M.S. THESIS

Measuring intratumor heterogeneity using network entropy by RNA-seq data.

전사체 정보와 엔트로피를 이용하여 종양 내 이질성도를 측정하는 방법

FEBRUARY 2017

INTERDISCIPLINARY PROGRAM IN BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Park Youngjune

M.S. THESIS

Measuring intratumor heterogeneity using network entropy by RNA-seq data.

전사체 정보와 엔트로피를 이용하여 종양 내 이질성도를 측정하는 방법

FEBRUARY 2017

INTERDISCIPLINARY PROGRAM IN BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Park Youngjune

Measuring intratumor heterogeneity using network entropy by RNA-seq data.

전사체 정보와 엔트로피를 이용하여 종양 내 이질성도를 측정하는 방법

지도교수 김 선

이 논문을 이학석사학위논문으로 제출함

2016년 11월

서울대학교 대학원

생물정보학 협동과정

박 영 준

박영준의 석사학위논문을 인준함

2016년 12월

위 원 장 이병재 교수님 (인)

부 위 원 장 김선 교수님 (인)

위 원 박태성 교수님 (인)

Abstract

Measuring intratumor heterogeneity using network entropy by RNA-seq data.

Park Youngjune
Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University

Intratumor heterogeneity (ITH) is observed at different stages of tumor progression, metastasis and reouccurence, which can be important for clinical applications. We used RNA-sequencing data from tumor samples, and measured the level of ITH in terms of biological network states. To model complex relationships among genes, we used a protein interaction network to consider gene-gene dependency. ITH was measured by using an entropy-based distance metric between two networks, nJSD, with Jensen-Shannon Divergence (JSD). With nJSD, we defined transcriptome-based ITH (tITH). The effectiveness of tITH was extensively tested for the issues related with ITH using real biological data sets. Human cancer cell line data and single-cell sequencing data were investigated to

verify our approach. Then, we analyzed TCGA pan-cancer 6,320 patients. Our result was in agreement with widely used genome-based ITH inference methods, while showed better performance at survival analysis. Analysis of mouse clonal evolution data further confirmed that our transcriptome-based ITH was consistent with genetic heterogeneity at different clonal evolution stages. Additionally, we found that cell cycle related pathways have significant contribution to increasing heterogeneity on the network during clonal evolution. We believe that the proposed transcriptome-based ITH is useful to characterize heterogeneity of a tumor sample at RNA level.

Keywords: Tumour heterogeneity, Clonal evolution, Network topology, Entropy, Gene expression analysis, RNA sequencing, Data mining, Diagnosis.

Student Number: 2015-20505

Contents

Abstract i

Contents iii

List of Figures v

List of Tables ix

Chapter 1 Introduction 1

1.1 Cancer Genomics 1

1.1.1 Cellular Information 1

1.1.2 Reading cellular information with sequencing . 2

1.1.3 Cancer informatics and tumor heterogeneity . . 3

1.2 Intratumor heterogeneity 4

1.3 Recent works for inference intratumor heterogeneity 5

1.4 Motivation 7

1.4.1 Transcriptome based approach 7

1.4.2 Entropy based metric for network analysis . . 8

1.4.3 Our work 9

Chapter 2 Materials & Methods 11

2.1 Gene expression data 11

2.2 Pathway and protein interaction network 12
2.3 *in silico* simulation with cancer cell lines data 12
2.4 tITH calculation of TCGA patients
and comparison to gITH results 12
2.5 Calculation of nJSD 13
2.6 Calculation of transcriptome-based ITH 14

Chapter 3 Results **18**
3.1 Proof of concept of state A with *in silico*
simulated data and single-cell sequencing data 18
3.2 tITH showed comparable result with gITH 20
3.3 Relationship between nJSD and tumor purity 22
3.4 Clinical potentials of tITH 23
3.5 tITH detected clonal evolution in xenograft model . 25

Chapter 4 Discussion **37**

Bibliography **42**

Chapter 5 Appendix **50**
Appendix 1 50
Appendix 2 55
Appendix 3 58

요약 **64**

List of Figures

Figure 1.1 Pathway ambiguity model to analyze transcriptome-based ITH. The figure illustrates the heterogeneous tumor and its corresponding pathway status. While the clonal evolution produces different subclonal populations, pathway is getting ambiguous. Here, different clones are associated with their differentially activated pathway. In this context, measuring network perturbation by network entropy implies measuring pathway ambiguity. As ITH getting worse, the entropy of network increases. . . 10

Figure 2.1 nJSD calculation example. A network consist of 18 genes and 17 edges. (a) All 3 networks have the same gene expression level in total, but the activated paths are different. (b) An example of calculating JSD of the gene "g". nJSD of each network is calculated based on normal network. With gene expression of neighbors, interacting probability is defined and JSD is calculated. Mean JSD of all genes in network is nJSD. Detailed information can be found in the Method section. 16

Figure 2.2 Overproduction during clonal evolution made

ambiguous network status. (A) The consequence of clonal evolution is single tumor with heterogeneous population of cancer clones. The red shade which has smallest area on darwin's tree would be early cancer and it's sequencing result is represented as red circled 'T'. Orange one has larger area than red one, of course, orange one has more diverse population. Lime one has the most diverse population. We set a maximal state of black ambiguous like lime one, most diverse population of cancer clones, and measured tITH. (B) Network represents tumor with diver population. Distance between each state measured with nJSD, described in Method 17

Figure 3.1 Heterogeneous sample show ambiguous network state like state A. (A) is result of *in silico* mixed data with 675 human cancer cell lines. (B) is result of real bio-data from single-cell sequencing study. The bulk tumor sequencing data is more closer to state A than each of single-cell data in three different LUAD data set. "SC" represents single-cell data, "pt" represents patients derived tumor data and "pooled" represents pooled tumor cell data. Z-score test was performed. (C) Protein-interaction network of *TP53* gene and its neighbors of *in silico* mixed data, X2 and X256. black We highlighted top 20 genes in terms of difference between two conditions 29

Figure 3.2 Gene expression distribution of simulated data. X2

represents mixed data of couple cancer cell line data. X256 represents mixed data of 256 different cancer cell line data . 30

Figure 3.3 TCGA pan-cancer data and ITH. (A) Boxplot shows that the inter-tumor types differences of tITH distribution. (B) tITH and the number of subclones is positively correlated. 648 patients in 7 different cancer types are analyzed (R^2 = 0.24, p-value < 2.2e-16). 31

Figure 3.4 Comparison gITH and tITH in cancer type separately. The boxes with red line is about tITH and the boxes with black line is about gITH. In seven different cancer types, the trends between gITH and tITH is similar. This results is obtained from TCGA 648 patients . . 32

Figure 3.5 Inter-PIN Correlation of tTIH. We applied 3 different PIN on tITH analysis of TCGA pan-cancer cohort (n=6,320). Three PIN results are highly correlated with each others. 33

Figure 3.6 Pan-cancer survival analysis of gITH and tITH. We divided patients into two groups with median of gITH and tITH value. (A, B) was analyzed with same patients group who had reported number of subclones from other research. (A) Kaplan-Meier plot of the two groups based on the subclone number in 5-year censored data, and (B) based on tITH in 5-year censored data. (C) is Kaplan-Meier plot of pan-cancer patients in 12 different cancer types. 35

Figure 3.7 tITH during tumor evolution. (A) Original experimental design of the data. Single cancer cell makes 5 different subclones. (B) As diverging subclones, the tITH is increasing. (C) is pathway-tITH of 6 positively correlated KEGG pathway. This pathway getting promiscuous as diverging subclones. (D) is pathway-tITH of 6 negatively correlated KEGG pathways. Those pathways are converging to certain perturbed status during tumor progression and evolution. 36

List of Tables

Table 3.1 Relationship between tITH and tumor purity score from ESTIMATE. Individual cancer type comparison of Pearson's correlation coefficients. 34

Chapter 1

Introduction

1.1 Cancer genomics

1.1.1 Cellular Information

Most of cells have blueprint of themselves for functions and manufacturing daughter cells. Nucleus is the blueprint container and chromosomes are a bundles of blueprints. DNA contains series of cellular information, from functional genes to regulatory elements, which are crucial parts for cellular activities. To build something out of blueprint, it is necessary to deliver a copy of blueprint to a builder. The courier of cellular information is RNA. RNA transfer information from DNA to protein factory [1], ribosome, or sometimes it is functional itself like miRNA, tRNA or rRNA. Ribosome produces proteins, functional parts of cells, with RNA information from DNA.

At a larger scale, the proteins interact each other and these relationships determine cellular activity. Complex structure of proteins is required for specific functions. The information of interaction between two proteins are obtained from molecular biology assays, the

most well-known assay is 'Yeast two-hybrid system' assay [2]. A Series of reactions in metabolic cycles are mediated by a set of enzymes and those enzymes are composed of multiple genes and those interactions. Functional units in a specialized cellular activity are named as biological pathway.

Cellular activities are interpretable with massive amount of information from gene-level analysis to biological pathway-level analysis. The cellular activities can be read via biochemical experiments, cloning, gel-electrophoresis or blots, and genetical experiments with model organisms, *E. coli*, *S. cerevisiae*, and *D. melanogaster*. Recently developed high-throughput sequencing technology is risen as a sophiscated method to read cellular information via directly reading genomic and transcriptomic sequences [3].

1.1.2 High-throughput sequencing technology

High-throughput sequencing technology, also known as Next Generation Sequencing (NGS), make it possible to read systematic cellular information. As sequencing cost drops at very high speed during last decades, many variation of sequencing technology are developed [4]. The basic one is "Whole-Genome Sequencing (WGS)", which is the reading protocol for total genomic information of each of chromosome. "Whole-Exome Sequencing (WES)" is another variation of WGS, which reads only exonic regions on genome. Using both WGS and WES, it has been possible to read genomic variations, kind

of typos in our genetic blueprint, and its association with diseases [5].

For transciptomic information, RNA sequencing is widely utilized in researches. Overall protocol of RNA sequencing is similar with WGS and WES, except that reverse-transcription step for converting RNA to DNA is added. RNA sequencing replaced microarray technology in quantification of gene-expression profiles. By reading mRNA, we assume that those mRNA quantity is correlated with genes' activity. Obviously, quantified gene-expression level is not guaranteed to represent protein abundance in a cell, however, we can measure genes' activity by reading cellular information from RNA sequencing. This transcriptome information is invaluable, as we can find many copy of blueprint of specific parts to infer the main functions. With similar concept, information from RNA sequencing is used to analyze cell's function or to understand reaction mechanism caused by stimulus.

Many biological science applied sequencing technology to obtain systematic view about cellular landscape [3]. Developing reading method, sequencing, for biological information, it becomes more important to analyze sequencing data in terms of overall biological systems.

1.1.3 Current trends in cancer genomics

The improvement of analyzing method of sequencing data allowed scientists to organize transnational research groups as a consortium

to produce a massive amount of sequencing data in public for worldwide researchers. One of the most successful transnational research groups is “The Cancer Genome Atlas (TCGA)” [6]. TCGA produced and shared over ten thousand patients data in various cancer types with multi-platform omics information. From genomic sequencing to protein abundance, the multiple omics data help scientists seek novel insight in cancer [7].

Although big data of cancer facilitated us to show new discoveries with higher statistical power, simultaneously it showed follow-up hurdle to overcome for precision medicine. Such as tumor heterogeneity. Inter-tumor heterogeneity is a nature that each patient poses different phenotype and response to therapeutic drugs. Intrinsic subtype of breast cancer is the best example for inter-tumor heterogeneity [8]. Intra-tumor heterogeneity is another nature of cancer arising from clonal evolution of tumor cells. Those heterogeneities of tumor is unique character of cancer and one of the main obstacles to treat cancer patients.

1.2 Intratumor heterogeneity

Cancer has a complex system consisting of different cancer clones that interact with each other and also with normal cells, known as intratumor heterogeneity (ITH) [9]. The complexity from ITH is a major hurdle to understanding of the dynamics of cancer systems and also difficult to predict therapeutic outcomes [10].

Intratumor heterogeneity is the consequence of clonal evolution of a

single tumor [11]. One of the main cause of this ITH is genomic instability of cancer cells [12]. High-throughput sequencing technology is widely used to measure ITH at molecular level. A recent study revealed that diverse clones with different genomic signatures co-exist in a single tumor [13]. Diversity of clones give evolutional advantage in metastasis [14]. Additionally, diverse subclones are known to be under high pressure of natural selection in therapeutic circumstance and even cause therapeutic resistance [15, 16]. This clonal evolution during chemotherapy makes current target-drug therapy difficult [10, 17, 18]. However, there still remains an evolutionary issue about selective process during neoplasia, *i.e,* which daughter cells are selected and survive. To this issue, a colon cancer study suggested a big bang model without selective sweeps and a liver cancer study proposed non-darwinian evolution in tumor [19, 20]. Whether or not selective force being present, overproduction of subclones highly-likely results in ITH.

1.3 Recent genetic ITH inferences

Molecular level ITH has been identified with multi-regional sequencing [19, 21]. Although this multi-regional sequencing is at the forefront of ITH studies, single-cell genomics has emerged as the most credible technology [22]. Single-cell sequencing has an advantage on direct sequencing of each clone [23]. However, experimental cost of single-cell sequencing is too high for clinical applications. Thus, researchers have developed computational methods

to infer ITH with bulk-tumor sequencing data as an aggregated metadata of each clone's genomic information. In general, daughter cells carry exactly the same parental genomic information. However, their DNA replication system malfunctions, often in cancer, and leaves de novo mutational signatures, furthermore copy number alterations (CNA) and loss of heterozygosity (LOH) [24, 25]. Those genomic alterations remain from generation to generation, thus enabling the backtracing genomic signatures [21, 26]. On the same principle, inferring subclones from the genomic landscape of bulk tumor sequencing is a widely used strategy [27-29]. Computational methods, such as PyClone and EXPANDS, are current state-of-the-art tools that use mutational information to infer subclonal populations [30, 31]. Clinical relevance of inferred ITH was also highlighted in related to prognostic outcomes [32, 33].

Although the ITH inference based on genomic information were successful, there remain a few more issues that need further investigation. For example, a study reported that patients with a moderate number of subclones (3 or 4 clones) implicates a higher risk than more heterogeneous patients (above 4 clones) [32]. They discussed that there is a trade-off between the advantage of diversity and the cost of generating inviable daughter cells, however as mentioned earlier the selective sweep during cancer progression is still in questions. To understand better in tumor heterogeneity and clonal evolutionary process, we need to investigate three issues when genomic information is used for ITH prediction. First, it is a difficult

to define whether a somatic mutation as either a driver or a passenger mutation in terms of cancer genome evolution [34]. The study about neutral evolution of tumor proposed that driver mutation can be altered differently in a certain context [35]. As a result, inference of ITH with driver gene mutations may not reflect true subclonal population. Second, the mutational information alone is insufficient to identify cellular activities of subclones in cancer. Furthermore, cell plasticity needs to be considered in ITH since phenotype of cancer subclones can be altered without inheritable genomic variations [36]. A colon cancer study revealed that different phenotype can exist with no differences in genotype [37]. Lastly, cancer microenvironment is important in clonal evolution, tumor progression and metastasis [9, 38, 39]. According to current researches, different clonal activities and surrounding stromal and immune cells effects on cancer progression [40, 41]. This finding was also confirmed in a single-cell sequencing study [42]. However, mutational lineage analysis could only detect heterogeneity of cancer clones, not other effects from microenvironmental factors. Therefore, we believe that, in addition to the current DNA-based ITH inference, measuring ITH at the RNA level can provide a new insight on ITH and its clinical applications.

1.4 Motivation

1.4.1 Transcriptome based approach

To investigate the functional differences of heterogeneous clones, we

developed a method for ITH inference using RNA-sequencing data. There are two major reasons why RNA-sequencing data was used. First, RNA-sequencing data is ubiquitous as much as genomic data. Second, like mutations, transcriptome is also used in evolutionary studies [43, 44]. However, there is a challenge for analysis at the RNA level. Complex gene-gene dependency needs to be considered [45]. Thus, we used a biological network which is the most effective tool for modeling the complex gene-gene relationship - protein interaction network (PIN) and pathway information [46-48].

1.4.2 Entropy based metric for network analysis

Given a network, an effective metric is needed to quantify differences in the network perturbation to reflect both expression levels of individual genes and their relationships such as network topology and also co-expression of genes. We used an information theoretic approach to measure network state. This approach was successful in measuring network perturbations in terms of gene expressional changes [49-51]. This entropy measure was also effective for detecting network state transition from the normal state to the disease state [52, 53]. A recent application of the network entropy successfully showed the difference between primary tumor and metastatic tumor [54]. Additionally, Signaling entropy studies by 'Teschendorff group' identified relationship between network entropy and differentiation potential, additionally the prognostic importance [55-57].

1.4.3 Our work

Our hypothesis is that a heterogeneous tumor will have more ambiguity in network than a homogeneous one (Fig. 1.1). Thus, we developed a novel measurement of ITH with transcriptome data using information theory, network-based Jensen-Shannon Divergence (nJSD) [58]. Our approach was extensively tested for issues related with ITH. For proof of concept, we used human cancer cell line data and single cell sequencing data. Then, the pan-cancer cohort data was analyzed. Our result was in agreement with widely used genome-based ITH inference methods. Additionally, our approach was also tested for immune cell infiltration. Finally, analysis of mouse clonal evolution showed that our network perturbation inference was consistent with ITH at different clonal evolution stages.

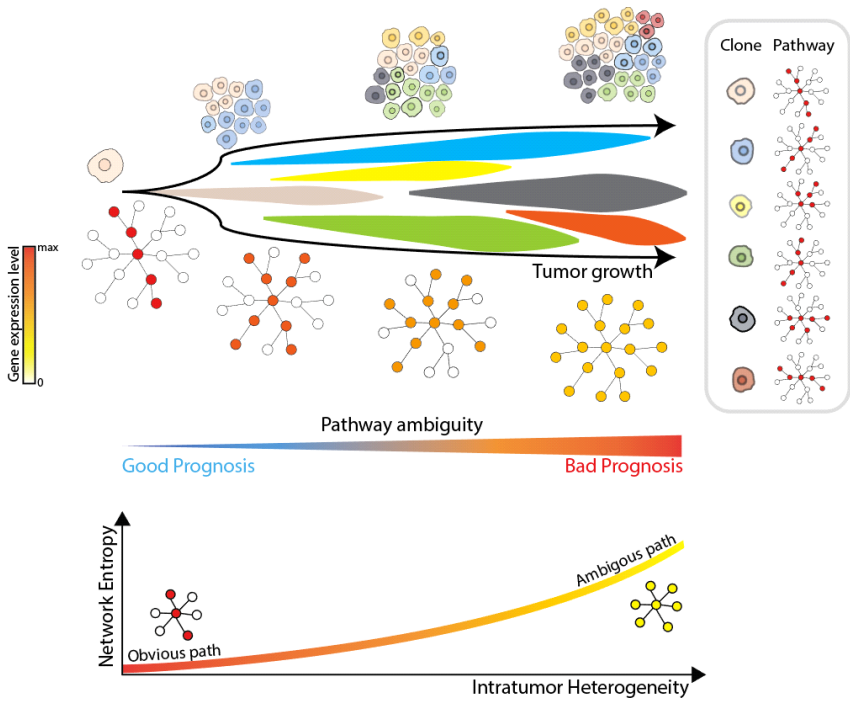


Figure 1.1 Pathway ambiguity model to analyze transcriptome-based ITH

The figure illustrates the heterogeneous tumor and its corresponding pathway status. While the clonal evolution produces different subclonal populations, pathway is getting ambiguous. Here, different clones are associated with their differentially activated pathway. In this context, measuring network perturbation by network entropy implies measuring pathway ambiguity. As ITH getting worse, the entropy of network increases.

Chapter 2

Materials and Methods

2.1 Gene Expression Data

We used four different datasets to examine the usefulness of nJSD. The lung adenocarcinoma (LUAD) single-cell RNA sequencing data from GEO under the accession number of GSE69405 [59]. Human cancer cell lines data, 675 different human origin cell lines, was obtained from GEO under the accession number of GSE30611 [60]. The xenograft-tumor data was obtained from GSE63630. The Ensembl gene ID was converted to gene symbol using 'mygene 2.3.0' python package. The log transformed gene expression data was downloaded from the supplementary data of the published research [61].

The pan-cancer data, TCGA RNA Seq V2, was obtained from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). We selected cancer types which have more than 10 normal samples: Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Head and Neck squamous cell

carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Prostate adenocarcinoma (PRAD), Thyroid carcinoma (THCA) and Uterine Corpus Endometrial Carcinoma (UCEC).

2.2 Pathway and Protein Interaction network

Protein interaction network was constructed with STRING v9 data [62]. Total 479,635 edges and 10,100 genes consist PIN. BioPlex and HINT were used in inter-PIN comparison study [63, 64]. BioPlex network has 10,963 genes and 56,554 edges and HINT has 12,194 genes and 53,126 edges. The KEGG pathway data contained 295 pathways and 6,969 genes.

2.3 *in silico* simulation with cancer cell lines data

We created *in silico* heterogeneous tumor data with gene-expression data of 675 human cancer cell lines. Randomly selected 2, 4, 8, 16, 32, 64, 128, 256, and 512 out of 675 cell line gene-expression data was individually averaged into a single gene-expression profile. Each simulation data had 1,000 gene-expression profiles.

2.4 tITH calculation of TCGA patients and comparison to gITH results

We calculated tITH and pathway-tITH based on mean expression

level of multiple normal samples of each cancer type. The intratumor heterogeneity, number of clones, in TCGA patients data were obtained from a previously published research [32]. This intratumor heterogeneity information was calculated based on mutations using state-of-the-art tools, PyClone and EXPANDs. The tumor purity information of TCGA patients is obtained from a previously published research [65]. This tool for tumor purity estimation, ESTIMATE, produces score about immune cell infiltration, stromal cell population and tumor purity. The TCGA pan-cancer clinical data were downloaded from TCGA data portal. Cox regression model analysis was done by using R library 'survival' [66].

2.5 Calculation of nJSD

Jensen-Shannon Divergence is the measure similar to Kullback-Leibler divergence with some modifications to make JSD symmetric and bounded in a finite value [67]. nJSD is the sum of entropy values measured at each of the genes in a protein interaction network. To define entropy of each gene, it is necessary to define a probability distribution using gene expression values. We used log2-normalized gene expression values and assumed that the protein interactions were under the law of mass action.

Let e_i denotes the expression level of gene-i and a set of neighbor genes of gene-i is J_i. Then, a probability of interaction between two genes is defined as

$$p_{ij} = \begin{cases} \frac{e_j}{\sum_{l \in J_i} e_l}, & \text{if } j \in J_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

With pi j, we defined a probability distribution of gene-i which has $\{1, 2, \cdots, j, \cdots, n\}$ neighbors on PIN on a sample X.

$$PD_i(X) = \{p_{i1}, \cdots, p_{ij}, \cdots, p_{in}\} \quad (2)$$

Let the l^{th}-element in the probability distribution $PD_i(X)$ be $PD_i(X)_l$, then the Kullback-Liebler Divergence of gene-i between normal (N) and tumor (T) is defined [68].

$$KLD_i(PD_i(T)||PD_i(N)) = \sum_{l=1}^{n} PD_i(T)_l \log \frac{PD_i(T)_l}{PD_i(N)_l} \quad (3)$$

Then the JSD of gene-i between normal and tumor was defined as

$$\begin{aligned} &JSD_i(PD_i(N)||PD_i(T)) \\ &\quad = \frac{1}{2} KLD_i(PD_i(N)||PD_i(M)) \\ &\quad + \frac{1}{2} KLD_i(PD_i(T)||PD_i(M)) \end{aligned} \quad (4)$$

where $PD_i(M) = \frac{1}{2}(PD_i(N) + PD_i(T))$.

Finally, nJSD was defined as an average JSD of all genes. Graphical example of calculation of nJSD was described in (Fig. 2.1).

2.6 Calculation of transcriptome-based ITH

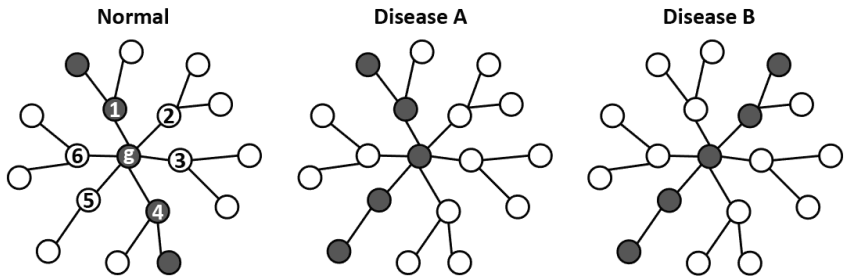
To define transcriptome-based ITH (tITH), we set a maximally

ambiguous network where whole gene-expression values were equal. nJSD was applied as a distance measure between two network states. Here, we defined tITH with two distance values, distance from normal data to cancer data (NT) and distance from cancer data to maximally ambiguous network (TA) (Fig. 2.2). This distance based approach was inspired by recent study about cancer evolution that described embryonic stem cell as cancer evolutionary destination [69]. Combining NT and TA into a single metric, we defined the transcriptome-based ITH and we named it as tITH in comparison with genomic ITH (gITH).

$$tITH = \frac{NT}{NT + TA} \tag{5}$$

To investigate tITH at the pathway level, pathway-tITH was defined using only the gene set in a specific pathway. With this metric, we were able to precisely quantify pathway perturbation value ranging from 0 to 1.

A



B

	Normal	Disease A	Disease B
Gene Expression of neighbor of "g"	{ 4, 1, 1, 4, 1, 1 }	{ 4, 1, 1, 1, 4, 1 }	{ 1, 4, 1, 1, 4, 1 }
Probability Distribution of "g"	$\{\frac{4}{12},\frac{1}{12},\frac{1}{12},\frac{4}{12},\frac{1}{12},\frac{1}{12}\}$	$\{\frac{4}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{4}{12},\frac{1}{12}\}$	$\{\frac{1}{12},\frac{4}{12},\frac{1}{12},\frac{1}{12},\frac{4}{12},\frac{1}{12}\}$
	Normal - Normal	**Normal - Disease A**	**Normal - Disease B**
JSD of "g"	0.0	0.08	0.16
Average JSD of all genes (nJSD)	0.0	0.010	0.020

Figure 2.1 nJSD calculation example

A network consist of 18 genes and 17 edges. (A) All 3 networks have the same gene expression level in total, but the activated paths are different. (B) An example of calculating JSD of the gene "g". nJSD of each network is calculated based on normal network. With gene expression of neighbors, interacting probability is defined and JSD is calculated. Mean JSD of all genes in network is nJSD. Detailed information can be found in the Method section

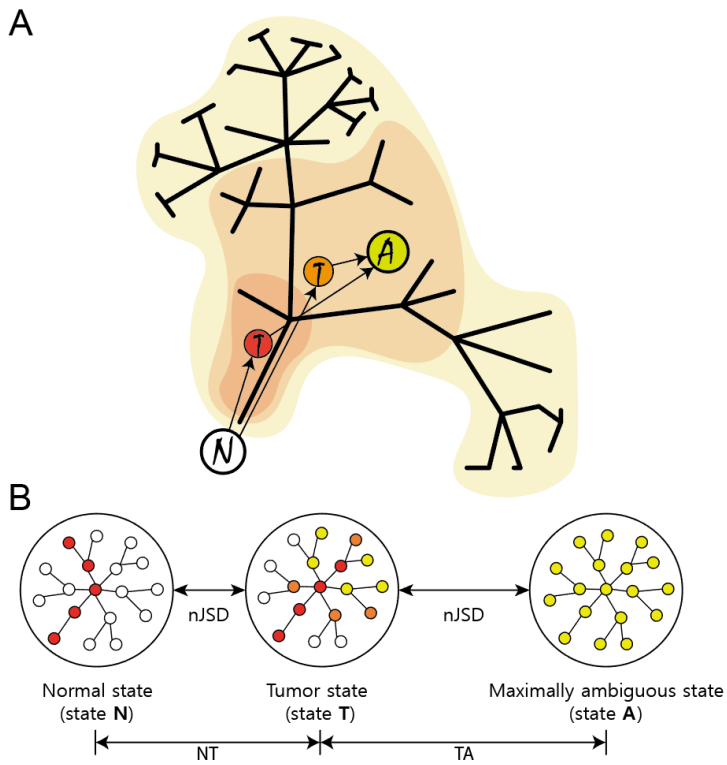


Figure 2.2 Overproduction during clonal evolution made ambiguous network status

(A) The consequence of clonal evolution is single tumor with heterogeneous population of cancer clones. The red shade which has smallest area on darwin's tree would be early cancer and it's sequencing result is represented as red circled 'T'. Orange one has larger area than red one, of course, orange one has more diverse population. Lime one has the most diverse population. We set a maximal state of black ambiguous like lime one, most diverse population of cancer clones, and measured tITH. (B) Network represents tumor with diver population. Distance between each state measured with nJSD, described in Method.

Chapter 3

Results

3.1 Proof of concept of state A with *in silico* simulated data and single-cell sequencing data

To calculate tITH, we assumed a maximally ambiguous state (state A). We made *in silico* data to investigate the relationship between sample heterogeneity and network ambiguity using single-cell sequencing data.

Each of cancer cell lines of different molecular characteristics such as drug resistance was cultured in well-controlled environments, so as to consider each cell line as unique clone in heterogeneous tumor. With 675 human cancer cell lines, we created heterogeneous tumor gene-expression data and calculated distance to state A by nJSD. The distance to state A decreased as more cell lines were mixed (Fig. 3.1A). This indicates that data from bulk-cell sequencing of tumor with diverse clones will show ambiguous network state similar to the state A. The difference in network status can be easily identified by visualizing the networks with gene expression values

(Fig. 3.1C). In a plot of the gene expression value *vs.* the number of genes at a certain expression level, shape of the line in the cell line mixed data shifting to the one in the state A from the skewed shape in individual cell lines (Fig. 3.2).

The relationship between sample heterogeneity and network ambiguity was re-examined with single-cell sequencing data as a real biological data. Diverse mutational patterns among different clones could contribute to differentially activated paths in network across clones, thus we expect that the bulk-cell tumor data will show more ambiguous network status than each of single-cell data. The LUAD data set consists of 3 different experimental sets (H358, LCT-PT-45 and LCT-PT-45Re with 35, 44 and 50 single-cell sequencing, respectively and additional pooled cell sequencing data) [59]. We compared patient derived bulk tumor and pooled sample with its single-cell sequencing data in three different LUAD data set. nJSD of 10,051 genes were calculated and compared in terms of distance from state A. Patient derived bulk tumor, "pt", was significantly closer to state A than single cells in TS-45 data (Z-score = -4.01, p-value = 0.00003; Fig. 3.1B). Pooled sample also had lower nJSD to state A than single cells in H358, LCT-PT-45 and LCT-PT-45Re data (Z-score = -5.27 and p-value < 0.00001, Z-score = -3.75 and p-value = 0.000087 and Z-score = -3.08 and p-value = 0.001042).

With *in silico* simulated data and single-cell sequencing data, it was possible to identify that the heterogeneous sample had

ambiguous network state than homogeneous one. Our tITH measurement would showed how their gene-expression profile changed from normal tissue to maximally ambiguous state using nJSD.

3.2 tITH showed comparable result with gITH

This experiment is to show how well tITH agrees with gITH, genomic information based ITH, using the pan-cancer data set from TCGA. To compare tITH with gITH inferred clonal information was obtained from the study using PyClone and EXPAND [32]. Since different mutational pattern will influence network perturbations genetic heterogeneity ought to be associated with tITH [70, 71].

tITH of pan-cancer cohort, 5,630 patients, showed intratumor heterogeneity of each patient. Among the tumor types, there were inter-tumor type differences in the distribution of nJSD (Fig. 3.3A). Notably, this inter-tumor type difference of tITH has the similar tendency to that of intratumor genetic ITH study [32]. THCA, PRAD and KIRC showed less heterogeneity than LUAD, HNSC, BLCA and LUSC (Fig. 3.4). Next, we compared gITH results with our tITH result and found a positive correlation between genetic heterogeneity and tITH (Fig. 3.3B).

When we analyzed the cancer types separately, 4 cancer types showed similar patterns with the pan-cancer result (p-value < 0.05), HNSC (162 patients, $r = 0.20$), KIRC (64 patients, $r = 0.27$), LUAD (76 patients, $r = 0.28$) and LUSC (84 patients, $r = 0.35$), but other 3

cancer types had no relationship (p-value > 0.05), BLCA (111 patients, r = 0.13), PRAD (93 patients, r = 0.14), THCA (60 patients, r = -0.04). Pan-cancer trends were similar between tITH and gITH. However, there were some cancer types has weak correlation between tITH and gITH. This might due to the effect of cancer microenvironments, inter-tumor type difference and also small size of data set. In order to make sure that the pan-cancer result is not dependent on specific topology of PIN, we performed analysis using other PINs such as BioPlex and HINT [63, 64]. Analysis of inter-PIN correlation showed that tITH was not dependent on a specific PIN topology (Fig. 3.5).

In all 12 cancer data sets, the greater the number of subclones, the greater tITH values. Thus we were curious to know whether this correlation is a global trends in many different pathways in cellular mechanism. Using pathway-tITH, an average of nJSD values of genes in a pathway, we identified the pathways that pathway-tITH was correlated gITH. Most of the pathways, 255 out of 291, were significant in terms of representing ITH at the p-value < 0:001 (Pearson's correlation test). Among the pathways identified by the pathway-tITH analysis, cell cycle and central dogma related pathways were highly correlated with the number of clones obtained from genomic ITH inference (Appendix 1). This is consistent with the previous reports using genomic information that focused on variations of driver genes in cell cycle and central dogma related pathways [30, 72]. Additionally, histological study reported cell

cycle marker is correlated with gITH [32].

3.3 Relationship between nJSD and tumor purity

Gene expression data was used to estimate tumor purity and immune cell infiltrations [65, 73, 74]. Tumor purity is a measurement of sample contamination from other cell types, and immune cell infiltration is a score of immune cell proportion in a tumor sample. For tumor purity information, we used three scores – stromal score, immune score, and tumor purity score – by ESTIMATE in 1,557 patients across 8 different cancer types [65].

We compared the tITH of each patient with the stromal score, immune score and the tumor purity score of the above mentioned methods. The stromal score was negatively associated with tITH ($r = -0.502$, p-value $< 2.2e-16$, Pearson's correlation test). The immune score had a weak negative association ($r = -0.203$, p-value = 5.08e-16). The increasing proportion of specific cell types, such as stromal cells and immune cells, led to decreased tITH as expected. The tumor purity score was associated with tITH ($r = 0.288$, p-value $< 2.2e-16$), which is intuitive that the purer tumor will contain the more diverse clones. When analyzing cancer types separately, stromal scores had consistently negative correlation with tITH (Table 1). Immune score had negative association with tITH, and this result was consistent with the findings in a previous study [33]. Only the kidney cancer showed different patterns in tumor purity and immune score. This inverse pattern with other cancer types might be due to

the higher immune cytolytic activity in kidney cancer [75].

3.4 Clinical potentials of tITH

Next, we were further curious if tITH has a prognostic power. We used the pan-cancer data comes with clinical information such as patient survival. We investigated the clinical utility of tITH in three different ways: survival analysis at the whole cancer cell level, pathway level, and analysis at the effect of immune cell filtration perspective.

The first experiment was to investigate the prognostic power of tITH at whole cancer level (n = 606) using the Cox regression model, in comparison to that of the gITH information predicted by EXPAND. To test the prognostic power of gITH and tITH for each cancer type, we built two univariate Cox models. One was done with gITH and 5-year survival information.

The other was done with tITH and 5-year survival information. The univariate Cox model using gITH was not significant (p-value = 0.3370, c-index = 0.543) while the model using tITH was statistically significant (p-value = 0.0006, c-index = 0.604). It is reported that the mutation based subclone number has a nonlinear association with survival [32], which supports why the gITH model was not successful in separating patients groups into good or poor groups in Kaplan-Meier survival analysis. However, tITH has linear association with survival and tITH model was successful in separating patients groups (Fig. 3.6A, B). Next, we performed cox proportional hazard

test with a bigger data set including patients without subclone information, a total of 5,628 patients. Pan-cancer univariate cox model using tITH values found clinical utility with significant statistics (c-index=0.64, p-value < 2e-16). Kaplan-Meier survival analysis with poor and good group also shows significant separating of two groups (Fig. 3.6C).

The second experiment was to identify pathways that are relevant to or useful for the patient survival prediction. We analyzed pathway-tITH values measured for each pathway and cox proportional hazard test were done for each of the KEGG pathways. If cox model for a pathway was significant in terms of p-value < 0.001, then the pathway was selected as one that is significant for prognosis. A list of pathways that may have prognostic power is listed (Appendix 2). In particular, mRNA surveillance pathway that controls mRNA abundance has the greatest c-index of 0.63. Additionally, Ribosome biogenesis in eukaryotes, Olfactory transduction, and RNA transport pathways also showed good prognostic power.

The third experiment was performed using two scores - immune score, and tumor purity score - from ESTIMATE result (n=1,558). Previous results showed a negative correlation between tITH and immune score (Table 1). Because, recently prognostic importance of immune related cells was reported, we wanted to test that does immune score has dominant effect on clinical utility of tITH [76]. Following the gITH study reported independency of gITH from

immune cell infiltration, we reproduced independency of tITH excluding effect of the immune cell infiltration [33]. We performed two variable cox proportional hazard test with both the immune score and tITH. The immune score showed weak significance (p-value = 0.52101), while tITH had significance statistics (p-value = 0.00018). Therefore, it seems that the prognostic power is more likely from heterogeneity of tumor, rather than dominant effect of immune cell proportion. When a tumor purity score from ESTIMATE was applied as covariate with tITH, cox model was improved (c-index of tITH univariate cox model : 0.568→tITH + purity score : 0.604).

3.5 tITH detected clonal evolution in xenograft model

We have shown that tITH can effectively measure the tumor heterogeneity from cohort data. Now we investigate whether tITH can measure the clonal evolution during tumor progression. Thus, we analyzed a time-series data of xenografted tumor by tITH in the extended time scale of tumor growth and divergence of subclones.

During the tumor progression from single-cell clone to metastatic tumor, the bulk-tumor sequencing data (12 time points) revealed the emergence of different clonal populations [61]. We measured tITH of 11 time point data with first time point data (MCF10A), MCF10A-*HRAS*, XT1-XT8, M1 and M2, and compared with reported number of clones from original research (Fig. 3.7A). The perturbed network status of MCF10A-*HRAS* was closer to the network of

MCF10A than the xenografted tumor samples (XT1-XT8, M1 and M2). However, we observed the abrupt elevation of tITH between MCF10A-*HRAS* and XT1 (from 0.109 to 0.346). The environmental change from the culture plate to the mouse system was the first big evolutional force which might result in the dramatic network perturbation change [77]. After the xenograft, the clones continuously produced different lineages and evolved. The authors of the original research reported that the very first clone generated 5 major subclonal populations, based on mutational lineage analysis. We were able to observe the increase in tITH during that clonal divergence. The divergence of major clones occurred at MCF10A-*HRAS*→XT1, XT2→XT3, XT5→XT6, and XT8→M. tITH values steadily increased as the number of subclones increased (Fig. 3.7B). tITH increased at MCF10A-*HRAS*→XT1 (from 0.109 to 0.346), XT2→XT3 (from 0.349 to 0.350), XT5→XT6 (from 0.349 to 0.356) and XT8→M (from 0.358 to 0.384 at M1, and 0.371 at M2).

Again, we performed the pathway-tITH analysis to further explore the relationship between clonal diversity and tITH. The correlation analysis showed that a number of pathways were either positively or negatively correlated with the number of subclones. Among those pathways, metabolic pathways were highly ranked in the correlation analysis. The metabolic pathway is well known to be important in cancer mechanism [78, 79]. For example, Oxidative phosphorylation is involved in metabolic reprogramming in cancer cells [80] and its heterogeneity is also observed clear in our correlation analysis (Fig.

3.7C). The whole pathway list is in (Appendix 3). As there are many kinds of metabolic pathways related to cancer, we focus on other pathways, excluding metabolic pathways for further pathway-tITH analysis.

We found that, in 131 out of 291 pathways, pathway-tITH values were positively correlated with tumor progression over time ($r > 0.3$, from XT1 to XT8-M2). Cell cycle and central dogma related pathways such as Parkinson's disease, Ribosome, mRNA surveillance pathway, Cell cycle, and DNA replication were at the top of the positively correlated pathway list in *HRAS* mutated cell lines (Fig. 3.7C). This finding was confirmed in a study that reports the central dogma and cell cycle related pathways became more heterogeneous as the mutated *HRAS* activated MAP kinase cascades downstream and effects on transcriptional control and cell growth [81]. This result implies that ITH is highly related with the aberration in the flow of cellular information and cell cycle transition from the quiescent stem-cell like state to accelerated proliferative state. Like Cell cycle pathway, Parkinson's disease pathway became more heterogeneous as clones diverged. Although the pathway is a kind of brain disease, it contained many genes related to cell cycle [82]. Recent cohort study revealed the relationship between parkinson's disease and cancer [83]. Also in molecular-level studies, the *PARK2* and *LRRK2* genes well known in parkinson's disease were revealed that those genes were related with cell cycle pathways [84, 85].

There were 60 of 291 pathways negatively correlated with tumor

evolution measured by pathway-tITH ($r < -0.3$). The tITH values of these pathways were steady or decreased as the number of subclones increased (Fig. 3.7D). Thus, we conjecture that these pathways were not affected from ITH but the perturbation at the early stage of carcinogenesis remained and converged among different clones. Pathway-tITH of six pathways had a negative association with the number of subclones, especially Fanconi anemia pathway showed dramatic decrease (Fig. 6D). This indicates that Fanconi anemia pathway related to DNA repair system was heterogeneous in early time points but converged for some reasons such as the influence of the host system or the process of clonal evolution (Fig. 3.7D). In the original study, missense mutations on *RAD54B* and *PMS1* were reported, and they were connected by direct edges to Fanconni anemia pathway in STRING PIN (*PMS1-FAN1* and *RAD51-RAD54B*) [61]. This pathway is a genetic disease about DNA repair genes - *BRCA1*, *RAD51*, *PMS2* and FANC proteins - which are highly related to cancers. Like Fanconi anemia pathway, some pathways, such as Hedgehog signaling pathway, NF kappa B signaling pathway, Adherens junction and immune related pathways, were converged to a certain state of cancer during tumor growth ($r < -0.6$ from XT1 to XT8-M2).

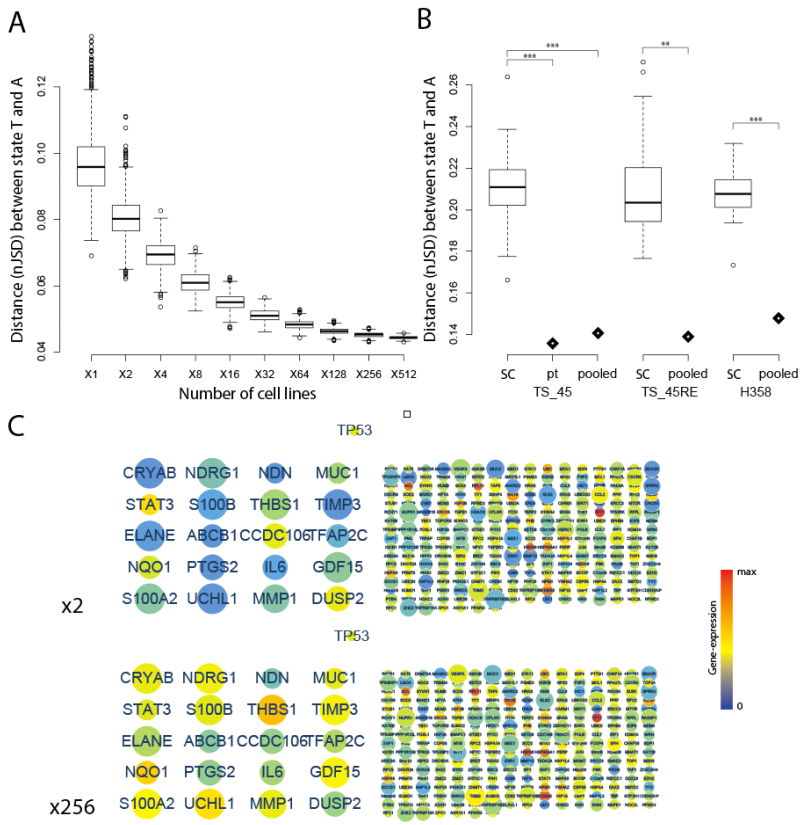


Figure 3.1 Heterogeneous sample show ambiguous network state like state A

(A) is result of *in silico* mixed data with 675 human cancer cell lines. (B) is result of real bio-data from single-cell sequencing study. The bulk tumor sequencing data is more closer to state A than each of single-cell data in three different LUAD data set. "SC" represents single-cell data, "pt" represents patients derived tumor data and "pooled" represents pooled tumor cell data. Z-score test was performed. (C) Protein-interaction network of *TP53* gene and its neighbors of *in silico* mixed data, X2 and X256. black We highlighted top 20 genes in terms of difference between two conditions.

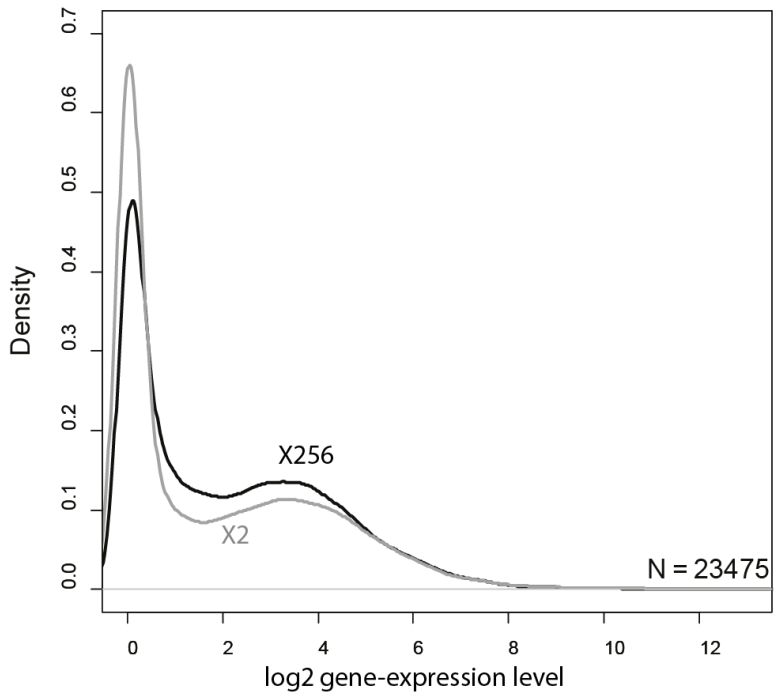


Figure 3.2 Gene expression distribution of simulated data

X2 represents mixed data of couple cancer cell line data. X256 represents mixed data of 256 different cancer cell line data.

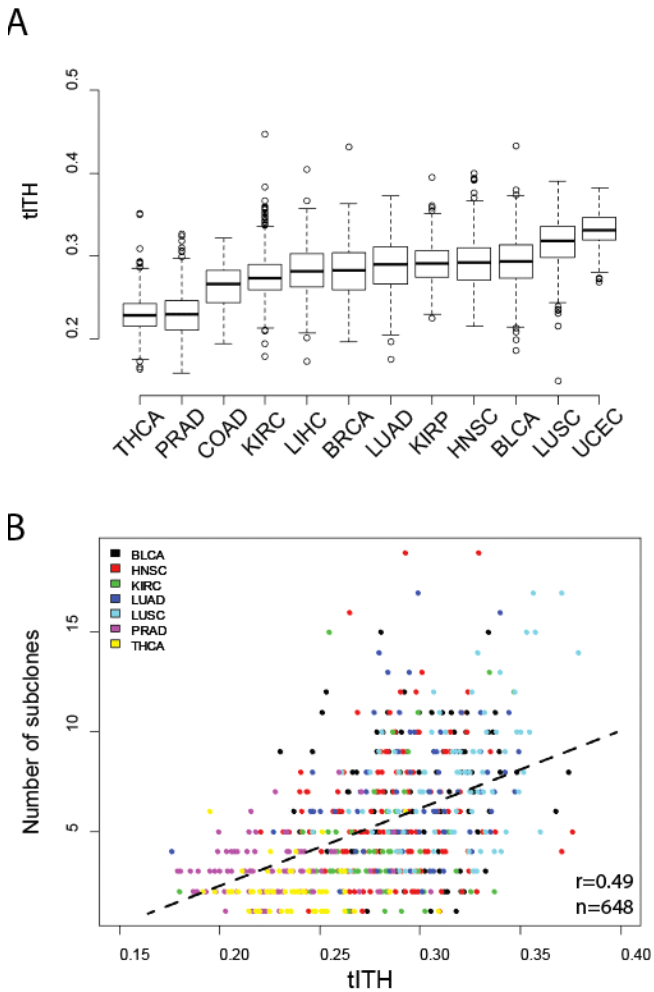


Figure 3.3 TCGA pan-cancer data and ITH

(A) Boxplot shows that the inter-tumor types differences of tITH distribution. (B) tITH and the number of subclones is positively correlated. 648 patients in 7 different cancer types are analyzed (R^2 = 0.24, p-value < 2.2e-16).

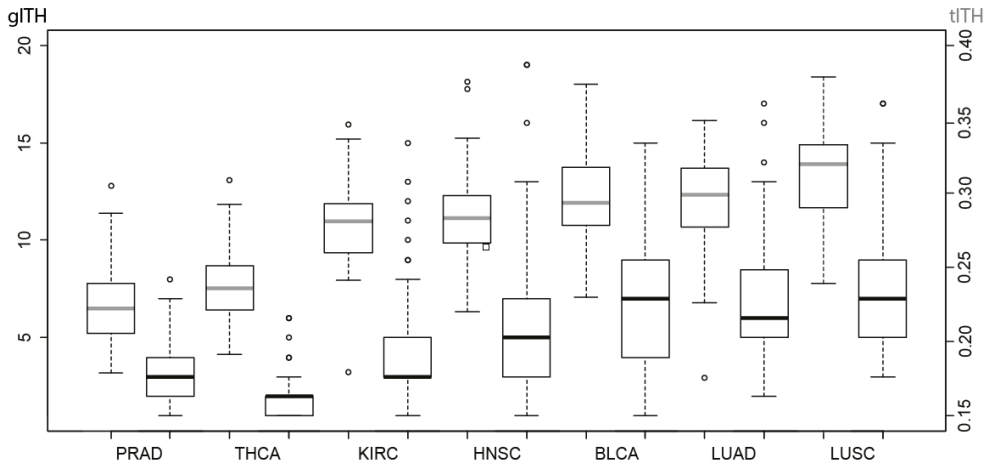


Figure 3.4 Comparison gITH and tITH in cancer type separately

The boxes with red line is about tITH and the boxes with black line is about gITH. In seven different cancer types, the trends between gITH and tITH is similar. This results is obtained from TCGA 648 patients

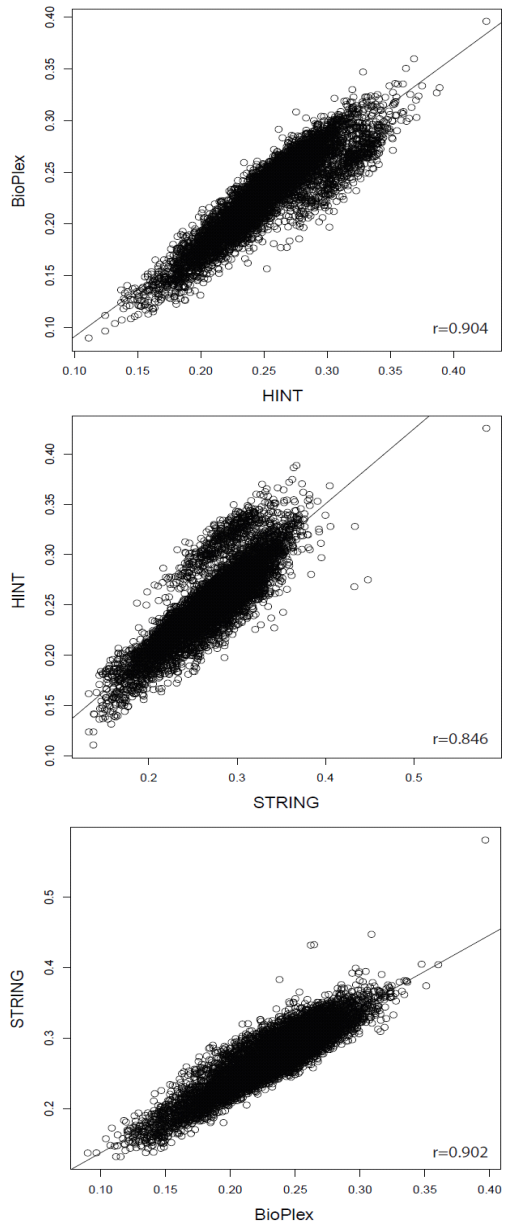


Figure 3.5 Inter-PIN Correlation of tTIH

We applied 3 different PIN on tITH analysis of TCGA pan-cancer cohort (n=6,320). Three PIN results are highly correlated with each others.

	BLCA	BRCA	COAD	HNSC	LUAD	LUSC	KIRC
Patients (n)	95	471	18	291	228	129	326
Purity (r)	0.330	0.417	0.466	0.459	0.230	0.505	0.074
Immune Score (r)	-0.389	-0.184	-0.481	-0.251	-0.382	-0.487	0.177
Stromal Score (r)	-0.575	-0.585	-0.616	-0.515	-0.384	-0.476	-0.342

Table 3.1 Relationship between tITH and tumor purity score from ESTIMATE

Individual cancer type comparison of Pearson's correlation coefficients.

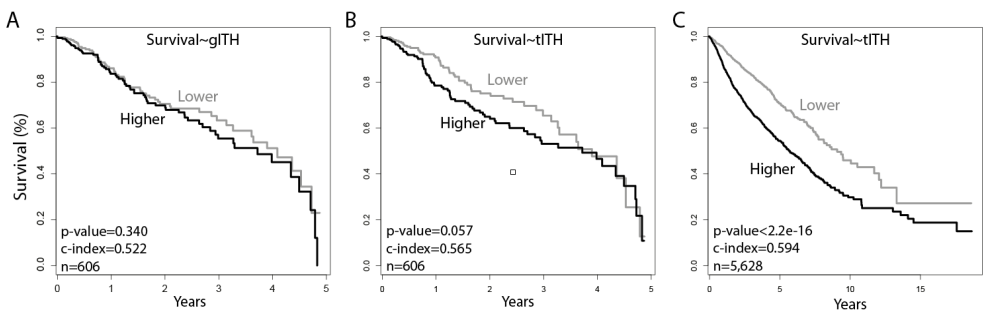


Figure 3.6 Pan-cancer survival analysis of gITH and tITH

We divided patients into two groups with median of gITH and tITH value. (A, B) was analyzed with same patients group who had reported number of subclones from other research. (A) Kaplan-Meier plot of the two groups based on the subclone number in 5-year censored data, and (B) based on tITH in 5-year censored data. (C) is Kaplan-Meier plot of pan-cancer patients in 12 different cancer types.

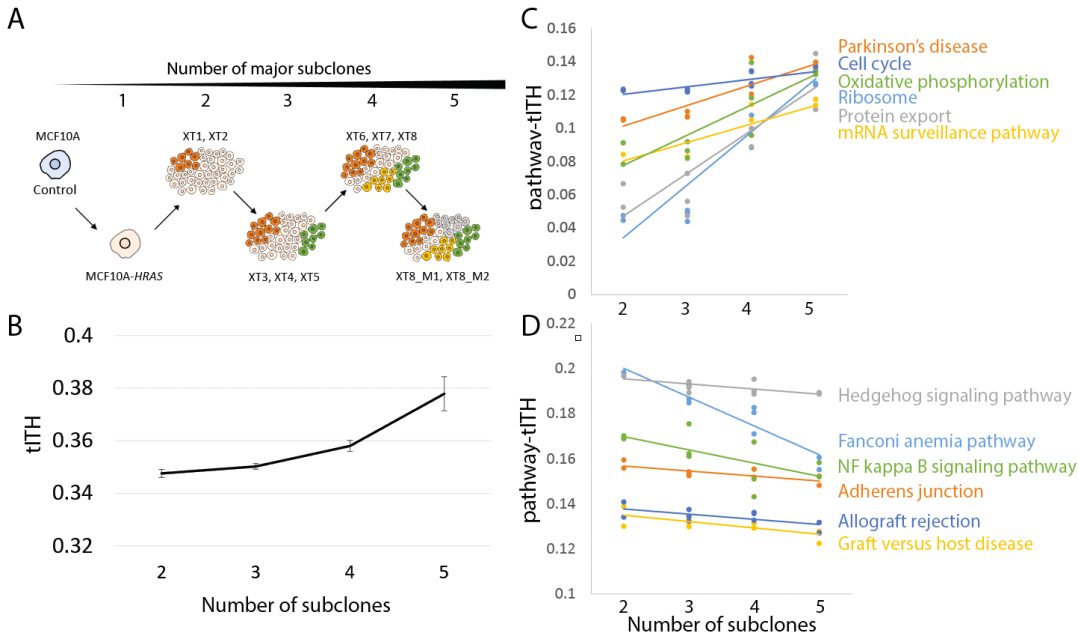


Figure 3.7 tITH during tumor evolution

(A) Original experimental design of the data. Single cancer cell makes 5 different subclones. (B) As diverging subclones, the tITH is increasing. (C) is pathway-tITH of 6 positively correlated KEGG pathway. This pathway getting promiscuous as diverging subclones. (D) is pathway-tITH of 6 negatively correlated KEGG pathways. Those pathways are converging to certain perturbed status during tumor progression and evolution.

Chapter 4

Discussion

Cancer evolution has become an important issue in understanding cancer biological mechanisms. An cancer evolution study by He *et. al.* [69], using JSD as a distance measure, reported that embryonic stem cell is the destination in cancer evolution. Their finding is that unicellularity is key characteristic of cancer [86-88]. Accordingly, the relationship between ESC and cancer was well studied, but the clinical application is yet to be a reality [89, 90]. In this respect, 'Teschendorff group' showed a possibility of clinical application in terms of cancer evolution [57]. Their study focused on signaling pathways, or regimes in their term, and observed the reverse differentiation of cancer cells [56]. The study used a network perturbation concept in terms of entropy and reported the association between differentiation potential and network entropy, thus they defined the measure termed as Signaling entropy. This way, the study showed the clinical importance of the signaling entropy in cancer, and reported the relationship with ITH [57]. The study

discussed about a possibility of measuring ITH using the network entropy.

However, signaling entropy was difficult to distinguish differentiation potential from ITH because the network measure is the metric of entire PIN. On the contrary, our entropy measure is gene centric and then combines gene level information to pathway level and also to the entire PIN level. We used the metric to measure ITH in terms of subclone diversity, rather than focusing on the differential potential.

Our analysis results in Fig. 4B with the tITH approach in the xenograft tumor evolution data can be interpreted as an increment of differentiation potential in terms of the reverse-evolution hypothesis of the previous works. To investigate further, we analyzed heterogeneous data by incrementally adding transcriptome data form different cancer cell lines (Fig. 3). Differentiation potentials in different cell lines may not be significantly different, thus the increased network entropy observed in our study with the mixed cell line data may be from other factors, possibly from some pathways, rather than differentiation related pathways. This hypothesis was supported by the analysis of the clonal evolution data. tITH values of differentiation potential related pathways were either steady or slightly decreased (Fig. 6D). This implies that cancer clones lose their control of differentiation at the early carcinogenesis stage. After tumorigenesis, continuous increment in tITH values may be the result from heterogeneity in cell cycle and central dogma related pathways. This conjecture is supported by the previous work with histological

information [32].

Although our study highlights the importance of cell cycle pathways in terms of tumor heterogeneity, reverse differentiation is well documented and important factor in cancer evolution.77 In cellular mechanism, multiple signaling pathways work as control switch between differentiation and proliferation state [88]. In our analysis, tITH value of the Hedgehog signaling pathway, an important information transmitor during embryogenesis [91], remains unchanged for different number of subclones, while cell cycle related pathways become more ambiguous. However, it is still unclear how the loss of differentiation and the accelerated cell proliferation interplay. Clones may have loss of differentiation because of the fast and uncontrolled cell cycle, however simultaneously cancer stem cell population did not differentiate even they were quiescent state of cell cycle [93, 94]. The aberration of the master regulator of differentiation and proliferation – like hedgehog signaling that we found – may be the main cause of dysregulation of differentiation. A breakthrough in cancer may be in there [87].

Our method successfully measured ITH with transcriptome data and network information, but our method did not use whole features of transcriptome data. Biological network information includes only a small number of genes; 20,000~40,000 transcripts are generally observed in the whole transcriptome data but only ~10,000 genes are used in the interaction data. There could be another approaches with de novo network construction with edge probability using statistical

approaches [95, 96]. As more comprehensive interaction data, including regulatory data, is available, our method can be more accurate in predicting ITH.

Although our study was able to show tumor heterogeneity using four datasets, our computational methods still need to be improved. Our approach focused on detecting ITH with bulk RNA-sequencing data. There are other notable methods which deconvolute gene-expression data to identify population of specific cell types [97]. Especially, immune related cell population in tumor sample were well studied [39, 76]. These methods, although successful in de-composing cell populations, require a reference gene expression profile. The requirement for a reference gene expression profile makes difficult to measure ITH since the number of clones is not known; this is a typical chicken and egg problem. Our current method is to measure the heterogeneity in a systematic view but it is not designed to de-compose cancer clones. As a future study, we are working on a computational method that can both de-compose clones and measure the status of heterogeneity.

We propose a new approach, tITH, to inference ITH using RNA-seq data by nJSD and compared with gITH. Our tITH was in agreement with gITH. Since our method is to measure the status of gene expression, it is possible to perform functional or pathway-level analysis. With xenograft model, we found importance of cell cycle related pathways in ITH. Other signaling pathway showed converging tendency during clonal evolutions. In addition, we showed that tITH

achieved better performance than gITH in cox regression model analysis for survival prediction. We believe that ITH should be investigated at the full spectrum of the central dogma, *i.e.*, at DNA, RNA, and protein levels. Our tITH can be useful for ITH inference using RNA-sequencing data of the bulk tumor, which may be useful for developing cost effective molecular diagnosis methods.

Bibliography

1. Crick, F. et al. Central dogma of molecular biology. Nature 227, 561 – 563 (1970).
2. Fields, S. & Song, O.-k. A novel genetic system to detect protein protein interactions (1989).
3. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. Molecular cell 58, 586 – 597 (2015).
4. Wetterstrand K. A., DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), Available at: www.genome.gov/sequencingcostsdata. Accessed [Nov. 2016].
5. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. New England Journal of Medicine 363, 166 – 176 (2010).
6. Tomczak, K., Czerwinska, P., Wiznerowicz, M. et al. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemp Oncol (Pozn) 19, A68 – A77 (2015).
7. Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158, 929 – 944 (2014).
8. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology 27, 1160 – 1167 (2009).
9. Tabassum, D. P. & Polyak, K. Tumorigenesis: it takes a village. Nature Reviews Cancer 15, 473 – 483 (2015).
10. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer cell 27, 15 – 26 (2015).
11. Nowell, P. C. The clonal evolution of tumor cell populations. Science 194, 23 – 28 (1976).

12. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501, 338–345 (2013).
13. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nature biotechnology 34, 155–163 (2016).
14. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. Nature 520, 353–357 (2015).
15. Turner, N. C. & Reis-Filho, J. S. Genetic heterogeneity and cancer drug resistance. The lancet oncology 13, e178–e185 (2012).
16. Greaves, M. & Maley, C. C. Clonal evolution in cancer. Nature 481, 306–313 (2012).
17. Bozic, I. et al. Evolutionary dynamics of cancer in response to targeted combination therapy. Elife 2, e00747 (2013).
18. Almendro, V. et al. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. Cell reports 6, 514–527 (2014).
19. Sottoriva, A. et al. A big bang model of human colorectal tumor growth. Nature genetics 47, 209–216 (2015).
20. Ling, S. et al. Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. Proceedings of the National Academy of Sciences 112, E6496–E6505 (2015).
21. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England journal of medicine 366, 883–892 (2012).
22. Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. Molecular cell 58, 598–609 (2015).
23. Patel, A. P. et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014).
24. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. Genetics 148, 1667–1686 (1998).
25. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells.

Nature genetics 47, 1402 – 1407 (2015).

26. Campbell, P. J. et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proceedings of the National Academy of Sciences 105, 13081 – 13086 (2008).
27. Carter, S. L. et al. Absolute quantification of somatic dna alterations in human cancer. Nature biotechnology 30, 413 – 421 (2012).
28. Ha, G. et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome research 24, 1881 – 1893 (2014).
29. Shen, R. & Seshan, V. E. Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing. Nucleic acids research gkw520 (2016).
30. Roth, A. et al. Pyclone: statistical inference of clonal population structure in cancer. Nature methods 11, 396 – 398 (2014).
31. Andor, N., Harness, J. V., Mueller, S., Mewes, H. W. & Petritsch, C. Expands: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics 30, 50 – 60 (2014).
32. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature medicine 22, 105 – 113 (2016).
33. Morris, L. et al. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. Oncotarget 7, 10051 – 10063 (2016).
34. Cheng, F. et al. A gene gravity model for the evolution of cancer genomes: a study of 3,000 cancer genomes across 9 cancer types. PLoS Comput Biol 11, e1004497 (2015).
35. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. Nature genetics (2016).
36. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. Nature 501, 328 – 337 (2013)
37. Kreso, A. et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science 339, 543 – 548 (2013).

38. Marusyk, A. et al. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature 514, 54 - 58 (2014).
39. Kitamura, T., Qian, B.-Z. & Pollard, J. W. Immune cell promotion of metastasis. Nature Reviews Immunology 15, 73 - 86 (2015).
40. Semenza, G. L. Cancer - stromal cell interactions mediated by hypoxia-inducible factors promote angiogenesis, lymphangiogenesis, and metastasis. Oncogene 32, 4057 - 4063 (2013).
41. Cleary, A. S., Leonard, T. L., Gestl, S. A. & Gunther, E. J. Tumour cell heterogeneity maintained by cooperating subclones in wnt-driven mammary cancers. Nature 508, 113 - 117 (2014).
42. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. Science 352, 189 - 196 (2016).
43. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. Nature 478, 343 - 348 (2011).
44. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of rna-seq expression data across species, tissues and studies. Genome biology 16, 1 (2015).
45. Cordell, H. J. Detecting gene - gene interactions that underlie human diseases. Nature Reviews Genetics 10, 392 - 404 (2009).
46. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8, e1002375 (2012).
47. Creixell, P. et al. Pathway and network analysis of cancer genomes. Nature methods 12, 615 (2015).
48. Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A. & Ideker, T. The cancer cell map initiative: defining the hallmark networks of cancer. Molecular cell 58, 690 - 698 (2015).
49. Berretta, R. & Moscato, P. Cancer biomarker discovery: the entropic hallmark. PLoS One 5, e12262 (2010).
50. Breitkreutz, D., Hlatky, L., Rietman, E. & Tuszynski, J. A. Molecular signaling network complexity is correlated with cancer patient survivability. Proceedings of the National Academy of Sciences 109, 9209 - 9212 (2012).
51. Csermely, P. et al. Cancer stem cells display extremely large evolvability:

alternating plastic and rigid networks as a potential mechanism: network models, novel therapeutic target strategies, and the contributions of hypoxia, inflammation and cellular senescence. In Seminars in cancer biology, vol. 30, 42 – 51 (Elsevier, 2015).

52. Liu, R. et al. Identifying critical transitions and their leading biomolecular networks in complex diseases. Scientific reports 2 (2012).
53. Cheng, F., Liu, C., Shen, B. & Zhao, Z. Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. BMC Systems Biology 10, 65 (2016).
54. Menichetti, G., Bianconi, G., Castellani, G., Giampieri, E. & Remondini, D. Multiscale characterization of ageing and cancer progression by a novel network entropy measure. Molecular BioSystems 11, 1824 – 1831 (2015).
55. Teschendorff, A. E. & Severini, S. Increased entropy of signal transduction in the cancer metastasis phenotype. BMC systems biology 4, 1 (2010).
56. Banerji, C. R. et al. Cellular network entropy as the energy potential in waddington's differentiation landscape. Scientific reports 3 (2013).
57. Banerji, C. R., Severini, S., Caldas, C. & Teschendorff, A. E. Intra-tumour signalling entropy determines clinical outcome in breast and lung cancer. PLoS Comput Biol 11, e1004115 (2015).
58. Lin, J. Divergence measures based on the shannon entropy. Information Theory, IEEE Transactions on 37, 145 – 151 (1991).
59. Kim, K.-T. et al. Single-cell mrna sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol 16, 127 (2015).
60. Klijn, C. et al. A comprehensive transcriptional portrait of human cancer cell lines. Nature biotechnology 33, 306 – 312 (2015).
61. Chen, H., Lin, F., Xing, K. & He, X. The reverse evolution from multicellularity to unicellularity during carcinogenesis. Nature communications 6 (2015).
62. Franceschini, A. et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research 41, D808 – D815 (2013).
63. Huttlin, E. L. et al. The bioplex network: a systematic exploration of the human

interactome. Cell 162, 425 - 440 (2015).

64. Das, J. & Yu, H. Hint: High-quality protein interactomes and their applications in understanding human disease. BMC systems biology 6, 92 (2012).
65. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications 4 (2013).
66. Therneau, T. M. A Package for Survival Analysis in R (2015). URL http://CRAN.R-project.org/package=survival. Version 2.38.
67. Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. IEEE Transactions on Information theory (2003).
68. Kullback, S. & Leibler, R. A. On information and sufficiency. The annals of mathematical statistics 22, 79 - 86 (1951).
69. Chen, H. & He, X. The convergent cancer evolution toward a single cellular destination. Molecular biology and evolution 33, 4 - 12 (2016).
70. Cheng, F. et al. Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. Molecular biology and evolution 31, 2156 - 2169 (2014).
71. Jia, P. & Zhao, Z. Impacts of somatic mutations on gene expression: an association perspective. Briefings in bioinformatics bbw037 (2016).
72. Tamborero, D. et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific reports 3, 2650 (2013).
73. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. Nature communications 6 (2015).
74. Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. Nature genetics 47, 312 - 319 (2015).
75. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 160, 48 - 61 (2015).
76. Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nature medicine 21, 938 - 945 (2015).
77. Martinez-Garcia, R. et al. Transcriptional dissection of pancreatic tumors engrafted in mice. Genome medicine 6, 1 (2014).

78. Warburg, O. et al. On the origin of cancer cells. Science 123, 309–314 (1956).
79. Ward, P. S. & Thompson, C. B. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer cell 21, 297–308 (2012).
80. Gatenby, R. A. & Gillies, R. J. Why do cancers have high aerobic glycolysis? Nature Reviews Cancer 4, 891–899 (2004).
81. Chang, L. & Karin, M. Mammalian map kinase signalling cascades. Nature 410, 37–40 (2001).
82. West, A. B., Dawson, V. L. & Dawson, T. M. To die or grow: Parkinson's disease and cancer. Trends in neurosciences 28, 348–352 (2005).
83. Lin, P.-Y. et al. Association between parkinson disease and risk of cancer in taiwan. JAMA oncology 1, 633–640 (2015).
84. Gong, Y. et al. Pan-cancer genetic analysis identifies park2 as a master regulator of g1/s cyclins. Nature genetics 46, 588 (2014).
85. Looyenga, B. D. et al. Chromosomal amplification of leucine-rich repeat kinase-2 (lrrk2) is required for oncogenic met signaling in papillary renal and thyroid carcinomas. Proceedings of the National Academy of Sciences 108, 1439–1444 (2011).
86. Davies, P. C. & Lineweaver, C. H. Cancer tumors as metazoa 1.0: tapping genes of ancient ancestors. Physical biology 8, 015001 (2011).
87. Aktipis, C. A. et al. Cancer across the tree of life: cooperation and cheating in multicellularity. Phil. Trans. R. Soc. B 370, 20140219 (2015).
88. Greaves, M. Evolutionary determinants of cancer. Cancer discovery 5, 806–820 (2015).
89. Ben-Porath, I. et al. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. Nature genetics 40, 499–507 (2008).
90. Kumar, S. M. et al. Acquired cancer stem cell phenotypes through oct4-mediated dedifferentiation. Oncogene 31, 4898–4911 (2012).
91. Klusza, S. & Deng, W.-M. At the crossroads of differentiation and proliferation: Precise control of cell-cycle changes by multiple signaling pathways in drosophila follicle cells. Bioessays 33, 124–134 (2011).

92. Lewis, E. B. A gene complex controlling segmentation in drosophila. In Genes, Development and Cancer, 205 – 217 (Springer, 1978).
93. Roesch, A. et al. A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth. Cell 141, 583 – 594 (2010).
94. Moore, N. & Lyle, S. Quiescent, slow-cycling stem cell populations in cancer: a review of the evidence and discussion of significance. Journal of oncology 2011 (2010).
95. Zhang, W., Zeng, T. & Chen, L. Edgemarker: identifying differentially correlated molecule pairs as edge-biomarkers. Journal of theoretical biology 362, 35 – 43 (2014).
96. Zhang, W., Zeng, T., Liu, X. & Chen, L. Diagnosing phenotypes of single-sample individuals by edge biomarkers. Journal of molecular cell biology mjv025 (2015).
97. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods 12, 453 – 457 (2015).

Chapter 5

Appendix 1

Correlation between gITH and KEGG pathway-tITH

KEGG pathway name	p-value	c-index
Fanconi anemia pathway	0.543198	0
Cell cycle	0.538414	0
p53 signaling pathway	0.522782	0
Nucleotide excision repair	0.51885	0
Homologous recombination	0.513607	0
DNA replication	0.507205	0
Mismatch repair	0.497014	0
Base excision repair	0.487229	0
Ubiquitin mediated proteolysis	0.479705	0
Oocyte meiosis	0.474504	0
Spliceosome	0.47327	0
Olfactory transduction	0.471522	0
Proteasome	0.469864	0
MicroRNAs in cancer	0.468819	0
RNA polymerase	0.465411	0
Colorectal cancer	0.46512	0
RNA degradation	0.461172	0
Endometrial cancer	0.460541	0
Protein processing in endoplasmic reticulum	0.459743	0
Pancreatic cancer	0.453766	0
RNA transport	0.449303	0
Bladder cancer	0.446358	0
Shigellosis	0.445264	0
Small cell lung cancer	0.443958	0
Non small cell lung cancer	0.433144	0
Progesterone mediated oocyte maturation	0.432866	0
Neurotrophin signaling pathway	0.431419	0
Bacterial invasion of epithelial cells	0.430792	0
Pathogenic Escherichia coli infection	0.430407	0
Acute myeloid leukemia	0.425022	0
Chronic myeloid leukemia	0.424043	0
Central carbon metabolism in cancer	0.423329	0

Glioma	0.421634	0
Cholinergic synapse	0.419081	0
Hepatitis C	0.417997	0
Dopaminergic synapse	0.417469	0
MAPK signaling pathway	0.415803	0
Legionellosis	0.414864	0
mRNA surveillance pathway	0.41447	0
Huntington's disease	0.414193	0
Glutamatergic synapse	0.411933	0
Prostate cancer	0.411479	0
Basal transcription factors	0.40701	0
Circadian entrainment	0.406464	0
Thyroid hormone signaling pathway	0.403801	0
AGE RAGE signaling pathway in diabetic complications	0.4037	0
Arrhythmogenic right ventricular cardiomyopathy	0.402922	0
Alcoholism	0.399739	0
Neuroactive ligand.receptor interaction	0.399632	0
Ras signaling pathway	0.399035	0
Pathways in cancer	0.397076	0
Insulin secretion	0.395829	0
Signaling pathways regulating pluripotency of stem cells	0.394286	0
Platelet activation	0.393453	0
Aldosterone synthesis and secretion	0.391316	0
Retrograde endocannabinoid signaling	0.391295	0
Epstein Barr virus infection	0.38948	0
Renal cell carcinoma	0.388641	0
ECM receptor interaction	0.388539	0
Influenza A	0.38838	0
Hepatitis B	0.384782	0
VEGF signaling pathway	0.384504	0
Inositol phosphate metabolism	0.384074	0
Rap1 signaling pathway	0.38373	0
Toxoplasmosis	0.383228	0
GABAergic synapse	0.382887	0
2-Oxocarboxylic acid metabolism	0.381961	0
Glycosaminoglycan degradation	0.38121	0
Cocaine addiction	0.381063	0
Thyroid cancer	0.380732	0
TGF-beta signaling pathway	0.379923	0
ErbB signaling pathway	0.379639	0
Leukocyte transendothelial migration	0.378326	0
Apoptosis	0.377509	0
HIF-1 signaling pathway	0.377479	0
Salivary secretion	0.376542	0
mTOR signaling pathway	0.375862	0
Amyotrophic lateral sclerosis (ALS)	0.375781	0
Amphetamine addiction	0.375413	0
Phospholipase D signaling pathway	0.374733	0
PI3K-Akt signaling pathway	0.3747	0
Parkinson.s disease	0.373681	0
cAMP signaling pathway	0.372912	0
Galactose metabolism	0.372654	0
Long-term potentiation	0.372223	0
Calcium signaling pathway	0.371914	0
Transcriptional misregulation in cancer	0.371452	0
GnRH signaling pathway	0.370108	0
Adrenergic signaling in cardiomyocytes	0.369759	0
Oxytocin signaling pathway	0.367641	0
HTLV-I infection	0.367472	0
NOD-like receptor signaling pathway	0.366893	0
Herpes simplex infection	0.366749	0

Viral carcinogenesis	0.365942	0
Thyroid hormone synthesis	0.364638	0
Pertussis	0.363626	0
Focal adhesion	0.360435	0
Alzheimer's disease	0.359795	0
Regulation of lipolysis in adipocytes	0.359118	0
Aldosterone.regulated sodium reabsorption	0.358607	0
Ovarian steroidogenesis	0.358517	0
TNF signaling pathway	0.357906	0
Morphine addiction	0.357752	0
Fc epsilon RI signaling pathway	0.357669	0
Sphingolipid signaling pathway	0.35719	0
Chagas disease .American trypanosomiasis.	0.355683	0
T cell receptor signaling pathway	0.354586	0
Nicotine addiction	0.353161	0
Ribosome biogenesis in eukaryotes	0.352439	0
Proteoglycans in cancer	0.351403	0
Prion diseases	0.35064	0
Circadian rhythm	0.349259	0
Pancreatic secretion	0.347105	0
Leishmaniasis	0.34584	0
Axon guidance	0.343406	0
Biosynthesis of unsaturated fatty acids	0.343036	0
Serotonergic synapse	0.342571	0
cGMP.PKG signaling pathway	0.341802	0
Gastric acid secretion	0.341441	0
Prolactin signaling pathway	0.339914	0
Type II diabetes mellitus	0.339712	0
Sulfur metabolism	0.339337	0
Adherens junction	0.3385	0
Long.term depression	0.336863	0
Fatty acid metabolism	0.33675	0
Complement and coagulation cascades	0.336148	0
Gap junction	0.334532	0
Notch signaling pathway	0.334525	0
Vascular smooth muscle contraction	0.334482	0
Inflammatory bowel disease (IBD)	0.333411	0
Cell adhesion molecules (CAMs)	0.332478	0
Fc gamma R.mediated phagocytosis	0.331579	0
AMPK signaling pathway	0.330607	0
Inflammatory mediator regulation of TRP channels	0.330465	0
Melanoma	0.32983	0
Renin secretion	0.329241	0
Osteoclast differentiation	0.327064	0
Hippo signaling pathway	0.327003	0
Estrogen signaling pathway	0.322809	0
Viral myocarditis	0.321909	0
Biotin metabolism	0.321717	0
Bile secretion	0.320649	0
Salmonella infection	0.319155	0
Taste transduction	0.318834	0
Measles	0.315029	2.22E-16
Regulation of actin cytoskeleton	0.314968	2.22E-16
Endocytosis	0.314288	2.22E-16
Steroid hormone biosynthesis	0.314118	2.22E-16
Endocrine and other factor regulated calcium reabsorption	0.313207	4.44E-16
Tuberculosis	0.312944	4.44E-16
Amoebiasis	0.310864	4.44E-16
Vasopressin regulated water reabsorption	0.310822	4.44E-16
Antigen processing and presentation	0.30821	8.88E-16
African trypanosomiasis	0.306446	1.33E-15

Systemic lupus erythematosus	0.304492	2.22E-15
Linoleic acid metabolism	0.303653	2.44E-15
Chemokine signaling pathway	0.303244	2.66E-15
Dilated cardiomyopathy	0.303052	2.89E-15
Carbon metabolism	0.299682	5.77E-15
Cytokine-cytokine receptor interaction	0.293459	2.24E-14
NF-kappa B signaling pathway	0.293448	2.24E-14
FoxO signaling pathway	0.292734	2.62E-14
Hypertrophic cardiomyopathy (HCM)	0.292325	2.84E-14
Carbohydrate digestion and absorption	0.291261	3.55E-14
Phosphatidylinositol signaling system	0.290749	3.95E-14
Autoimmune thyroid disease	0.290538	4.13E-14
Toll.like receptor signaling pathway	0.288183	6.75E-14
Basal cell carcinoma	0.28805	6.93E-14
Non-alcoholic fatty liver disease (NAFLD)	0.287008	8.62E-14
Melanogenesis	0.286022	1.05E-13
Metabolic pathways	0.28602	1.05E-13
Jak-STAT signaling pathway	0.285684	1.13E-13
Maturity onset diabetes of the young	0.283624	1.72E-13
N-Glycan biosynthesis	0.28282	2.03E-13
Arachidonic acid metabolism	0.278818	4.53E-13
Ubiquinone and other terpenoid-quinone biosynthesis	0.277541	5.84E-13
Malaria	0.277534	5.84E-13
Adipocytokine signaling pathway	0.2775	5.88E-13
Protein digestion and absorption	0.277412	5.99E-13
Thiamine metabolism	0.275254	9.16E-13
Fatty acid elongation	0.273499	1.29E-12
Tyrosine metabolism	0.271899	1.76E-12
RIG-I-like receptor signaling pathway	0.270115	2.49E-12
Staphylococcus aureus infection	0.26735	4.22E-12
Phagosome	0.267153	4.38E-12
Purine metabolism	0.265737	5.73E-12
Fructose and mannose metabolism	0.264932	6.67E-12
Asthma	0.264037	7.90E-12
Lysosome	0.262264	1.10E-11
Butanoate metabolism	0.261433	1.28E-11
Propanoate metabolism	0.260922	1.41E-11
Insulin signaling pathway	0.260768	1.45E-11
Wnt signaling pathway	0.260354	1.57E-11
Sulfur relay system	0.259979	1.68E-11
Choline metabolism in cancer	0.259628	1.79E-11
Phototransduction	0.259618	1.79E-11
Allograft rejection	0.257439	2.67E-11
Cardiac muscle contraction	0.256867	2.97E-11
Hematopoietic cell lineage	0.256734	3.04E-11
B cell receptor signaling pathway	0.253968	5.01E-11
Glucagon signaling pathway	0.251352	8.00E-11
Drug metabolism	0.250011	1.01E-10
Pyrimidine metabolism	0.2439	2.94E-10
Natural killer cell mediated cytotoxicity	0.242677	3.63E-10
Graft-versus-host disease	0.241678	4.31E-10
One carbon pool by folate	0.239588	6.14E-10
Proximal tubule bicarbonate reclamation	0.236248	1.07E-09
Insulin resistance	0.230231	2.88E-09
Phenylalanine metabolism	0.227338	4.58E-09
SNARE interactions in vesicular transport	0.22693	4.89E-09
Type I diabetes mellitus	0.223122	8.91E-09
Glyoxylate and dicarboxylate metabolism	0.22189	1.08E-08
Tight junction	0.218739	1.76E-08
Valine. leucine and isoleucine degradation	0.214694	3.24E-08
Pentose phosphate pathway	0.213677	3.78E-08

Ribosome	0.211416	5.29E-08
Selenocompound metabolism	0.208628	7.96E-08
Cysteine and methionine metabolism	0.20447	1.45E-07
Glycerophospholipid metabolism	0.203322	1.71E-07
Fatty acid biosynthesis	0.202231	2.00E-07
Ether lipid metabolism	0.201413	2.24E-07
Hedgehog signaling pathway	0.199123	3.08E-07
Peroxisome	0.199088	3.10E-07
Rheumatoid arthritis	0.192012	8.14E-07
Tryptophan metabolism	0.189435	1.15E-06
Metabolism of xenobiotics by cytochrome P450	0.186727	1.64E-06
Biosynthesis of amino acids	0.186594	1.67E-06
alpha-Linolenic acid metabolism	0.183863	2.37E-06
Arginine and proline metabolism	0.183429	2.51E-06
Cytosolic DNA-sensing pathway	0.183267	2.56E-06
Sphingolipid metabolism	0.183014	2.64E-06
Starch and sucrose metabolism	0.176766	5.79E-06
Other glycan degradation	0.172609	9.62E-06
Intestinal immune network for IgA production	0.171943	1.04E-05
Glutathione metabolism	0.170791	1.20E-05
Lysine degradation	0.169491	1.40E-05
Taurine and hypotaurine metabolism	0.168238	1.62E-05
Non homologous end joining	0.167119	1.85E-05
Renin-angiotensin system	0.163524	2.80E-05
beta.Alanine metabolism	0.162945	2.99E-05
Dorso-ventral axis formation	0.162337	3.20E-05
Alanine. aspartate and glutamate metabolism	0.160234	4.06E-05
Chemical carcinogenesis	0.156403	6.21E-05
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	0.1561	6.42E-05
Synaptic vesicle cycle	0.151847	0.000102
Phenylalanine. tyrosine and tryptophan biosynthesis	0.151712	0.000103
Primary immunodeficiency	0.15109	0.00011
Nitrogen metabolism	0.147334	0.000163
Folate biosynthesis	0.146211	0.000184
Glycolysis Gluconeogenesis	0.141239	0.000304
Glycerolipid metabolism	0.139644	0.000356
Aminoacyl-tRNA biosynthesis	0.138696	0.00039
ABC transporters	0.136497	0.000484
Epithelial cell signaling in Helicobacter pylori infection	0.132463	0.00071
D-Glutamine and D-glutamate metabolism	0.130519	0.000852
Fatty acid degradation	0.122399	0.00177
Other types of O-glycan biosynthesis	0.121004	0.001999
Caffeine metabolism	0.120338	0.002117
PPAR signaling pathway	0.119478	0.002279
Primary bile acid biosynthesis	0.115201	0.003269
Butirosin and neomycin biosynthesis	0.113666	0.003711
Fat digestion and absorption	0.111449	0.004444
Arginine biosynthesis	0.106716	0.006464
Histidine metabolism	0.10261	0.008847
Retinol metabolism	0.09335	0.017285
Pentose and glucuronate interconversions	0.093179	0.017492
Vitamin digestion and absorption	0.092795	0.017963
Glycosaminoglycan biosynthesis	0.077491	0.04829

Appendix 2

TCGA pan-cancer cox model analysis with pathway-tITH

KEGG pathway name	p-value	c-index
mRNA surveillance pathway	0	0.630359
Ribosome biogenesis in eukaryotes	0	0.625562
Olfactory transduction	0	0.617829
RNA transport	0	0.610876
Purine metabolism	0	0.604419
Influenza A	2.66E-13	0.598796
Basal cell carcinoma	2.12E-12	0.594936
Homologous recombination	0	0.592557
Cell cycle	1.36E-13	0.590326
Hepatitis C	5.83E-10	0.589985
Carbon metabolism	1.96E-11	0.58984
Nucleotide excision repair	1.07E-14	0.589662
Fanconi anemia pathway	2.89E-15	0.589434
Fatty acid elongation	0	0.589344
p53 signaling pathway	8.61E-11	0.5889
Endometrial cancer	4.22E-08	0.587383
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	6.71E-13	0.586171
Colorectal cancer	2.30E-07	0.584896
Inflammatory bowel disease (IBD)	4.84E-09	0.584664
Phosphatidylinositol signaling system	2.05E-07	0.584458
Mismatch repair	1.79E-14	0.583654
Pathogenic Escherichia coli infection	1.82E-08	0.582843
Oocyte meiosis	8.47E-08	0.582668
DNA replication	2.43E-14	0.581975
Fc epsilon RI signaling pathway	1.70E-06	0.581655
Prion diseases	4.72E-09	0.580111
Leishmaniasis	4.76E-08	0.579405
Bladder cancer	3.14E-07	0.577553
Phototransduction	2.03E-09	0.577298
Serotonergic synapse	1.42E-08	0.577109
Asthma	4.60E-08	0.576683
Progesterone-mediated oocyte maturation	3.45E-06	0.576461
HTLV-I infection	6.83E-07	0.576229
Proteasome	7.18E-10	0.576095
Pancreatic cancer	1.74E-06	0.575228
MicroRNAs in cancer	1.71E-06	0.575029
VEGF signaling pathway	1.54E-05	0.574417
Taste transduction	1.51E-06	0.574338
Huntington's disease	3.49E-07	0.574285
ABC transporters	6.31E-06	0.573862
GABAergic synapse	6.00E-08	0.573662

RNA polymerase	1.00E-07	0.573635
Spliceosome	8.60E-07	0.573504
Inositol phosphate metabolism	2.87E-06	0.57272
Pathways in cancer	1.28E-05	0.572523
AMPK signaling pathway	4.61E-06	0.571889
Non-small cell lung cancer	4.30E-06	0.571851
Sphingolipid signaling pathway	9.01E-05	0.571239
Neurotrophin signaling pathway	0.000214	0.571205
AGE-RAGE signaling pathway in diabetic complications	1.95E-05	0.570945
Hedgehog signaling pathway	6.17E-06	0.570177
Thiamine metabolism	2.52E-07	0.56998
Chagas disease (American trypanosomiasis)	2.17E-05	0.569889
Prostate cancer	1.92E-05	0.569712
Aldosterone synthesis and secretion	8.54E-05	0.569512
Dopaminergic synapse	1.32E-05	0.569255
Biosynthesis of unsaturated fatty acids	1.47E-05	0.567897
Glioma	0.000206	0.567444
Chronic myeloid leukemia	6.44E-05	0.567329
Osteoclast differentiation	8.88E-05	0.567213
Retrograde endocannabinoid signaling	7.28E-06	0.567114
Fatty acid degradation	0.000218	0.566975
African trypanosomiasis	2.84E-06	0.566963
Signaling pathways regulating pluripotency of stem cells	2.40E-06	0.566839
Shigellosis	0.000106	0.566431
Acute myeloid leukemia	9.92E-05	0.56584
beta-Alanine metabolism	2.34E-07	0.565677
Phospholipase D signaling pathway	0.00055	0.565189
Apoptosis	0.000191	0.565041
Hepatitis B	3.53E-05	0.564969
Estrogen signaling pathway	2.89E-05	0.564662
Platelet activation	0.000356	0.56462
T cell receptor signaling pathway	0.000179	0.564331
Drug metabolism	1.52E-08	0.564085
Proteoglycans in cancer	7.33E-05	0.56404
PI3K-Akt signaling pathway	8.31E-05	0.563607
Glyoxylate and dicarboxylate metabolism	1.16E-06	0.563425
Herpes simplex infection	7.99E-05	0.563399
Gap junction	0.000439	0.56326
One carbon pool by folate	8.81E-06	0.563019
Wnt signaling pathway	7.01E-05	0.562992
Carbohydrate digestion and absorption	2.02E-05	0.562814
Propanoate metabolism	1.66E-12	0.562486
Cholinergic synapse	3.74E-05	0.562178
Toxoplasmosis	0.000732	0.562146
Jak-STAT signaling pathway	0.000333	0.562139
Ubiquitin mediated proteolysis	3.15E-05	0.561526
Morphine addiction	4.05E-05	0.561089
Small cell lung cancer	0.000142	0.561081
Cell adhesion molecules (CAMs)	9.55E-05	0.561003
Glutathione metabolism	0.000102	0.560622
Long-term depression	0.000157	0.560575
Hematopoietic cell lineage	6.53E-06	0.560484
Axon guidance	0.000485	0.560309
Type I diabetes mellitus	4.48E-06	0.560268
Hippo signaling pathway	7.58E-05	0.559419
Insulin signaling pathway	0.000379	0.559071
Circadian rhythm	0.000187	0.558208
Natural killer cell mediated cytotoxicity	0.000156	0.558188
Aminoacyl-tRNA biosynthesis	4.06E-10	0.557729
Pyrimidine metabolism	2.26E-06	0.55684
Butanoate metabolism	2.89E-06	0.556683

Allograft rejection	0.000367	0.556627
Glucagon signaling pathway	0.000573	0.556306
2-Oxocarboxylic acid metabolism	6.02E-09	0.556272
Fatty acid metabolism	0.000118	0.556181
Transcriptional misregulation in cancer	1.85E-05	0.556067
Melanogenesis	0.000233	0.555755
Glycolysis Gluconeogenesis	4.10E-05	0.555659
Maturity onset diabetes of the young	6.97E-07	0.555127
Notch signaling pathway	9.50E-06	0.554771
RNA degradation	5.21E-07	0.553646
Nicotine addiction	0.00065	0.553149
ECM-receptor interaction	0.000513	0.552858
Malaria	1.26E-05	0.552716
Galactose metabolism	4.17E-08	0.552271
Melanoma	0.000679	0.552252
Lysosome	0.000781	0.550892
alpha-Linolenic acid metabolism	3.18E-09	0.549792
Starch and sucrose metabolism	1.89E-05	0.549016
Base excision repair	0.00023	0.547704
Fructose and mannose metabolism	0.000327	0.54687
Amino sugar and nucleotide sugar metabolism	5.45E-09	0.545852
Histidine metabolism	0.000392	0.544255
Butirosin and neomycin biosynthesis	0.000508	0.542508
Protein export	8.51E-06	0.54221
Cardiac muscle contraction	2.80E-05	0.536448
Ether lipid metabolism	1.86E-05	0.530867
Dilated cardiomyopathy	0.000549	0.524913
Hypertrophic cardiomyopathy (HCM)	0.000608	0.524869
PPAR signaling pathway	0.000568	0.519411

Appendix 3

Xenograft data correlation between tITH and number of major clone

KEGG pathway name	correlation
Glycosphingolipid biosynthesis	0.940755
Other types of O-glycan biosynthesis	0.898637
Thyroid cancer	0.8954
Citrate cycle (TCA cycle)	0.893555
Ribosome	0.889963
Parkinson's disease	0.886861
Glycosaminoglycan biosynthesis	0.886838
Butirosin and neomycin biosynthesis	0.874196
Terpenoid backbone biosynthesis	0.870596
Protein export	0.861782
Protein processing in endoplasmic reticulum	0.850309
mRNA surveillance pathway	0.838567
Cell cycle	0.837966
Oxidative phosphorylation	0.837536
DNA replication	0.837356
N-Glycan biosynthesis	0.825608
ABC transporters	0.824433
Non-alcoholic fatty liver disease (NAFLD)	0.822856
Spliceosome	0.819715
Huntington's disease	0.819138
Insulin signaling pathway	0.813044
Carbon metabolism	0.804002
Alzheimer's disease	0.80034
Ubiquinone and other terpenoid-quinone biosynthesis	0.797543
Nucleotide excision repair	0.796506
Riboflavin metabolism	0.795827
Mucin type O-Glycan biosynthesis	0.793335
Other glycan degradation	0.789931
Primary immunodeficiency	0.789037
FoxO signaling pathway	0.788088
Propanoate metabolism	0.785753
T cell receptor signaling pathway	0.7857
Ovarian steroidogenesis	0.779908
Maturity onset diabetes of the young	0.77945
Proteasome	0.76476
Protein digestion and absorption	0.76322
Acute myeloid leukemia	0.761954
D-Glutamine and D-glutamate metabolism	0.761763
Epstein-Barr virus infection	0.760006
RNA transport	0.757161
Metabolic pathways	0.744838

NOD-like receptor signaling pathway	0.734884
Fatty acid metabolism	0.729824
Thyroid hormone signaling pathway	0.72822
Prolactin signaling pathway	0.727217
Circadian rhythm	0.722114
RNA degradation	0.722048
Proximal tubule bicarbonate reclamation	0.72067
Insulin resistance	0.714533
Pentose phosphate pathway	0.701829
2-Oxocarboxylic acid metabolism	0.697318
Cardiac muscle contraction	0.692155
Adipocytokine signaling pathway	0.689444
Non-small cell lung cancer	0.680639
Basal transcription factors	0.674808
Ubiquitin mediated proteolysis	0.672935
p53 signaling pathway	0.671621
ECM-receptor interaction	0.6706
RNA polymerase	0.669386
PPAR signaling pathway	0.662706
Glycolysis Gluconeogenesis	0.662485
Butanoate metabolism	0.661868
Amino sugar and nucleotide sugar metabolism	0.660142
Ribosome biogenesis in eukaryotes	0.653694
Natural killer cell mediated cytotoxicity	0.649047
Central carbon metabolism in cancer	0.643802
Colorectal cancer	0.636461
Bile secretion	0.634361
Pyrimidine metabolism	0.633991
Sphingolipid metabolism	0.625771
Phenylalanine metabolism	0.622046
Renal cell carcinoma	0.611568
Folate biosynthesis	0.604201
B cell receptor signaling pathway	0.600563
Taurine and hypotaurine metabolism	0.59082
mTOR signaling pathway	0.586535
Thiamine metabolism	0.584694
Salivary secretion	0.584237
Non-homologous end-joining	0.574616
Endometrial cancer	0.568403
Tight junction	0.568281
Hippo signaling pathway	0.563708
Prostate cancer	0.560112
Starch and sucrose metabolism	0.556205
Oocyte meiosis	0.550821
Synthesis and degradation of ketone bodies	0.549364
Chronic myeloid leukemia	0.548331
Endocytosis	0.541776
Notch signaling pathway	0.53886
Inflammatory bowel disease (IBD)	0.536678
Phagosome	0.530973
Focal adhesion	0.527931
Glyoxylate and dicarboxylate metabolism	0.525568
Glycine, serine and threonine metabolism	0.522503
Primary bile acid biosynthesis	0.521134
Dopaminergic synapse	0.519699
Mineral absorption	0.519321
PI3K-Akt signaling pathway	0.518572
Caffeine metabolism	0.516292
Malaria	0.513138
Neurotrophin signaling pathway	0.510463
Aldosterone-regulated sodium reabsorption	0.487196

Aminoacyl-tRNA biosynthesis	0.482688
Glioma	0.480063
Dorso-ventral axis formation	0.477935
Renin-angiotensin system	0.476722
HIF-1 signaling pathway	0.476562
beta-Alanine metabolism	0.465173
Calcium signaling pathway	0.46267
Leukocyte transendothelial migration	0.450913
Viral carcinogenesis	0.447315
Fc gamma R-mediated phagocytosis	0.447035
Nitrogen metabolism	0.440737
Olfactory transduction	0.436127
Peroxisome	0.43134
AGE-RAGE signaling pathway in diabetic complications	0.428546
Porphyrin and chlorophyll metabolism	0.412105
Pancreatic secretion	0.402525
Carbohydrate digestion and absorption	0.399834
Fc epsilon RI signaling pathway	0.397774
Lysine degradation	0.391736
Glucagon signaling pathway	0.387638
VEGF signaling pathway	0.383853
AMPK signaling pathway	0.381226
Neuroactive ligand-receptor interaction	0.380504
Alcoholism	0.378294
Phospholipase D signaling pathway	0.344609
Phenylalanine, tyrosine and tryptophan biosynthesis	0.344582
Signaling pathways regulating pluripotency of stem cells	0.340166
Ascorbate and aldarate metabolism	0.330573
Choline metabolism in cancer	0.317522
Regulation of autophagy	0.29974
cGMP-PKG signaling pathway	0.29863
Small cell lung cancer	0.298496
Fatty acid elongation	0.296931
Steroid hormone biosynthesis	0.296295
Purine metabolism	0.291484
Transcriptional misregulation in cancer	0.288571
Long-term depression	0.2833
Cysteine and methionine metabolism	0.282482
Proteoglycans in cancer	0.281433
African trypanosomiasis	0.274089
Staphylococcus aureus infection	0.273642
Asthma	0.271417
One carbon pool by folate	0.259222
Cholinergic synapse	0.25676
Renin secretion	0.245688
Cocaine addiction	0.242389
Melanogenesis	0.238737
Osteoclast differentiation	0.23843
Pyruvate metabolism	0.229871
Pancreatic cancer	0.228873
Adrenergic signaling in cardiomyocytes	0.226157
Insulin secretion	0.221674
Fatty acid degradation	0.218632
Hypertrophic cardiomyopathy (HCM)	0.215331
Pathways in cancer	0.214923
SNARE interactions in vesicular transport	0.213672
Hematopoietic cell lineage	0.213649
TGF-beta signaling pathway	0.212077
Regulation of lipolysis in adipocytes	0.211786
Biosynthesis of unsaturated fatty acids	0.203137
Hepatitis B	0.194702

MicroRNAs in cancer	0.18327
ErbB signaling pathway	0.17664
Biotin metabolism	0.17258
Vasopressin-regulated water reabsorption	0.16846
MAPK signaling pathway	0.152152
Shigellosis	0.151748
Amphetamine addiction	0.147335
Tuberculosis	0.132774
Vitamin digestion and absorption	0.130224
Chagas disease (American trypanosomiasis)	0.126222
Type II diabetes mellitus	0.119724
Bacterial invasion of epithelial cells	0.11818
Estrogen signaling pathway	0.113375
Complement and coagulation cascades	0.111294
Sphingolipid signaling pathway	0.098962
Melanoma	0.097176
Platelet activation	0.091386
Pertussis	0.080096
Bladder cancer	0.078838
Hepatitis C	0.07054
Aldosterone synthesis and secretion	0.069054
Circadian entrainment	0.068873
Toxoplasmosis	0.068092
Axon guidance	0.067047
Wnt signaling pathway	0.063613
Thyroid hormone synthesis	0.055527
Regulation of actin cytoskeleton	0.042105
Mismatch repair	0.042044
GnRH signaling pathway	0.037739
Taste transduction	0.027163
Histidine metabolism	0.02695
TNF signaling pathway	0.024764
Metabolism of xenobiotics by cytochrome P450	0.021511
Steroid biosynthesis	0.020126
Cell adhesion molecules (CAMs)	0.014778
Progesterone-mediated oocyte maturation	0.009273
Alanine, aspartate and glutamate metabolism	0.008506
Chemical carcinogenesis	0.005602
Dilated cardiomyopathy	0.004517
cAMP signaling pathway	-0.00222
Cytokine-cytokine receptor interaction	-0.00239
Chemokine signaling pathway	-0.00668
Tyrosine metabolism	-0.01549
Rap1 signaling pathway	-0.0204
Base excision repair	-0.0235
Glutamatergic synapse	-0.02655
Measles	-0.04763
Amyotrophic lateral sclerosis (ALS)	-0.07743
Antigen processing and presentation	-0.07876
Morphine addiction	-0.08066
Legionellosis	-0.08328
Endocrine and other factor-regulated calcium reabsorption	-0.09334
Arginine and proline metabolism	-0.09965
Basal cell carcinoma	-0.10148
Degradation of aromatic compounds	-0.10201
GABAergic synapse	-0.10246
Oxytocin signaling pathway	-0.10327
Lysosome	-0.14276
Vibrio cholerae infection	-0.17978
Retinol metabolism	-0.18189
Epithelial cell signaling in Helicobacter pylori infection	-0.18316

Retrograde endocannabinoid signaling	-0.18904
HTLV-I infection	-0.24934
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	-0.2509
Phototransduction	-0.26034
Ras signaling pathway	-0.26837
Influenza A	-0.2712
Fat digestion and absorption	-0.29984
Gastric acid secretion	-0.30423
Pathogenic Escherichia coli infection	-0.30976
Amoebiasis	-0.31746
Nicotine addiction	-0.31937
Fatty acid biosynthesis	-0.32793
Sulfur metabolism	-0.34093
Collecting duct acid secretion	-0.35524
Serotonergic synapse	-0.35664
Jak-STAT signaling pathway	-0.35717
Systemic lupus erythematosus	-0.3698
Leishmaniasis	-0.38091
Glycerophospholipid metabolism	-0.38445
Vascular smooth muscle contraction	-0.39226
Biosynthesis of amino acids	-0.39459
alpha-Linolenic acid metabolism	-0.39635
Arginine biosynthesis	-0.3966
Autoimmune thyroid disease	-0.40767
Ether lipid metabolism	-0.40927
Synaptic vesicle cycle	-0.4169
Pentose and glucuronate interconversions	-0.42335
Selenocompound metabolism	-0.42636
Salmonella infection	-0.43017
RIG-I-like receptor signaling pathway	-0.43259
Cyanoamino acid metabolism	-0.43405
Tryptophan metabolism	-0.43441
Pantothenate and CoA biosynthesis	-0.43764
Gap junction	-0.4454
Valine, leucine and isoleucine degradation	-0.45785
Prion diseases	-0.46297
Inflammatory mediator regulation of TRP channels	-0.46697
Inositol phosphate metabolism	-0.51161
Long-term potentiation	-0.52951
Intestinal immune network for IgA production	-0.54055
Apoptosis	-0.543
Cytosolic DNA-sensing pathway	-0.55091
Viral myocarditis	-0.57345
Phosphatidylinositol signaling system	-0.57793
Herpes simplex infection	-0.5894
Drug metabolism	-0.59088
Toll-like receptor signaling pathway	-0.59272
Homologous recombination	-0.62655
Linoleic acid metabolism	-0.63239
NF-kappa B signaling pathway	-0.63378
Glutathione metabolism	-0.64568
Allograft rejection	-0.67149
Glycerolipid metabolism	-0.6887
Glycosaminoglycan degradation	-0.69246
Galactose metabolism	-0.69836
Sulfur relay system	-0.72322
Graft-versus-host disease	-0.72397
Hedgehog signaling pathway	-0.72967
Fructose and mannose metabolism	-0.73444
Rheumatoid arthritis	-0.76823
Adherens junction	-0.77763

Vitamin B6 metabolism	-0.78391
Arachidonic acid metabolism	-0.84811
Nicotinate and nicotinamide metabolism	-0.85898
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	-0.87017
Type I diabetes mellitus	-0.87078
Fanconi anemia pathway	-0.94469

요약

암의 발생, 항암제 저항성, 암의 재발과 같은 암의 특성을 설명하는데 있어서, 암 클론의 진화 모델(Clonal evolution)은 충분한 설명력을 가진다고 인정받고 있다. 이와 같은 이유로 암 클론의 진화 과정에서 생겨나는 종양 내 이질성도(Intratumor heterogeneity)를 측정하는 문제가 암의 특성을 이해하는데 점점 중요한 주제가 되고 있다. 최근 유전체의 돌연변이 정보를 이용하여 종양 내 이질성도를 측정하는 방법론들이 많이 제시가 되었고, 이를 활용하여 The Cancer Genome Atlas (TCGA)의 대규모 환자 데이터 분석 결과가 발표되었다. 종양 내 이질성도와 환자의 예후간의 상관관계를 조사해본 결과를 보면, 높은 이질성도를 갖는 환자가 낮은 이질성도를 갖는 환자보다 예후가 안 좋다는 것이 밝혀졌다. 이번 논문에서는 이런 종양 내 이질성도를 전사체 수준에서 측정해내는 새로운 방법을 제시하고자 하였다. 전사체 정보를 효율적으로 이용하기 위해서 생물학적 네트워크인 단백질 간 관계 네트워크와 패스웨이 정보를 활용하였다. 네트워크 정보와 전사체 정보를 통합하기 위하여 엔트로피 기반의 젠슨-샤논 발산값(Jensen-Shannon Divergence)를 응용하여 적용하였다. 이를 통하여 대규모 환자군 데이터에서 기존의 유전체를 활용한 방법론과 비교하여 합리적인 결과를 얻어내었으며, 환자의 예후를 예측하는 것에서는 전사체를 활용하는 본 논문의 방법이 더 좋은 결과를 얻었다. 암의 진화 모델의 데이터에서도 유전체의 결과와 일치하는 관측을 하였다. 이를 통하여 볼 때 전사체를 활용하는 종양 내 이질성도를 측정하는 방법이 이질성도를 측정하는 새로운 효율적인 방법이 될 수 있음을 보였다.

주요어: 종양 내 이질성, 클론의 진화, 네트워크 위상, 무질서도, 유전자 발현양, RNA 시퀀싱, 데이터 마이닝, 진단

학 번: 2015-20505