Big-data analysis unearths the general regulatory regime in normal human genome and cancer

Anyou Wang[1], Hai Rong[1,2]

[1] The Institute for Integrative Genome Biology, University of California at Riverside, Riverside, CA 92521, USA, and [2]Department of Microbiology and Plant Pathology, University of California at Riverside, Riverside, CA 92521, USA

[*]Correspondence:
A Wang anyou.wang@alumni.ucr.edu

Gene regulation interprets most variations of biological phenotype and remains a crucial topic in biology. Conventionally, manipulating gene sequences like knockout helps to infer gene regulation, but these inferences suffer several pitfalls like transcript compensation[1] , leading to biased results. An unbiased regulation has rarely been appreciated. Here, we develop a software, FINET[2], to infer unbiased regulatory networks from massive data, including all human RNAseq data publicly available from Sequence Read Archive (SRA, 274469 samples) and The Cancer Genome Atlas (TCGA, 11574 samples), and unearth the general regulatory rules in normal genome and cancer as deposited[3]. Generally, the genome is positively regulated. Regulators primarily self-regulate their targets in the same annotated category, like processed-pseudogenes regulating processed-pseudogenes. At normal, ribosomal proteins drive the regulatory network, and proteins tightly control the genome and primarily regulate the remote proteins across chromosomes, but rarely regulate local targets (<1M bp), yet cancer noncoding RNAs, especially pseudogenes, strongly activate the cancer genome and induce local targets, including noncoding RNAs and proteins. As a result, the whole regulatory regime switches from a normal remote protein-controlled domain to a cancerous local noncoding RNA-activated niche. This parallels with our recent discovery from clinical data revealing noncoding RNAs as the deadliest drivers for cancer[4], instead of proteins as conventionally thought. This refreshes the fundamental basis of cancer research and therapy. Our overall finding provides a systems version of the natural regulatory regime in human genome, which helps to correct the biased notions standing in current literature.

To generate a general gene regulatory network, we needed to obtain data with all conditions and a software. SRA and TCGA provides such data, referred to as normal and cancer respectively (materials and methods). We developed FINET[2], which combines julia language with elastic-net and stability-selection methods, and can efficiently and accurately infer regulatory interactions with >90% true positive ratio (true positives/total true callings), dramatically improving when compared to current software[5]. We used FINET to search all possible gene regulations by treating each gene as target to select its regulators from the rest of all annotated genes (materials and methods). This search generated a comprehensive network for normal and cancer, respectively, and contained 19721 nodes (genes) with 63878 interactions, and 25402 nodes with 61772 interactions (**Fig. 1A**). As expected, our networks were much less complex than normal because they only collected the reliable positive interactions with generality in all conditions.

To understand the core controllers of the normal and cancer networks, we investigated each network centrality and modules[6] (materials and methods). Ribosomal proteins dominate the top 20 centrality in normal, while telomerase ranks at the top in the cancer (**Fig. 1A**). At normal, the number one centrality, RPS23, works as a bridge in a module (**module #35 in table S1**) and as an inducer to RPL39, which strongly activated a hub in the module, SSBP1 (R^2_{adj}=0.48, p< 2.2e-16, **Fig. 1B-1C**). SSBP1 did not interact with any members of this module in STRING database, although the interactions between ribosomal proteins have also been identified in STRING **(Fig. S1).** This suggests that the current knowledge of gene regulations is seriously biased to genes with similar functions, but lacks depth and investigation about interactions between seemingly irrelevant genes. In addition, only 7% of interactions were mapped to the STRING database (materials and methods), suggesting the novelty of our network.

We clustered network modules into either protein or noncoding module (materials and methods). At normal, protein modules (proteins > 50% in a module members) occupied 60.52% of total 38 modules (**table S1**) and noncoding (noncodings >50% in module components) only took 28.94% (**Fig. S1B**), while cancer modules significantly changed their compositions (p-value = 0.02963, Pearson's Chi-squared test with Yates' continuity correction, referred as chisq-test thereafter), with protein reduced to 47.29% and

noncoding increased to 45.94% of total 74 cancer modules (**Fig. S1B**, **table S2**). In addition, the overall regulators and targets also significantly shifted from normal proteins to cancer noncoding RNAs (p-value = 0.06917,chisq-test, **Fig. S1C**).

Based on gene ontology (GO), these top 20 centralities in normal networks perform crucial functions in translation (RPL18, RPL3, RPL30, RPL39, RPS10, RPS23, RPS28). Consistently, the functions for the whole normal network and modules (**table S1**) were also relevant to translation and negative regulations, suggesting ribosomal proteins as the active, tightly regulated core of the human genome. In contrast, telomere maintenance guided by telomerase served as the function for top 20 cancer centrality (HNRNPA1, HNRNPC, HNRNPU). The whole cancer network and modules (**table S2**) also executed relevant functions and regulation of the cell cycle and alternative splicing, but these proteins activated in cancer were actually driven by noncoding RNAs as shown below and in our another study[4].

To understand general regulation direction (positive or negative), we examined the coefficient distribution of regulators. In total, more than 99% of regulators in both normal and cancer carried positive coefficients referred to as the inducer, and less than 1% as repressors (**Fig S2A-S2B**). This evidence indicates that the human genome is positively regulated in both normal and cancer, consistent with GENCODE observation[7]. However, cancer has much stronger inducers, with significantly higher coefficients than normal ($p<1.1e-9$ **Fig S2C**). Furthermore, a cancer genome's lost repressors (**Fig S2D**), leads to increased protein activation ($R^2_{adj}=0.53, p<0.0155$, **Fig 1D**). These together suggest the cancer genome, in a highly activated state, featured greater, stronger inducers and fewer repressors.

To understand the general gene regulatory rule of regulators, we examined the distribution of regulator numbers and their targets in top 4 abundant gene categories: protein-coding, processed-pseudogene (p-pseudogene), lincRNA, and antisense RNA. When protein-coding genes served as regulators at both normal and cancer, their targets were mostly proteins(>87%), and the noncoding gene targets are very limited. On the other hand, when proteins worked as targets, their regulators were also proteins (>87%, **Figure 1E**). Surprisingly, when p-pseudogenes served as regulators, their targets were also mostly (>60%) p-pseudogenes (**Figure 1F**). The reverse was also true. Furthermore, similar patterns of self-regulation occurred in lincRNAs and antisense RNAs (**Figure S3A-S3B**). This evidence places self-regulation in the same gene category as one of the general regulatory rules in biology.

To appreciate the target location rule debated **recently**[8 7 9], we investigated the target site distribution of top four gene categories. Generally in both normal and cancer, all regulators prefer to target their targets densely in two locations: either outside the chromosome as trans-regulators, or less than 1M bp as cis-regulators (**Figure 2A-2B**). Yet the target density in these two locations were quite different between normal and cancer. At normal, p-pseudogene, protein and lincRNA served as trans-regulators and mostly regulated their targets located outside the chromosome, 80%, 60%, 55% respectively (**Figure 2A**), while antisense RNAs worked as cis-regulators and targeted mostly (>50%) local (<1M bp) targets. In addition, 25% and 10% of antisense RNAs and lincRNAs respectively regulated their targets with overlapped sequences, while p-pseudogenes and proteins rarely regulated their targets with overlapped sequences. In cancer, however, most regulators of all categories turn to regulate their targets locally (<1M bp **Figure 2B**). Specifically, more than 80% of lincRNAs and antisense RNAs and significant proteins (45%) worked locally. This suggests the target site rule switching from remote in normal to local in cancer.

To appreciate the strongest regulatory rule and to make our pattern robustC, we examined the composition of the top regulators and their targets. Based on coefficient, the top 300 inducers and repressors were selected respectively for normal and cancer. At normal, proteins worked as the strongest

inducers. From the top 300 to top 10 inducers, proteins occupied 60% out of top 300 and 50% at top 10 (**Figure 3A-3C**). lincRNAs came next and occupied ~20%. Yet in cancer, proteins even did not show up (<10%) and noncoding RNAs dominated the top inducers instead. For example, p-pseudogenes counted 70% out of top 10 (Figure 3A), suggesting noncoding RNAs as the primary rulers in cancer genome. Interestingly, these noncoding RNAs almost purely targeted proteins (>98%) in cancer, while in normal, proteins, p-pseudogene, and lincRNA became targets (**Figure 3B, 3C**). This suggested that proteins work as targets instead of fundamental drivers for cancer, and the conventional practice to treat proteins as cancer drivers is a mistake. Consistently, analyzing clinical data also reveals noncoding RNAs as the deadliest regulators for all types of cancers[4].

As for the strongest repressors, almost all repressors were proteins at normal, while at cancer repressors contained antisense RNAs, p-pseudogene, and proteins, with at least 10% at each (**Fig S4A**). Surprisingly, regardless of repressors, almost all targets were proteins >85%, and any noncoding RNA category was too low to show (<10%) (**Fig. S4B**). Identifying proteins as targets for both inducers and repressors in cancer interprets why the current observations have focused on proteins, yet treating targets as cancer drivers is a big mistake. Our discoveries identifying noncoding RNAs as key cancer drivers evidently presents a novel basis for understanding cancer, eventually making all cancers curable.

Our overall results establish fundamental rules and resources of general gene regulations in the human genome.

Availability:
software FINET and network were respectively deposited in github[2] and project website[3]

Materials and methods included in supplemental data.

Figure legends

Figure 1. Central rulers for human genome and gene self-regulatory law. A, general regulatory network and its top 20 central rulers for normal (left) and cancer (right). B, a module containing top 1 ruler (RPS23) in normal. Edge color represents strength of regulation, darker, stronger, and arrow as regulatory directions, arrow as positive and square as negative. C, linear regression between SSBP1 and RPL39. D, linear regression between delta protein percentage and delta repressor percentage (normal percentage-cancer percentage). E-F, profiling self-regulated gene categories for protein (E) and p-pseudogene (F). Regulators and their targets were counted separately. For example, in protein(E) at normal, proteins work as regulators and count their targets, marked as normal-target; in reverse, proteins serve as targets and count their regulator, denoted as normal-regulators(E).

Figure 2. Target distance rules. A-B, target location distribution of top abundant gene categories (>10%) in normal (A) and cancer (B). outchro: target locating outside chromosome, M as million inside the chromosome.

Figure 3. Gene categories of top strongest inducers and their targets. A, profiling gene categories of top 300 inducers at normal and cancer, from top 300 to top 10 inducer. B, the gene categories of targets of top 300 inducers of normal and cancer. C, top 10 inducer lists at normal (upper) and cancer (bottom).

References

1.El-Brolosy, M. A. *et al.* Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**, 193 (2019).
2.Wang, A. & Hai, R. FINET: Fast Inferring NETwork. *bioRxiv* 733683 (2019). doi:10.1101/733683
3.Wang, A. & Hai, R. human regulatory network http://combai.org/network/. (2019).
4.Wang, A. & Hai, R. Noncoding RNAs serve as the deadliest regulators for cancer. (2019).
5.Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
6.Systematically dissecting the global mechanism of miRNA functions in mouse pluripotent stem cells | BMC Genomics | Full Text. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1706-y. (Accessed: 2nd October 2019)
7. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
8.Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
9.Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510 (2019).

Fig1

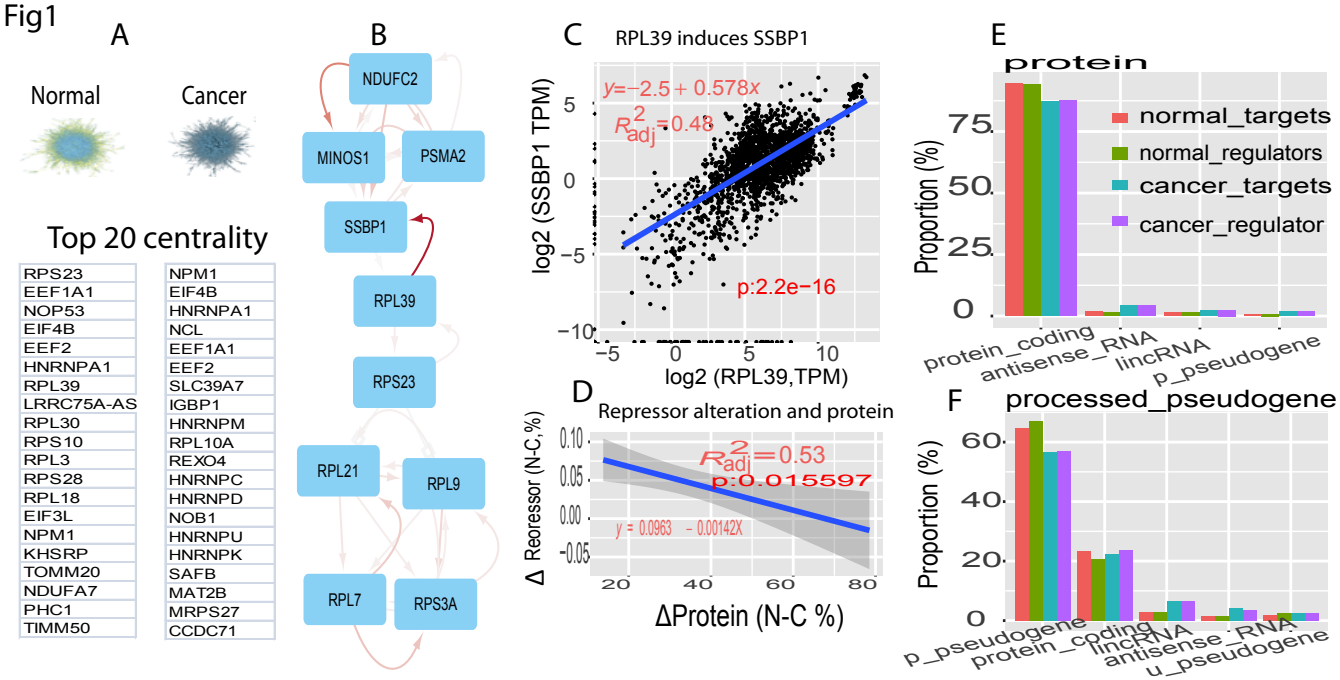

A
Normal
Cancer
Top 20 centrality
RPS23
EEF1A1
NOP53
EIF4B
EEF2
HNRNPA1
RPL39
LRRC75A-AS
RPL30
RPS10
RPL3
RPS28
RPL18
EIF3L
NPM1
KHSRP
TOMM20
NDUFA7
PHC1
TIMM50
NPM1
EIF4B
HNRNPA1
NCL
EEF1A1
EEF2
SLC39A7
IGBP1
HNRNPM
RPL10A
REXO4
HNRNPC
HNRNPD
NOB1
HNRNPU
HNRNPK
SAFB
MAT2B
MRPS27
CCDC71
B
NDUFC2
MINOS1
PSMA2
SSBP1
RPL39
RPS23
RPL21
RPL9
RPL7
RPS3A
C
RPL39 induces SSBP1
y=−2.5 + 0.578x
R2adj=0.48
p:2.2e−16
log2 (SSBP1 TPM)
log2 (RPL39,TPM)
D
Repressor alteration and protein
R2adj=0.53
p:0.015597
y = 0.0963 − 0.00142X
Δ Reoressor (N-C,%)
ΔProtein (N-C %)
E
protein
normal_targets
normal_regulators
cancer_targets
cancer_regulator
Proportion (%)
protein_coding
antisense_RNA
lincRNA
p_pseudogene
F
processed_pseudogene
Proportion (%)
p_pseudogene
protein_coding
lincRNA
antisense_RNA
u_pseudogene

Fig2

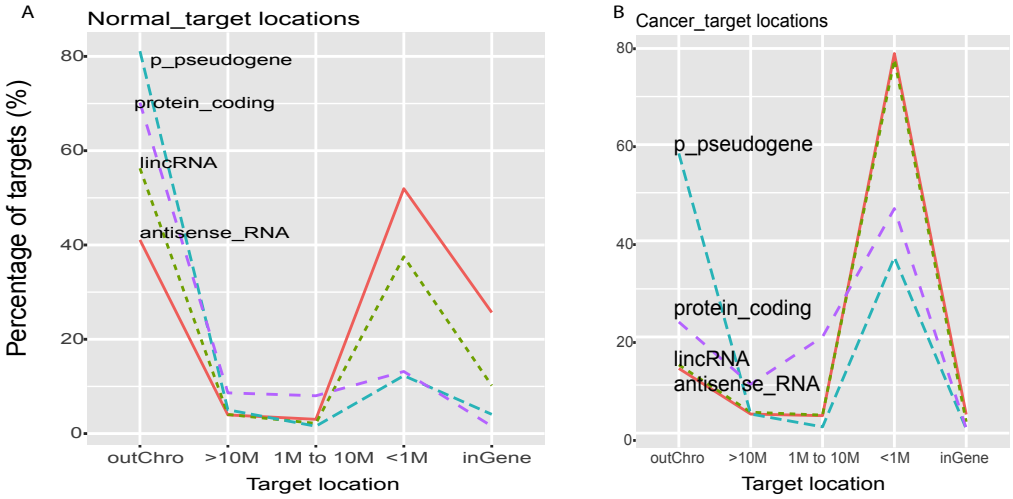
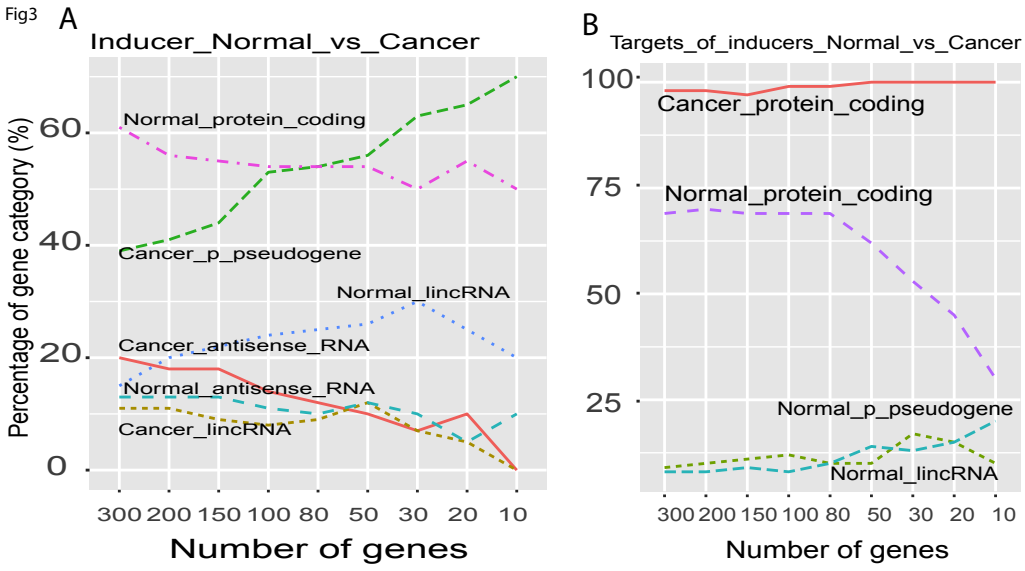

A
Normal_target locations
Percentage of targets (%)
80
60
40
20
0
p_pseudogene
protein_coding
lincRNA
antisense_RNA
outChro
>10M
1M to 10M
<1M
inGene
Target location
B
Cancer_target locations
80
60
40
20
0
p_pseudogene
protein_coding
lincRNA
antisense_RNA
outChro
>10M
1M to 10M
<1M
inGene
Target location

Fig3



C

Top 10 inducers and their targets at normal

Inducer	icat	Target	tcat	coef
HLA-DQA1	protein	MT-TD	Mt_tRNA	1679.481
METTL9	protein	EEF1A1P5	p_pseudogene	486.7019
AL117190.1	lincRNA	SNORD114-20	snoRNA	480.983
CFL1P1	tu_pseudogene	FBP1	protein	426.4937
AC012363.1	antisense_RNA	RNU1-122P	snRNA	359.018
MTX1P1	u_pseudogene	MT-CO1	protein	304.6804
ANKRD36B	protein	MALAT1	lincRNA	289.9498
AC004069.1	lincRNA	MT-TF	Mt_tRNA	275.1722
CCNK	protein	MT-ND4L	protein	251.5023
MRPL45	protein	RPL41P1	p_pseudogene	240.9145

Top 10 inducers and their targets at cancer

inducer	icat	target	tcat	coef
CTB-129P6.7	TEC	APOE	protein	11054.65
ASS1P9	p_pseudogene	ASS1	protein	10482.21
PABPC3	protein	PABPC1	protein	9182.533
HMGA1P1	p_pseudogene	HMGA1	protein	6368.683
MIR483	miRNA	IGF2	protein	6038.805
KRT18P28	p_pseudogene	KRT18	protein	5771.494
RP11-144L1.8	p_pseudogene	HSP90AA1	protein	4301.799
KRT18P19	p_pseudogene	KRT18	protein	4114.136
RPL10P3	p_pseudogene	RPL10	protein	3784.052
KRT18P11	p_pseudogene	KRT18	protein	3723.988