Tumor subclonal progression model for cancer hallmark acquisition

Yusuke Matsui[(1)], Satoru Miyano[(2)], Teppei Shimamura[(1)]

(1) Nagoya University Graduate School of Medicine
Division of Systems Biology, 65 Tsurumai-cho Showa-ku, Nagoya, 466-8550, Japan,
ymatsui@med.nagoya-u.ac.jp

(2) Institute of Medical Science, The University of Tokyo
Laboratory of DNA Information Analysis, Human Genome Center, 4-6-1
Shirokanedai, Minatoku, Tokyo 108-8639, Japan

Keywords: Cancer evolution, Multi-regional sequence, Pathway alternation, Probabilistic causality, clear cell renal cell carcinomas.

Abstract. Recent advances in the methodologies of reconstructing cancer evolutionary trajectories opened the horizon for deciphering the subclonal populations and their evolutionary architectures under the cancer ecosystems. An important challenge of the cancer evolution studies is connecting genetic aberrations in subclones to clinically interpretable and actionable target of subclones for individual patients. In this paper, we present a novel method for constructing tumor subclonal progression model for cancer hallmark acquisition using multi-regional sequencing data. We parepare a subclonal evolutionary tree inferred from variant allele frequencies and estimate the pathway alternation probabilities from large scale cohort genomic data. We then construct an evolutionary tree of pathway alternation that takes account of selectivity of pathway alternations by the notion of probabilistic causality. We show the effectiveness of our method using a dataset of clear cell renal cell carcinomas.

1 Scientific Background

Cancer is a heterogeneous genetic disease characterized by the dynamic evolution through acquiring genomic aberrations. The clonal theory of cancer proposed by Nowell [1] hypothesized that acquisition of the mutation in cancer follows natural selection in the Darwinian model, in which cancer obtains the advantages of biological fitness under the selective pressure.

The development of multi-regional sequencing techniques has provided new perspectives of genetic heterogeneity [2]. From the studies of multi-regional sequencing, spatially distinct regions within the same tumor acquire different set of somatic single nucleotide variants (SSNVs) and it is called intra-tumor heterogeneity. Recently, reconstructing method of cancer evolutionary structures are extensively studied. Since the cell population of each region is admixture of the normal and tumor cells, distinct regions are deconvolued into cell sub-populations called subclones and then they are assigned to tree structures under the the constraint that is derived from infinite site assumption [3].

However, identifications of clinically actionable subclone targets for individual patients remained unresolved. One of the reasons is the difficulties for identifying the most plausible biological event from the limited number of SSNVs, which is due to the sequencing depth and low frequencies of mutation.

To overcome the situation, we present a novel method to infer a tumor progression model for cancer hallmark acquisitions. We estimate pathway alternation probabilities and probabilistic causalities between pathway alternations using large scale cohort genomic data and we integrate them with cancer subclonal evolutioanry trees. We demonstrate the effectiveness of our method using the acutual dataset of clear cell renal cell carcinomas.

2 Materials and Methods

Our method consists of 3 steps: (1) Constructing skelton of cancer subclonal evolutionary tree (2) Estimation of pathway alternation probability and probabilistic causality (3) Constructing progression model of pathway alternations.

At first, we construct an a priori evolutionary tree of pathway alternation progression model, called a skelton, to decompose the cell population into subclones and infer the subclonal evolutionary structures for each patient based on multi-regional VAFs. Secondly, we estimate the pathway alternation probability based on large cohort data to identify the most likely pathway alternations in the subclones and to infer the probabilistic causation among the pathway alternations in the subsequent step. In the last step, we construct the tumor progression model of pathway alternations based on the skelton and pathway alternation probabilities. After scanning each subclone whether at least one SSNV is included in the given pathways, we remove the subclones in the first case or we identify the unique pathway alternation from those pathways int the second case under 3 assumptions since there are cases of no candidate pathway alternation or multiple candidate pathway alternations per subclone.

(Assumption 1) No pathway alternation occurs twice in the course of cancer evolution.

(Assumption 2) Any pathway alternation never lost.

The assumption 1 and 2 mean that if a given pathway alternation occured in any subclone, it occurs exactly one time in the course of tumor progression. We reconstruct the progression model from ancestral subclones and we never use the pathway alternations in those subclones for their descendant subclones. In addition to the two assumptions, we assume a selective pressures between the pathway alternations.

(Assumption 3) There is a selective pressure between the pathway alternations.

We model this based on the notion of the probabilistic causation that can be estimated from the large scale cohort datasets, that is similar approach to [5]. Using the three assumptions, we identify a unique pathway alternation from the multiple candidates of pathway alternations with the strongest probabilistic causation. In the following section, we describe more details of our method.

2.1 Constructing skelton of cancer subclonal evolutionary tree

Skelton represents the subclonal evolutionary tree based on variant allele frequencies (VAFs) obtained from a single patient with multiple regions by a bulk sequencing. The VAFs are approximately proportional to the sizes of cell population with the set of SSNVs, however, in the settings of bulk data, each region might be admixture of normal and tumor cells and need to deconvolute the cell populations into sub-populations. Identified sublones are assigned to tree structures employing the 2 assumptions: (i) a mutation cannot recur during the course of cancer evolution, and (ii) no mutation can be lost [6]. Several approaches are implemented to deal with the problem. Using one of the algorithms LICHeE [3], we construct the skelton from the VAFs for each patient. Let be $T_i^0 = (V_i, E_i)$ as a skelton of a patient $i; i = 1, 2, \ldots, n$ with a set of vertices $V_i = \{v_{ij}; i = 1, 2, \ldots, n, j = 1, 2, \ldots, \eta_i\}$ and edges $E_i = \{e_{ik}; i = 1, 2, \ldots, p, k = 1, 2, \ldots \nu_i\}$ where vertices and edges represent subclones with a set of SSNVs and evolutionary relations, respectively. Without loss of generality, $v_{i,j=1}$ always represents normal cells. Each vertex has a set of labels that can be obtained from a mapping $L : v_{i,j} \mapsto L_{i,j}$, *e.g.*, $L_{i,j} = \{\text{SSNV1, SSNV2, SSNV3}\}$.

2.2 Estimation of pathway alternation probability and probabilistic causality

To execute phenotypic characterization of each subclone, we need to identify the most closely related pathway alternations for subclones. There are mainly two approaches for detection of pathway alternations; One is the knowledge based gene enrichment analysis such as Fisher's exact test and the other is the de novo oriented approach where the alternation patterns are mapped to large scale protein networks and identify the subnetworks as a driver pathway with cost functions such as the tendancy of mutual exclusivity. In this study, we focus on the knowledge based approach since the biological validation for the de novo pathways are usually difficult to perform quickly.

Using the large scale cohort data because of the nature of limitation of sample size in experimental data, we estimate the pathway alternation probability using SLALPenrich [4] that is the state-of-art method for identifying pathway alternation incorporating the background pathway alternation probabilities and the mutual exclusivities. Let P denote the pathway list and then the main output of SLAPenrich is the P-value $p_i; i = 1, 2, \ldots, |P|$ for each pathway i that represents the significance of the enrichment of mutations and optionally we can obtain pathway alternation probabilities for each samples, that is, $\rho_{i,j} = \Pr(X_{i,j} \geq 1); i = 1, 2, \ldots, |P|, j = 1, 2, \ldots, N$ where $X_{i,j}$ represents the number of SSNVs that are included in the pathway i in the sample j.

We define a binary variable $y_{i,j}$ if $\rho_{i,j} > t$ then $y_{i,j} = 1$ otherwise $y_{i,j} = 0$ with a given threshold $0 \leq t \leq 1$. We evaluate the probabilistic causation though probability raising score $S_{k\to l}; k, l = 1, 2, \ldots, |P|, k \neq l$, defined as

$$S_{k\to l} = \Pr(Y_k|Y_l = 1) - \Pr(Y_k|Y_l = 0) \tag{1}$$

where Y_k represents alternation status variable of pathway k. We empirically estimate the $S_{k\to l}$ by

$$\tilde{S}_{k\to l} = \frac{\sum_{j=1}^{N} y_{k,j|y_{l,j}=1} - \sum_{k=1}^{N} y_{k,j|y_{l,j}=0}}{P}. \tag{2}$$

We consider zero or negative values are non causal relation,*i.e.,* $\tilde{S}_{k\to l} = 0$ if $\tilde{S}_{k\to l} \leq 0$.

2.3 Constructing progression model of pathways alternation

Now we are ready to construct the subclonal evolutionary tree for pathway alternations. Given a subclone, we at first scan the SSNVs to identify the candidate pathway alternations. If at least one SSNVs is included in a pathway, it is called as a candidate pathway alternation. Let $P_k; k = 1, 2, \ldots, |P|$ be the genes included in the pathway k and $Z_{i,j,k}$ be the candidate pathway alternation status of subclone j in patient i where $Z_{i,j,k} = 1$ if $L(v_{i,j}) \subseteq P_k; i = 1, 2, \ldots, n, j = 1, 2, \ldots, \eta_i$ otherwie $Z_{i,j,k} = 0$. In case of $Z_{i,j,k} = 0$ for all k, we regard the subclone as non-functional one and remove the node $v_{i,j}$ and the corresponding edges from the skelton T_i^0.

From the candidate pathway alternations $Z_{i,j,k}$, we identify the unique pathway alternation. In case that ancestral subclone is normal cells, we select pathway with the smallest P-value, *i.e.,*

$$\text{argmin}_k\, p_k \text{ for } Z_{i,j,k} = 1 \text{ and } k \notin Q \tag{3}$$

Otherwise, we select pathway with the largest probability rasing score, *i.e.,*

$$\text{argmax}_k\, S_{l\to k} \text{ for } Z_{i,j,k} = 1 \text{ and } k \notin Q \tag{4}$$

where l is the pathway alternation of the ancestral subclone and Q is a set of pathways alternations that have already appeared in the ancestral subclones. The assumption (1) and (2) are assured by condition of $k \notin Q$. If there is no corresponding pathway alternation because all the candidate pathway alternations have already happened in the ancestral subclones, then we remove the subclone as non functional one.

2.4 Dataset

A dataset from the study of clear cell renal cell carcinomas (ccRCCs) [7] were used for the analysis. Whole exome multi-regional bulk sequencing was performed for 8 individuals with clinical information and 587 out of 602 mutations were remained after filtering mutations with depth less than $100\times$.

The estimation of pathway alternation probabilities followed SLAPenrich procedures described in [4]. The 417 KIRC (corresponding to ccRCC) samples from The cancer genome atlas (TCGA) and international cancer genome consortium (ICGC) and high confidence variants identified in the study [8] were used for estimation of the pathway alternation probabilities. Pathway gene sets were downloaded from the pathway Commons data portal (v8, 2016/04) and gene sets containing less than 4 or more than 1,000 genes were discarded. After merging the gene sets that correspond to the same pathway across the multiple data source or have a large overlap defined by Jarccard index ≥ 0.8, we obtained 1,911 pathway gene sets. Cancer hallmarks were manually curated and assign them to the 456 pathways [4].

3 **Results**

We reconstructed the skeltons from VAFs by LICHeE using the same parameter with their experimental settings that are described in [3] and finally 8 skeltons were obtained. Next we estimated the pathway alternation probabilities based on SLAPenrich and obtained P-values for 209 pathways and pathway alternation probabilities for 417 patients. We evaluated the probabilistic causalities with the threshold $t = 0.1$.

We show the results of constructed the ccRCC progression models for cancer hallmark acquisition in Figure 1. At first, we simply count the number of cancer hallmarks observed in the common ancestral subclones (trunk) and subclones without any descendants (private) shown in Table 1.

Table 1: Counts of subclones (patients) with each cancer hallmark. For example, '10 (7)' in Sustaining Proliferative Signaling means 10 subclones have the cancer hallmark and it was observed in 7 patients. The columns of trunk and private count the number of cancer hallmarks in the common ancestral subclone and subclones without any descendant, respectively.

Cancer hallmarks	Total number	Trunk	Private	Other
Sustaining Proliferative Signaling	10 (7)	1 (1)	5 (5)	4 (1)
Evading Growth Suppressors	6 (4)	0 (0)	2 (2)	4 (2)
Avoiding Immune Destruction	1 (1)	0 (0)	1 (1)	0 (0)
Enabling Replicative Immortality	2 (2)	0 (0)	1 (1)	(1)
Tumour-Promoting Inflammation	2 (2)	0 (0)	2 (2)	0 (0)
Activating Invasion and Metastasis	7 (5)	0 (0)	6 (5)	1 (0)
Inducing Angiogenesis	12 (6)	4 (4)	6 (4)	2 (1)
Genome Instability and Mutation	4 (3)	2 (2)	2 (1)	0 (1)
Resisting Cell Death	4 (4)	0 (0)	3 (3)	1 (1)
Deregulating Cellular Energetics	3 (3)	0 (0)	0 (0)	3 (3)

The patterns of cancer hallmark acquisitions seemed still heterogeneous between patients, however, there were several patterns related with phenotypes when we focus on the trunk and private subclones. In the trunk, the Inducing Angiogenesis (4 subclones were counted) were the most frequently observed cancer hallmark that was due to pathway alternations caused by VHL mutations. The second most frequently observed cancer hallmark was the Genome Instability and Mutation (2) caused by transcription factor related aberrations such as a FOXM1 Transcription Factor Network. In the private, Sustaining Proliferative Signaling (5) and Activating Invasion and Metastasis (6) were the most common event among the patients (5 out of 8 patients). In particular, untreated pa-

tients (RMH004, RMH008, and RK26) showed the Activating Invasion and Metastasis (6).

We also counted frequency of evolutionary paths of cancer hallmarks up to 2 descendants. The most frequent path is 'Normal - Inducing Angiogenesis- Deregulating Cellular Energetics' (3) and the second most frequent paths are 'Normal - Inducing Angiogenesis - Inducing Angiogenesis' (2), 'Normal - Genome Instability and Mutation - Inducing Angiogenesis' (2), 'Normal - Genome Instability and Mutation - Activating Invasion and Metastasis' (2), 'Normal - Genome Instability and Mutation - Evading Growth Suppressors' (2). These results give us biological and clinical implications beyond the SSNVs.

4 Conclusion

We developed the method of constructing personalized tumor progression models for cancer hallmark acquisition and showed effectiveness using the actual ccRCC dataset. In the example of ccRCC, identification of druggable target subclones evolved after pathway alternation of HIF activation with the VHL mutation is a clinically important problem and our model gave some implications. The cancer hallmark can help us reducing complexity of cancer development and characterazing phenotypes of the subclones. Our method effectively incorporate the cancer hallmarks into the current state-of-art tree reconstruction method of cancer subclonal evolution. As a future challenge, we construct an unified pipeline for constructing skeltons and estimating confident pathway alternation probabilities, and we extend this approach to other cancer types to reveal the phenotipic features of cancer evolution.

References

[1] Nowell PC. The clonal evolution of tumor cell populations. *Science*,194: 23–28, 1976.

[2] Gerlinger M, Rowan AJ, Horswell S, Larkin J, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N Engl J Med*,366 :883–92,2012.

[3] Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, et al. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol*,16:91,2015.

[4] Iorio F, Alonso LG, Brammeld J, Martincorena I, et al. Population-level characterization of pathway alterations with SLAPenrich dissects heterogeneity of cancer hallmark acquisition. *bioRxive*,doi: https://doi.org/10.1101/077701,2016.

[5] Caravagna G, Graudenzi A, Ramazzotti D, Sanz-Pamplona R et al. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proc Natl Acad Sci U S A.*,113:4025–4034,2016.

[6] Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*,149:979–993,2012.

[7] Gerlinger M, Horswell S, Larkin J, Rowan AJ, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet*,46:225–233,2014.

[8] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166:740–754,2016.

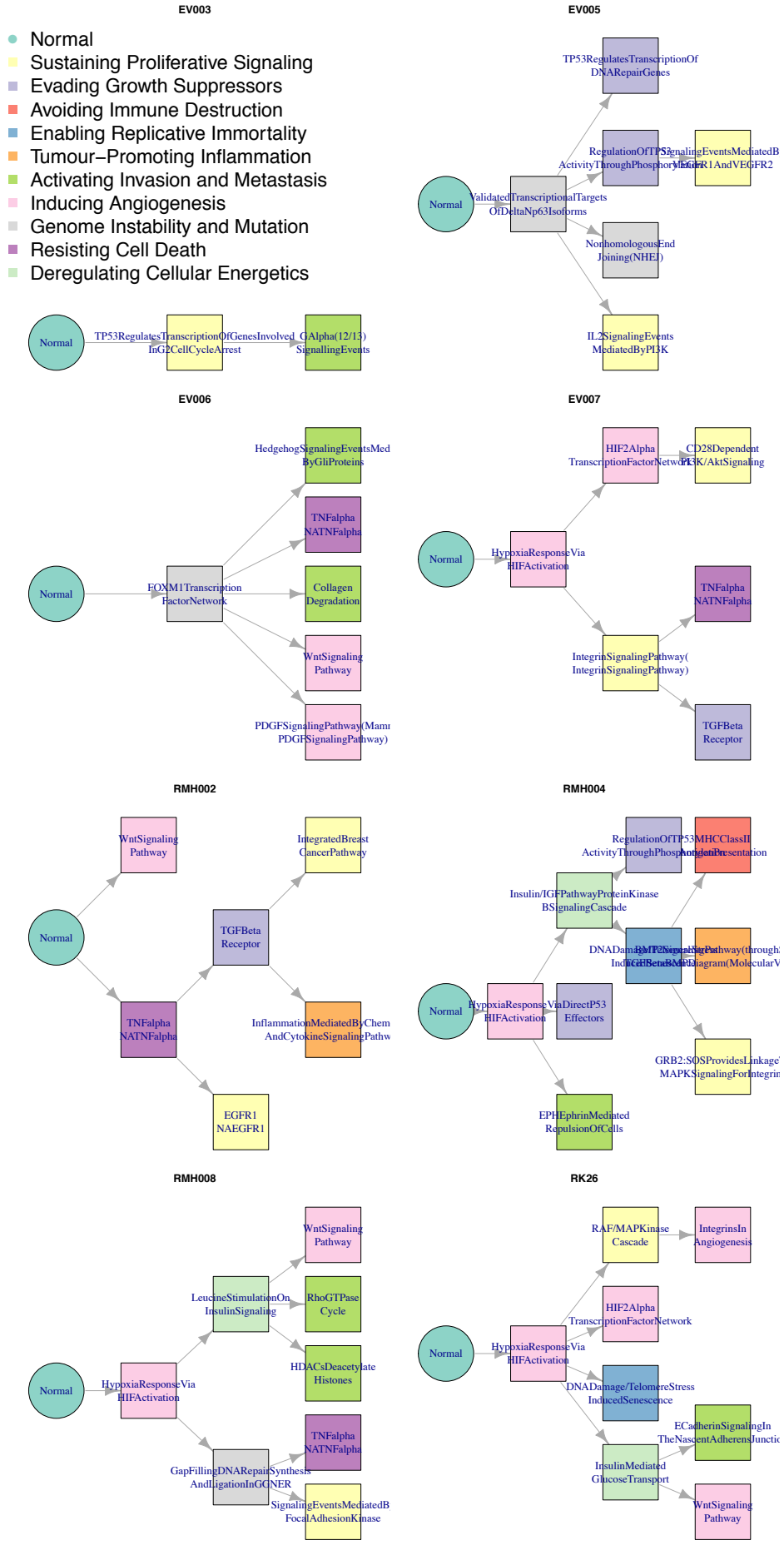

Figure 1: ccRCC progression models of cancer hallmarks acquisition. Circle and square shape indicate the normal and subclone population, respectively. Cancer hallmarks are represented as colors. Pathway names are described in the box.