situ* in the nuclear genome using molecular beacon probes. eLife, 6, e21660
125. Zhanghao, K., Chen, L., Yang, X., Wang, M., Jing, Z., Han, H., Zhang, M. Q., Jin, D., Gao, J. and Xi, P. (2016) Super-resolution dipole orientation mapping via polarization demodulation. Light Sci. Appl., 5, e16166
126. Shi, M., Gao, J. and Zhang, M. Q. (2017) Web3DMol: interactive protein structure visualization based on WebGL. Nucleic Acids Res., 45, W523–W527
127. Niu, J., Zhang, X., Li, G., Yan, P. X., Yan, Q., Dai, Q. H., Jin, D. Y., Shen, X. H., Wang, J. G., Zhang, M. Q., *et al.* (2020) Novel cytogenetic method to image chromatin interactions with sub-kilobase resolution: Tn5-FISH. J. Genet. Genomics, 47, 123
128. Li, X., Chen, W., Chen, Y., Zhang, X., Gu, J. and Zhang, M. Q. (2017) Network embedding-based representation learning for single cell RNA-seq data. Nucleic Acids Res., 45, e166
129. Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O. M., Zhang, M. Q., Jiang, R. and Chen, T. (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nat. Commun., 8, 22
130. Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., Von Hoff, D., Han, H., Zhang, M. Q. and Lin, W. (2019) SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. Nucleic Acids Res., 47, e48
131. Xiong, L., Xu, K., Tian, K., Shao,Y., Tang, L., Gao, G., Zhang, M., Jiang, T. and Zhang, Q. C. (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat. Commun., 10, 4576
132. MacArthor, B. (2021) Truth and beauty in physics and biology. Nat. Phys., 17, 149–151
133. Jiang, R., Zhang, X., and Zhang, M. Q., (2013) Basics of Bioinformatics: Lecture Notes of Graduate Summer School on Bioinformatics of China. Springer