

修士論文

トランスクリプトームデータを用いた肺腺癌クラスタ  
リングモデルの作成

前田 雄大

2018年2月1日

奈良先端科学技術大学院大学  
情報科学研究科

本論文は奈良先端科学技術大学院大学情報科学研究科に  
修士 (理学) 授与の要件として提出した修士論文である。

前田 雄大

審査委員：

|                       |         |
|-----------------------|---------|
| 金谷 重彦 教授              | (主指導教員) |
| 安本 慶一 教授              | (副指導教員) |
| 小野 直亮 准教授             | (副指導教員) |
| Md. Altaf-Ul-Amin 准教授 | (副指導教員) |
| 黄 銘 助教                | (副指導教員) |

# トランスクリプトームデータを用いた肺腺癌クラスタリングモデルの作成\*

前田 雄大

## 内容梗概

肺癌は、世界の癌関連死において最も一般的な原因であり、毎年 100 万人以上が死亡している。肺癌はその組織型により臨床経過が異なり、また治療に対する反応も異なるため、それらの治療にあたってより有効な分類が必要とされている。その中でも肺腺癌は最も一般的な組織学上の肺癌の分類型であり、多数の亜型が存在していることから、統一的なサブクラス分類方法が求められている。我々はここに、442 人の肺腺癌患者から採取された肺腺癌の遺伝子発現量を網羅的な測定結果であるトランスクリプトームデータから、Bayesian Hierarchical Clustering による分類を行った。その結果、11 の肺腺癌サブクラスを得た。複数のクラスターに ASCL1 や TTF1 などの臨床データと関連のある遺伝子の有意な発現変動がみられた。さらに FGA や CA12 遺伝子に肺腺癌の分類に寄与する可能性のある発現変動を発見した。

## キーワード

Bioinformatics, lung adenocarcinoma, transcriptome, Bayesian Hierarchical Clustering, Pathway analysis

---

\*奈良先端科学技術大学院大学 情報科学研究科 修士論文, NAIST-IS-MT1651096, 2018 年 2 月 1 日.

# Creation of lung adenocarcinoma clustering model using transcriptome data\*

Yuta Maeda

## Abstract

Lung cancer is the most common cause of cancer-related mortality in the world, more than 1 million people die each year. Lung cancer has different clinical course depending on its tissue type and response to treatment. So effective classification is required for more better treatment of them. Among them, lung adenocarcinoma is the most common histologically classified type of lung cancer, and since there are many subtypes, a consensus subclass classification method is required. We classified the gene expression level of lung adenocarcinoma collected from 442 lung adenocarcinoma patients by Bayesian Hierarchical Clustering from the transcriptome data, which is an exhaustive measurement result. As a result, 11 lung adenocarcinoma subclasses were obtained. Significant expression fluctuation of genes related to clinical data such as ASCL1 and TFF1 was found in multiple clusters. We also found expression fluctuations that could contribute to the classification of lung adenocarcinoma in FGA and CA12 genes.

## Keywords:

Bioinformatics, lung adenocarcinoma, transcriptome, Bayesian Hierarchical Clustering, Pathway analysis

---

\*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1651096, February 1, 2018.

# 目次

|                                                     |           |
|-----------------------------------------------------|-----------|
| <b>1. 序論</b>                                        | <b>1</b>  |
| 1.1 肺癌の組織学的な分類 . . . . .                            | 1         |
| 1.2 トランスクリプトーム解析 . . . . .                          | 4         |
| 1.3 トランスクリプトームデータによる肺腺癌の分類 . . . . .                | 4         |
| 1.4 本研究の目的 . . . . .                                | 6         |
| <b>2. 解析</b>                                        | <b>7</b>  |
| 2.1 データセット . . . . .                                | 7         |
| 2.1.1 トランスクリプトームデータ . . . . .                       | 8         |
| 2.1.2 臨床データ . . . . .                               | 10        |
| 2.2 解析を行う上での前提 . . . . .                            | 11        |
| 2.3 解析の流れ . . . . .                                 | 12        |
| <b>3. 分析方法</b>                                      | <b>14</b> |
| 3.1 Bayseian Hierarchical Clustering(BHC) . . . . . | 14        |
| 3.2 パスウェイ解析 . . . . .                               | 17        |
| <b>4. 結果と考察</b>                                     | <b>18</b> |
| 4.1 次元削減 . . . . .                                  | 18        |
| 4.2 検体のクラスタリング . . . . .                            | 19        |
| 4.2.1 クラスター1 . . . . .                              | 22        |
| 4.2.2 クラスター2 . . . . .                              | 22        |
| 4.2.3 クラスター3 . . . . .                              | 22        |
| 4.2.4 クラスター4 . . . . .                              | 23        |
| 4.2.5 クラスター5 . . . . .                              | 23        |
| 4.2.6 クラスター6と7 . . . . .                            | 24        |
| 4.2.7 クラスター8 . . . . .                              | 24        |
| 4.2.8 クラスター9 . . . . .                              | 24        |
| 4.2.9 クラスター10 . . . . .                             | 25        |

|                           |           |
|---------------------------|-----------|
| 4.2.10 クラスター 11 . . . . . | 26        |
| 4.3 パスウェイ解析 . . . . .     | 27        |
| 4.4 遺伝子のクラスタリング . . . . . | 28        |
| <b>5. 結論</b>              | <b>35</b> |
| 謝辞                        | 36        |
| 参考文献                      | 37        |
| 付録                        | 45        |

## 図 目 次

|    |                                                   |    |
|----|---------------------------------------------------|----|
| 1  | 肺がんの組織的分類 . . . . .                               | 1  |
| 2  | トランスクリプトームデータ（生データ）のヒストグラム . . . . .              | 8  |
| 3  | 対数をとったトランスクリプトームデータのヒストグラム . . . . .              | 8  |
| 4  | 想定するサンプル間で異なる肺腺癌の特徴の、トランスクリプトームデータへの反映 . . . . .  | 11 |
| 5  | 解析のアウトライン . . . . .                               | 13 |
| 6  | BHC の初期状態 . . . . .                               | 14 |
| 7  | ツリーの一部の模式図 . . . . .                              | 15 |
| 8  | BHC のアルゴリズム . . . . .                             | 16 |
| 9  | 累積寄与率のグラフ . . . . .                               | 18 |
| 10 | 検体のクラスタリング . . . . .                              | 19 |
| 11 | KEGG PATHWAY 上で 2 個以上の遺伝子が記録されていたパスウェイ . . . . .  | 27 |
| 12 | 遺伝子のクラスタリング . . . . .                             | 28 |
| 13 | ヒートマップの色と値との対応 . . . . .                          | 29 |
| 14 | サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 1 . . . . . | 30 |
| 15 | サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 2 . . . . . | 31 |
| 16 | サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 3 . . . . . | 32 |
| 17 | サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 4 . . . . . | 33 |
| 18 | サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 5 . . . . . | 34 |

## 表 目 次

|   |                                      |    |
|---|--------------------------------------|----|
| 1 | 肺腺癌の臨床治療上の分類 . . . . .               | 2  |
| 2 | WHO が定めた肺腺癌の分類 . . . . .             | 3  |
| 3 | 肺腺癌の分類とその命名 . . . . .                | 5  |
| 4 | Overall design of GSE68465 . . . . . | 7  |
| 5 | 臨床データ項目 . . . . .                    | 10 |
| 6 | 検体クラスターと臨床データの対応表 1 . . . . .        | 20 |
| 7 | 検体クラスターと臨床データの対応表 2 . . . . .        | 21 |



# 1. 序論

肺癌は報告されている中でも死亡数の多い癌であり、世界的に毎年何百万もの人間が肺癌によって死亡している [1]。疫学的な用語での肺癌は特定の疾患を指すものではなく、組織学的な肺、気管支、胸膜にできる腫瘍を指す [2]。肺癌はその組織型により臨床経過が異なり、また治療に対する反応も異なるため、それらのより治療にあたって有効な分類が必要とされている。

## 1.1 肺癌の組織学的な分類

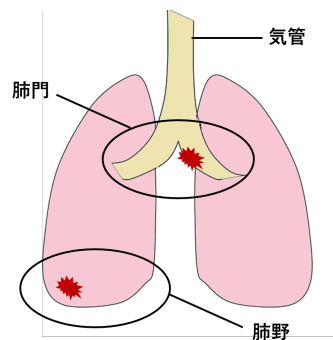


図 1 肺がんの組織的分類

赤のマークが肺癌を、丸の囲いがそれぞれ上が肺門、下が肺野の領域を指す

臨床的な観点からは、ほとんどの肺癌は小細胞肺癌 (small-cell lung carcinoma; SCLC) か非小細胞肺癌 (non-small-cell lung carcinoma; NSCLC) に組織学的な診療により分類される。この2つの区分には症状や細胞の大きさに顕著な差異がみられ、特に小細胞肺癌は放射線療法、化学療法が効果的なことから治療の方針を決めるにあたって有用なものとなっている [3][4]。非小細胞肺癌は肺癌患者の80%を占めており、治療の方針を決めるにあたって有用な組織学的な分類として管腔形成、または乳頭状増殖をする腺癌 (adenocarcinomas)、皮膚や食道の表面を被う扁平上皮組織に似た構造を示す扁平上皮癌 (squamous cell lung carcinomas; SQ)、

腺癌、扁平上皮癌のいずれの分化も示さない大細胞癌（large-cell lung carcinomas; LCLC）の3つが主要である [2]。

表 1 肺腺癌の臨床治療上の分類

|        | 組織分類  | 発生する部位 | 特徴                   |
|--------|-------|--------|----------------------|
| 非小細胞肺癌 | 腺癌    | 肺野部    | 女性の肺癌で多い<br>症状が出にくい  |
|        | 扁平上皮癌 | 肺門部    | 喫煙との関連が大きい           |
|        | 大細胞癌  | 肺野部    | 増殖が速い                |
| 小細胞肺癌  | 小細胞癌  | 肺門部    | 喫煙との関連が多きい<br>転移しやすい |

その中でも腺癌は数多くのサブタイプと組織的な変異が確認されており、WHO の最新の見解である “Histologic Typing of Lung and Pleural Tumors” によれば、診断上の分類は少なくとも 13 種類以上認められている [51]。しかし、これらの組織的な分類は常に厳密なものではなく、複数のサブタイプの特徴のいずれに近いものかを判断するのは検鏡者であるため、発見される組織型の割合は施設間で異なる。これらの検鏡者の判断による診断の揺れは治療にあたっての問題となっており、再現性のある分類基準が求められている。

これを受けて、有用な肺腺癌サブタイプを提唱するために、DNA マイクロアレイや次世代シーケンサといったスクリーニング分析に注目されている。これらのツールにより一度の分析で何千もの潜在的なバイオマーカの情報を測定することが可能となった [5]。スクリーニングの手法は特定の特徴と関連した遺伝子を探査する方法と、遺伝子発現量から支配的なパターンを探索するクラスタリングの 2 種類が存在する。腫瘍の分類においては、同じタイプの腫瘍で共通した遺伝的変異などの認められる場合がある。極端な例では、乳癌における優性遺伝子発現パターンによる分類は、エストロゲンレセプターの状態や、Her2 / neu の発現、BRCA の変異等の遺伝子型、および生存を含む臨床的に重要な指標を反映した乳癌サブタイプの再現可能性を示した [55, 56, 57]。肺腺癌においても、気管支肺胞癌（bronchioalveolar carcinoma; BAC）とその特徴を伴う腺癌と、上皮成長因子受容体（epidermal growth factor receptor; EGFR）遺伝子の突然変異が関係がみられたことから、初期癌の診断における非小細胞肺癌の分類と分子マーカとしての注目を集めている [59, 60, 61]。

表 2 WHO が定めた肺腺癌の分類

| Histologic Type and Subtypes                           | ICDO Code |
|--------------------------------------------------------|-----------|
| Adenocarcinoma                                         | 8140/3    |
| Lepidic adenocarcinoma                                 | 8250/3d   |
| Acinar adenocarcinoma                                  | 8551/3d   |
| Papillary adenocarcinoma                               | 8260/3    |
| Micropapillary adenocarcinoma                          | 8265/3    |
| Solid adenocarcinoma                                   | 8230/3    |
| Invasive mucinous adenocarcinoma                       | 8253/3d   |
| Mixed invasive mucinous and nonmucinous adenocarcinoma | 8254/3d   |
| Colloid adenocarcinoma                                 | 8480/3    |
| Fetal adenocarcinoma                                   | 8333/3    |
| Enteric adenocarcinoma                                 | 8144/3    |
| Minimally invasive adenocarcinoma                      |           |
| Nonmucinous                                            | 8256/3d   |
| Mucinous                                               | 8257/3d   |
| Preinvasive lesions                                    |           |
| Atypical adenomatous hyperplasia                       | 8250/0d   |
| Adenocarcinoma in situ                                 |           |
| Nonmucinous                                            | 8250/2d   |
| Mucinous                                               | 8253/2d   |
| Squamous cell carcinoma                                | 8070/3    |
| Keratinizing squamous cell carcinoma                   | 8071/3    |
| Nonkeratinizing squamous cell carcinoma                | 8072/3    |
| Basaloid squamous cell carcinoma                       | 8083/3    |
| Preinvasive lesion                                     |           |
| Squamous cell carcinoma in situ                        |           |

## 1.2 トランスクリプトーム解析

Transcriptome とは、転写を意味する Transcript と ome を組み合わせて作られた造語である。原則としてゲノムは同一個体のすべての細胞において共通であるが、組織ごとにタンパク質へと翻訳される領域、頻度は異なっている。そして、遺伝子から転写される mRNA の細胞内での存在量は、翻訳先のタンパク質の量と比例する。そのため、細胞中の mRNA の発現量を観測することで、様々な条件下での遺伝子発現パターンの瞬時的な状態を知ることができ、細胞の持つ機能や刺激に対する反応を、そこから類推することが可能である。その意味でトランスクリプトームデータは上流の、細胞が持つ高度な機能を実現するタンパク質ネットワークの一側面であるといえる。

DNA マイクロアレイとは遺伝子発現量を定量的に検出するための分析器具である。プローブと呼ばれる測定したい DNA と相補的なヌクレオチドの断片をプラスチックやガラス等の基板上に高密度に配置し、そこへ検体から採取した細胞から抽出した mRNA を逆転写酵素で cDNA に変換したものを流し込むことでハイブリタゼーションさせる。あらかじめ、cDNA には蛍光ラベルが施されており、プローブと結合した cDNA が多いほど、その RNA は細胞中で多く発現していたと考えられるため、チップ上のプローブの位置とその蛍光強度から RNA の発現量を評価することができる。

## 1.3 トランスクリプトームデータによる肺腺癌の分類

肺癌においては、様々な研究グループによって遺伝子発現プロファイリングが非小細胞肺癌の組織学的分類を再現できることを報告しており、いくつかの研究は腺癌がさらに分類され得ることも示した。しかし現状、これまでに報告されたこれらのサブタイプは研究によって大きく変化し、結果を一致させることや明確な結論に達することは困難である [58]。報告される肺腺癌のサブタイプの数に統一的な見解はない。もとより肺腺癌で確認されている臨床的、分子的、組織学的特徴は多種多様でありながら、新たなデータにおける確実なサブタイプの同定や異なる実験間で分類の再現性には問題が残されている。

先行研究では、DNA マイクロアレイを用いて得た遺伝子の発現量をもとに、臨床的に有意義な肺腺癌の分類を行ったことが報告されている [54]。報告では、3つのコホート研究によって得た 31, 72, 128 の計 231 人のサンプルから得た、2,553 種の遺伝子の発現量を用いて consensus clustering による分類を行い、既存研究により報告されていた疾患特異的生存率のような臨床治療において重要な要素と関連のある分類の再現に成功している。この時、提案された肺腺癌の 3つの分類は、他の肺癌の遺伝子発現パターンとの相関に因み、気管支肺胞癌 (bronchioalveolar carcinoma) に似た発現パターンを持つ bronchioid、扁平上皮癌 (sauamous cell lung cacinomas) に対する squamoid、大細胞癌 (large-cell lung carcinomas) に対する magnoid とそれぞれ命名された。これは、のちに The Cancer Gene Atlas research network ;TCGA reserch network により、より疾患の特徴と関連した名前である Terminal Respiratory Unit、Proximal Inflammatory、Proximal Proliferative へと変更することが提案されている [53]。図 3 に旧名と新名の対応を示す。

表 3 肺腺癌の分類とその命名

| 発現パターンの<br>モデルとなった肺癌                  | D. Neil Hayes 氏らにより<br>提案された肺腺癌の分類名 (2006) | TCGA Research Network<br>により提案された改名案 (2015) |
|---------------------------------------|--------------------------------------------|---------------------------------------------|
| 気管支肺胞癌<br>bronchioalveolar carcinoma  | bronchioid                                 | Terminal Respiratory Unit ; TRU             |
| 扁平上皮癌<br>sauamous cell lung cacinomas | squamoid                                   | Proximal Inflammatory ; PI                  |
| 大細胞癌<br>large-cell lung carcinomas    | magnoid                                    | Proximal Proliferative ; PP                 |

これらの分類は、“Differential pathogenesis of Lung Adenocarcinoma Subtypes Involving Sequence Mutations, Copy Number, Chromosomal Instability, and Methylation(2012)” 内で新たな実験結果による妥当性の検証と、突然変異との及び治療経過との関連性が報告された [52]。さらに、“Comprehensive molecular profiling of lung adenocarcinoma(2015)” 内で、トランスクリプトームデータに対する NMF による解析の結果、3つ以上のクラスが存在が示唆された [53]。

## 1.4 本研究の目的

本研究の目的は複数のデータベース間で同様に確認することが可能な肺腺癌の特徴および、関連する遺伝子の探索と、それによる新たな肺腺癌の分類の提唱である。そして、それらの特徴を用いての、種々の肺腺癌が持つ分子メカニズムの解明への糸口を発見することである。

## 2. 解析

### 2.1 データセット

本研究では、新たな肺腺癌の分類の提唱に GSE68465 のデータセットを用いる。GSE68465 は 442 人の患者から採取した肺腺癌の遺伝子発現を用いて、予後モデルを作成するために行われた実験のデータである。マイクロアレイ測定で得られた単独または複数の遺伝子発現に、基本的な臨床データ（ステージ、年齢、性別）を共変量として組み合わせたモデルにより、肺癌被験体の全生存期間を予測が可能かどうかを調査するために行われた実験であり [62]、肺腺癌に対する広範な病理学および臨床的注釈を有する利用可能なマイクロアレイデータの測定結果が提供されている<sup>1</sup>。

表 4 Overall design of GSE68465

|                |                                                               |
|----------------|---------------------------------------------------------------|
| n              | 442                                                           |
| Assay Type     | Gene Expression                                               |
| Provider       | Affimetrix                                                    |
| Array Designs  | HG-U133A                                                      |
| Organism       | Homo sapiens(ncbitax)                                         |
| Material Types | cell, synthetic_RNA, organism_part, whole_organism, total_RNA |
| Disease States | Lung Adenocarcinoma                                           |

---

<sup>1</sup><https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68465>

### 2.1.1 トランスクリプトームデータ

GSE68465 で公開されているトランスクリプトームデータは、DNA マイクロアレイによって測定されたものである。GSE68465 では Affimetrix 社のヒトゲノムアレイ HG-U133A が使用され、これはヒトの遺伝子を測定対象としたもので、22,283 のプローブを持つ。今回は Gene Symbol が参照可能な 21225 のプローブを用いて解析を行った。

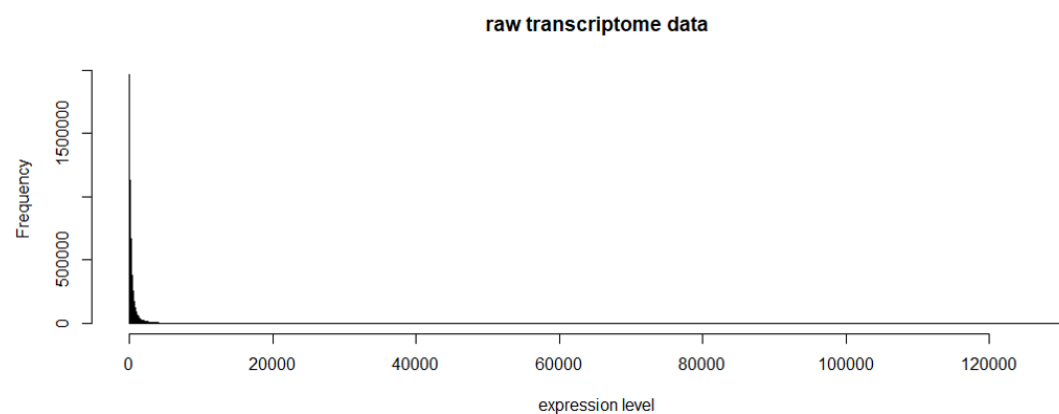


図 2 トランスクリプトームデータ（生データ）のヒストグラム

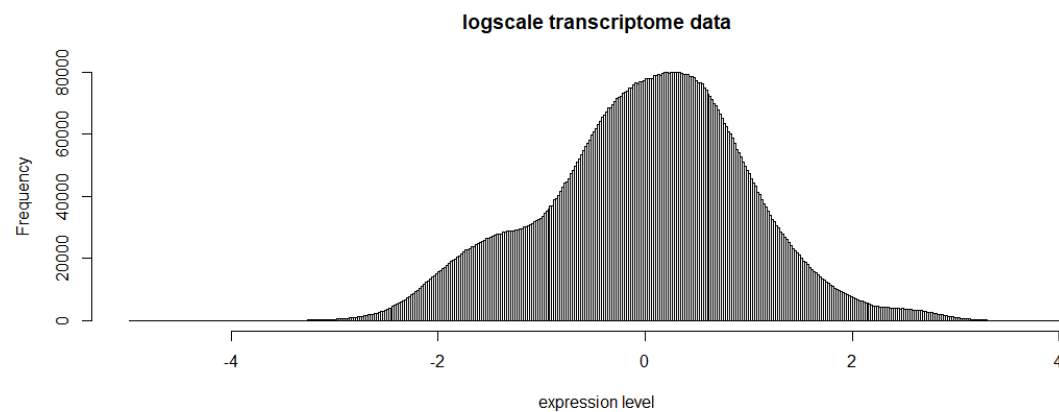


図 3 対数をとったトランスクリプトームデータのヒストグラム



図2に蛍光強度そのものの値のトランスクリプトームデータのヒストグラムを、  
図3に前処理として2の対数をとったものを示す。

### 2.1.2 臨床データ

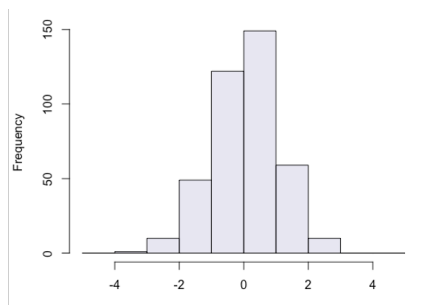
各サンプルは以下の表 5 に示す 11 の臨床データと紐づけられている。仮にサンプル間のこれらのパラメータの違いが肺腺癌に由来するものであれば、その特徴はタンパク質ネットワークの活性の違いとして、ひいてはトランスクリプトームデータ上に何らかの形で現れていることを期待する。

表 5 臨床データ項目

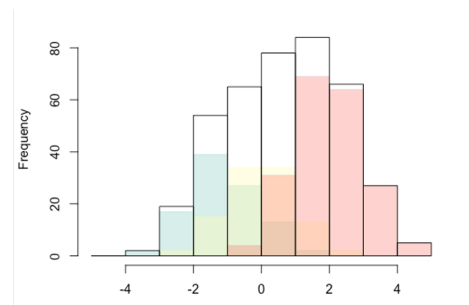
| 項目名                          | 備考                                    |
|------------------------------|---------------------------------------|
| sex                          |                                       |
| age                          |                                       |
| race                         |                                       |
| status after 5 years         | 5 年後生存確認の結果                           |
| adjuvant chemo               | アジュバント化学療法を受けたか                       |
| adjuvant radithional therapy | アジュバント放射線治療を受けたか                      |
| desease stage                | 数字は癌の進行度<br>T は癌の面積、N はリンパ節転移の度合いを示す。 |
| first progression            | 再発か否か、NO は再発                          |
| smoking history              | 喫煙してるか、していたか                          |
| histlogical grade            | 癌の分化の度合い                              |

## 2.2 解析を行う上での前提

本節では、クラスタリングを行う上で、標的となる肺腺癌のトランスクリプトームデータに仮定する3つの条件を述べる。まず、1つ目の仮説は、癌の持つ特徴、その背景にあるメカニズムの全てはタンパク質ネットワークを通して mRNA の発現量に反映されていること。次に、トランスクリプトームデータ測定時に乗ったノイズは正規分布に従うこと。最後に、同様な特徴、メカニズムの上に成り立つ肺腺癌の遺伝子発現パターンは、異なる検体であっても同じであること。この3つの仮定の下では、すべてのサンプルに共通して同様な発現パターンの遺伝子がある場合、その発現量の分布は図 (a) に示すような正規分布に従う。その反対に、異なる発現パターンを持つサンプル群が存在する場合、その発現パターンはその種類の数、図 (b) に示すような異なる平均値を持つ混合正規分布に従う。



(a) 1種類の発現パターンを持つ遺伝子の場合



(b) 複数の発現パターンを持つ遺伝子の場合

図 4 想定するサンプル間で異なる肺腺癌の特徴の、トランスクリプトームデータへの反映

そして、今回解析に用いるデータはすべて肺腺癌患者から採取したものであるため、すなわちこの集団内にある発現パターンの違いは全て癌を特徴づけるものであるということが、解析を進めるうえでの前提となる。

## 2.3 解析の流れ

研究のアウトラインを図5に示す。はじめに、サンプル同士を比較するために、トランスクリプトームデータに対して検体ごとに平均を0に分散1に標準化を施す。次に、各遺伝子ごとに Shapiro-Wilk 検定によりサンプル間の遺伝子発現量に正規性があるかを検定する。仮に正規性が認められた場合、その遺伝子はサンプル集団内で単一の発現パターンしかもたず、サンプル間の分類に用いるには適当でないと判断しデータセットから除外する。そして、サンプル集団内で発現パターンが複数存在する見込みの高い遺伝子のみを用いて、Principal Component Analysis (PCA) による次元削減を行う。今回扱うトランスクリプトームデータは442サンプルに対して21,225もの次元を持つが、癌にかかわりがある遺伝子は約25,000あるヒト遺伝子の1%以上との予測があり、21,225は多すぎる[49]。そのため、PCAによりサンプルの特徴量を主成分にまとめたのち、各主成分の寄与率で重みづけをした割合で、因子負荷量の高い遺伝子から順に選定していき、“A census of human cancer genes”で示されたタンパク質をコードするがん遺伝子の数である221を踏まえて、それより少し多い300次元以下にまで次元を減らす[49]。その後、抽出した遺伝子を用いて、サンプル間でBayesian Hierarchical Clustering(BHC)によるクラスタリングを行う。

次に、遺伝子同士でのBHCによるクラスタリングを行い、サンプル間でのクラスタツリーを用いてヒートマップを作成する。その後、遺伝子をGO解析及びパスウェイ解析にかけ、事前のサンプルの分類結果とヒートマップとを合わせて、遺伝子のクラスターの持つ意味を考察する。

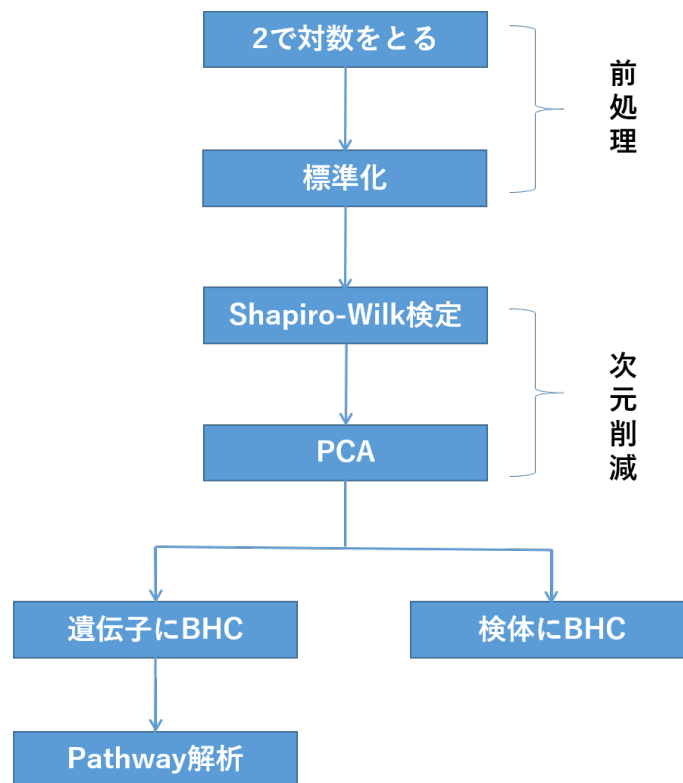


図 5 解析のアウトライン

### 3. 分析方法

#### 3.1 Bayesian Hierarchical Clustering(BHC)

Bayesian Hierarchical Clustering (BHC) は階層型クラスタリングの一種であり、ツリーをマージする際の評価値として確率モデルとマージ前後のデータ分布間との尤度が採用されている。 $n$  個のデータを持つデータセットの全体集合  $D = \{x^{(1)}, \dots, x^{(n)}\}$  を定義する。ここで、 $D_i \subset D$  は部分木  $T_i$  を構成する葉ノードを意味する。このアルゴリズムの初期状態では、図??に示すように  $n$  本の木  $\{T_i : i = 1 \dots n\}$  が1つのデータ  $D_i \{x^{(i)}\}$  を葉として持つ。

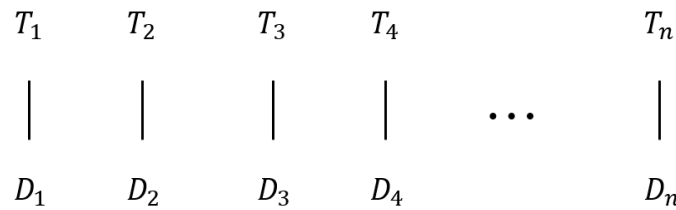


図 6 BHC の初期状態

ツリーをマージする際には、その時点で存在するすべてのツリーの組み合わせを想定する。図7に表す例として、 $T_i$  と  $T_j$  をマージすると、新しいツリー  $T_k$  となる。このとき、マージされる前の2つのツリーに属していたデータは新たなデータ集合  $D_k = D_i \cup D_j$  として統合される。

それぞれのマージを考えるにあたって、2つの仮説を比較する。1つ目の仮説は  $H_1^k$  は「 $D_k$  に属するすべてのデータは独立同分布であり、不明なパラメータ  $\theta$  を持つ確率モデル  $p(x|\theta)$  から生成されている。」である。ここでこの確率モデルが  $\theta = (\mu, \Sigma)$  をパラメータとした多変量ガウス分布であるとした場合を想定すると、異なる確率モデルがより適切であるかもしれないことを強調することもまた重要である。この仮説のもとでの確率を求めるために、確率モデルのパラメータ  $\theta$  の確率モデル  $p(\theta|\beta)$  パラメータであるハイパーパラメータ  $\beta$  を決める必要がある。ここに、仮説  $H_1^k$  のもとで  $D_k$  が生成される確率  $p(D_k|H_1^k)$  を計算するための要素がそろい、それは式(1)によって求まる。

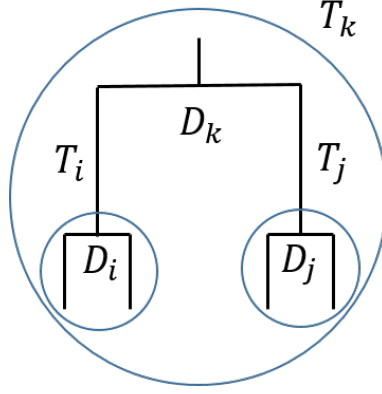


図 7 ツリーの一部の模式図

$$\begin{aligned}
 p(D_k|H_1^k) &= \int (D_k|\theta)p(\theta|\beta)d\theta \\
 &= \int \left[ \prod_{x^{(i)} \in D_k} p(x^{(i)}|\theta) \right] p(\theta|\beta)d\theta
 \end{aligned} \tag{1}$$

この値が、データが1つのクラスターにどれだけ適合するかを測定するモデルに基づいた基準となる。 $H_1^k$  の対立仮説は、「 $D_k$  は2つかそれ以上のクラスターを持つ」である。本来  $H_1^k$  のもとでの、 $D_k$  が生成される確率を求めるには、指数関数的な数となる  $D_k$  を2つ以上のクラスターに分割する全ての組み合わせを考慮して、計算する必要がある。しかし、 $D_k$  を分割する方法を部分木  $T_i$  と  $T_j$  に限定することにより、再帰を用いることで効率的に計算することが可能である。この制限された対立仮説  $H_2^k$  のもとで  $D_k$  が生成される確率は  $p(D_k|H_2^k) = p(D_i|T_i)p(D_j|T_j)$  と求めることが可能である。 $H_1^k$  と  $H_2^k$  のもとでのデータ  $D_k$  の、 $D_k$  内のデータがすべて1つのクラスターに属するという前提で重みづけされた条件付き確率を組み合わせることで、 $T_k$  に属するデータの周辺確率が以下の式2で求まる。なお、 $\pi_k = p(H_1^k)$  である。

$$p(D_k|T_k) = \pi_k p(D_k|H_1^k) + (1 - \pi_k) p(D_i|T_i) p(D_j|T_j) \tag{2}$$

この方程式は再帰的に定義され、第1項は  $D_k$  が1つのクラスター内にあるという仮説を、第2項は  $D_k$  内のツリー構造である、他のすべてのクラスターをに効

率的に合計したものを表している。マージされた場合の事後確率  $r_k = p(H_1^k | D_k)$  はベイズの法則を用いて式3により得られる。

$$r_k = \frac{\pi_k p(D_k | H_1^k)}{\pi_k p(D_k | H_1^k) + (1 - \pi_k) p(D_i | T_i) p(D_j | T_j)} \quad (3)$$

この方程式はマージすべき2つのツリーの決定に使用され、さらにどのタイミングでマージを終了すべきだったかを確認する際にも使われる。マージする前後でマージした後のほうがモデルに対する尤度が低い時、式3より  $r_k < 0.5$  となり、この時マージを終了し残っていた木の数が BHC により得たクラスターの数となる。アルゴリズムは図8に示される。

```

input: data  $\mathcal{D} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ , model  $p(\mathbf{x}|\theta)$ ,
        prior  $p(\theta|\beta)$ 
initialize: number of clusters  $c = n$ , and
         $\mathcal{D}_i = \{\mathbf{x}^{(i)}\}$  for  $i = 1 \dots n$ 
while  $c > 1$  do
    Find the pair  $\mathcal{D}_i$  and  $\mathcal{D}_j$  with the highest
    probability of the merged hypothesis:
        
$$r_k = \frac{\pi_k p(\mathcal{D}_k | \mathcal{H}_1^k)}{p(\mathcal{D}_k | T_k)}$$

    Merge  $\mathcal{D}_k \leftarrow \mathcal{D}_i \cup \mathcal{D}_j$ ,  $T_k \leftarrow (T_i, T_j)$ 
    Delete  $\mathcal{D}_i$  and  $\mathcal{D}_j$ ,  $c \leftarrow c - 1$ 
end while
output: Bayesian mixture model where each
        tree node is a mixture component
    The tree can be cut at points where  $r_k < 0.5$ 

```

図8 BHC のアルゴリズム

本研究では、R の BHC パッケージ `bclust`<sup>2</sup>を用いて解析を行った。事前分布は解析を行う上での前提に従って正規分布とした。

<sup>2</sup><https://cran.r-project.org/web/packages/bclust/index.html>



### 3.2 パスウェイ解析

任意の方法で抽出した発現変動遺伝子の機能や代謝経路などを調べる方法として、パスウェイ解析があげられる。パスウェイ解析では、代謝経路や質相互作用、遺伝子発現制御、共発現等のタンパク質や化合物等の分子間相互作用のネットワークをデータベース化した、パスウェイデータベースを用いて特定の遺伝子群がどのパスウェイに多く含まれているかを調べる。その際、特定のパスウェイに含まれていたことが統計学的に有意である事象か否かの検定を行う。

本研究では、パスウェイデータベースに KEGG PATHWAY を使用し、解析ツールに DAVID(The Database for Annotation, Visualization and Integrated Discovery) を用いてパスウェイ解析を行う。KEGG PATHWAY は web で公開されているバイオインフォマティクス研究用のデータベース、KEGG(Kyoto Encyclopedia of Genes and Genomes) のコアとなる分子間相互作用ネットワークのパスウェイマップデータベースである。パスウェイマップには遺伝子、化合物、化学反応、疾患情報など KEGG 内の他のデータベースのタームが関連付けられている。KEGG パスウェイでは代謝経路、遺伝情報処理、環境情報処理、細胞内プロセス、生体システム、疾患、創薬の分子間相互作用ネットワークのパスウェイが提供されている。また、DAVID は与えられた遺伝子群のアノテーションから、KEGG を含む様々なデータベースごとにそのアノテーションの組み合わせの偶然性を検定し、FDR や遺伝子が記録されていたパスウェイの数などの条件を満たすものをリストにして返す、web アクセス可能なプログラムである。

## 4. 結果と考察

### 4.1 次元削減

まず、発現に変動が見られる遺伝子をスクリーニングするため、Shapiro Wilk 検定による検定を行った。False Discovery Rate(FDR)  $q < 0.01$  として検定を行ったところ、21,225 ある遺伝子の発現量のうち 16,573 の遺伝子では正規分布から外れていることが認められ、4,652 の遺伝子は正規分布のノイズにのっという判定されたため、正規分布から外れている 16,573 の遺伝子について解析を行った。前述の特徴量を用いて PCA による解析を行った。図 9 に主成分と累積寄与率のグラフより累積寄与率 80%が達する第 1~230 主成分を用いて、さらに次元削減を行う。次元削減の目標である 300 に寄与率で重みづけをして、それぞれの主成分で因子負荷量の絶対値が大きい順に遺伝子を選定した。結果、選定された遺伝子に重複があったため 212 個の遺伝子が抽出された。これらの遺伝子は正規分から外れており、同じ肺腺癌患者内での遺伝子発現量において分散が高いものである。

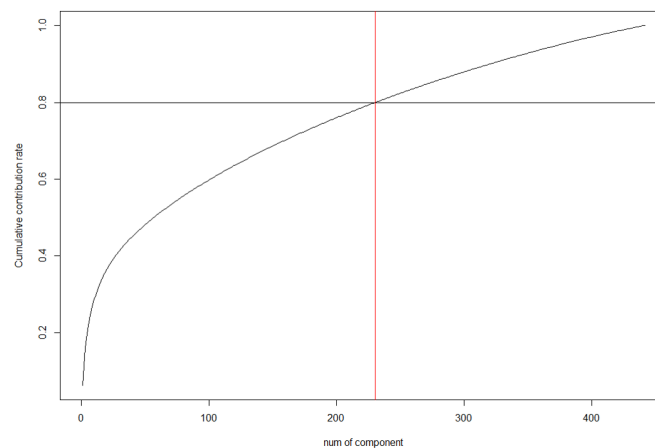


図 9 累積寄与率のグラフ

## 4.2 検体のクラスタリング

BHC によるサンプルごとのクラスタリングを行った。結果を図 10 に示す。縦軸は、クラスターをマージする前後で減少する事前分布モデルとの尤度を表しており、横軸はクラスターに属するサンプルの数に比例している。式 4 より  $r_k = 0.5$  となるマージ、すなわち木全体の尤度が最大になる閾値は高さ 35990 と判定された。これによって検体は 11 のクラスターに分類された。各クラスターと臨床データとの対応を表 6 及び表 7 に示す。BHC によって得た個々のクラスターは、事前分布である多変量正規分布から生成されたものとして解釈することが可能であり、また、これ以上マージするとそのモデルに対する尤度が下がるため、これらのクラスは独立したものと考えるべきである。

$$height = |\pi_k p(D_k | H_1^k) - (1 - \pi_k) p(D_i | T_i) p(D_j | T_j)| \quad (4)$$

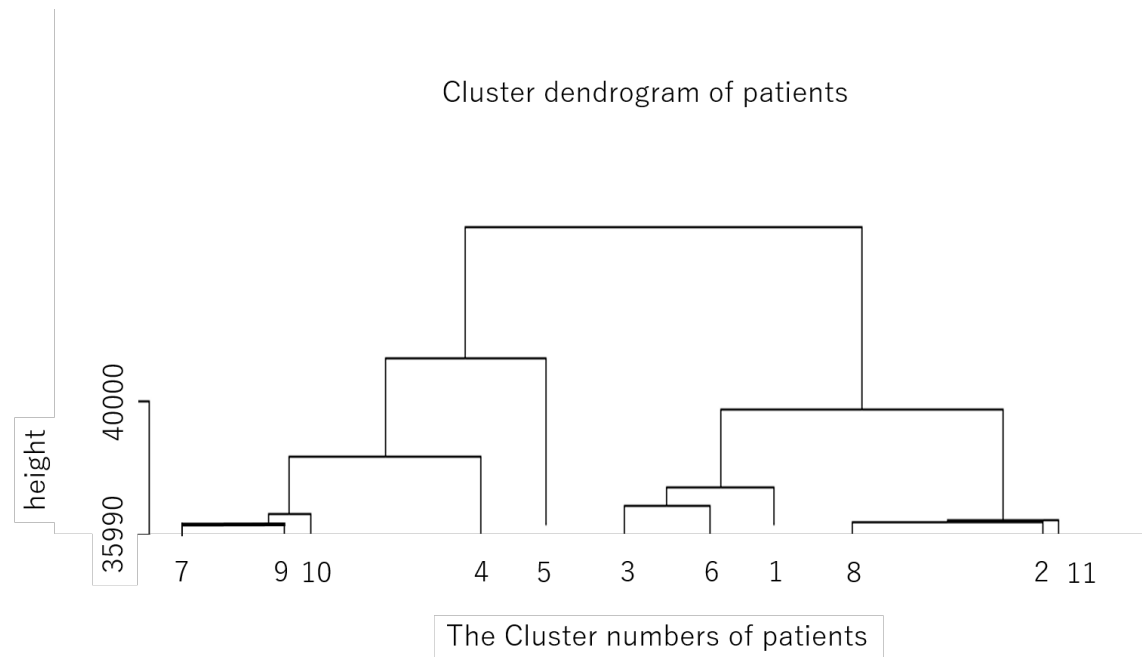


図 10 検体のクラスタリング

表 6 検体クラスターと臨床データの対応表 1

|                            |              | total | patients cluster number |    |    |    |    |    |    |    |    |    |    |
|----------------------------|--------------|-------|-------------------------|----|----|----|----|----|----|----|----|----|----|
|                            |              |       | 1                       | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 |
| number of patients         |              | 442   | 46                      | 97 | 18 | 86 | 54 | 43 | 39 | 26 | 13 | 13 | 7  |
| age                        |              | 64.5  | 68                      | 65 | 62 | 64 | 62 | 64 | 63 | 69 | 62 | 67 | 64 |
| SEX                        | Female       | 220   | 16                      | 72 | 8  | 33 | 24 | 25 | 18 | 12 | 5  | 4  | 3  |
|                            | Male         | 222   | 30                      | 25 | 10 | 53 | 30 | 18 | 21 | 14 | 8  | 9  | 4  |
| RACE                       | Asian        | 6     | 0                       | 1  | 0  | 1  | 1  | 3  | 0  | 0  | 0  | 0  | 0  |
|                            | Black        | 12    | 1                       | 0  | 0  | 5  | 1  | 2  | 2  | 1  | 0  | 0  | 0  |
|                            | Hawaiaa      | 1     | 0                       | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  |
|                            | Not Reported | 45    | 2                       | 12 | 2  | 10 | 4  | 1  | 5  | 3  | 0  | 1  | 5  |
|                            | Unknown      | 83    | 8                       | 13 | 3  | 19 | 9  | 4  | 13 | 2  | 2  | 8  | 2  |
|                            | White        | 295   | 35                      | 71 | 13 | 51 | 39 | 33 | 18 | 20 | 11 | 4  | 0  |
| Status after 5 years       | Alive        | 207   | 24                      | 54 | 8  | 35 | 29 | 18 | 14 | 15 | 5  | 3  | 2  |
|                            | Dead         | 235   | 22                      | 43 | 10 | 51 | 25 | 25 | 25 | 11 | 8  | 10 | 5  |
| adjuvant chemo             | No           | 341   | 37                      | 79 | 14 | 61 | 40 | 30 | 32 | 22 | 9  | 11 | 6  |
|                            | Unknown      | 12    | 1                       | 1  | 0  | 6  | 1  | 0  | 2  | 0  | 0  | 0  | 1  |
|                            | Yes          | 89    | 8                       | 17 | 4  | 19 | 13 | 13 | 5  | 4  | 4  | 2  | 0  |
| adjuvant radiation therapy | No           | 364   | 41                      | 85 | 14 | 63 | 43 | 33 | 33 | 24 | 10 | 12 | 6  |
|                            | Unknown      | 13    | 1                       | 1  | 0  | 7  | 1  | 0  | 2  | 0  | 0  | 0  | 1  |
|                            | Yes          | 65    | 4                       | 11 | 4  | 16 | 10 | 10 | 4  | 2  | 3  | 1  | 0  |

表 7 検体クラスターと臨床データの対応表 2

|                  |          | total | patients cluster number |    |    |    |    |    |    |    |    |    |    |
|------------------|----------|-------|-------------------------|----|----|----|----|----|----|----|----|----|----|
|                  |          |       | 1                       | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 |
| disease stage    | pN0pT1   | 114   | 16                      | 39 | 2  | 14 | 10 | 10 | 10 | 10 | 0  | 2  | 1  |
|                  | pN0pT2   | 162   | 18                      | 26 | 10 | 28 | 26 | 18 | 16 | 11 | 2  | 4  | 3  |
|                  | pN0pT3   | 16    | 0                       | 3  | 2  | 2  | 3  | 2  | 0  | 0  | 1  | 2  | 1  |
|                  | pN0pT4   | 7     | 0                       | 2  | 1  | 2  | 1  | 0  | 1  | 0  | 0  | 0  | 0  |
|                  | pN1pT1   | 24    | 3                       | 5  | 0  | 7  | 1  | 2  | 2  | 2  | 1  | 1  | 0  |
|                  | pN1pT2   | 54    | 4                       | 10 | 2  | 12 | 9  | 5  | 4  | 0  | 5  | 1  | 2  |
|                  | pN1pT3   | 7     | 0                       | 0  | 0  | 3  | 0  | 2  | 2  | 0  | 0  | 0  | 0  |
|                  | pN1pT4   | 2     | 1                       | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|                  | pN2pT1   | 11    | 1                       | 3  | 0  | 3  | 1  | 1  | 0  | 1  | 0  | 1  | 0  |
|                  | pN2pT2   | 34    | 3                       | 8  | 0  | 10 | 2  | 3  | 3  | 0  | 3  | 2  | 0  |
|                  | pN2pT3   | 5     | 0                       | 0  | 0  | 3  | 0  | 0  | 0  | 1  | 1  | 0  | 0  |
|                  | pN2pT4   | 3     | 0                       | 0  | 1  | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 0  |
|                  | pNXpT1   | 1     | 0                       | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|                  | pp       | 2     | 0                       | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 0  |
| --               | 71       | 5     | 15                      | 2  | 20 | 4  | 10 | 9  | 2  | 1  | 3  | 0  |    |
| 1st progression  | No       | 157   | 20                      | 37 | 7  | 23 | 23 | 16 | 11 | 12 | 5  | 2  | 1  |
|                  | Unknown  | 10    | 1                       | 2  | 1  | 0  | 2  | 1  | 0  | 2  | 1  | 0  | 0  |
|                  | Yes      | 204   | 20                      | 43 | 8  | 43 | 25 | 16 | 19 | 10 | 6  | 8  | 6  |
| smoking history  | --       | 2     | 0                       | 0  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|                  | smoking  | 32    | 3                       | 3  | 2  | 2  | 8  | 7  | 3  | 1  | 1  | 2  | 0  |
|                  | Never    | 49    | 0                       | 22 | 3  | 7  | 1  | 1  | 2  | 10 | 0  | 2  | 1  |
|                  | Smoked   | 268   | 35                      | 54 | 10 | 51 | 41 | 24 | 17 | 15 | 11 | 4  | 6  |
|                  | Unknown  | 91    | 8                       | 18 | 3  | 24 | 4  | 11 | 17 | 0  | 1  | 5  | 0  |
| histologic grade | --       | 7     | 0                       | 1  | 0  | 0  | 4  | 1  | 0  | 1  | 0  | 0  | 0  |
|                  | Moderate | 209   | 27                      | 49 | 11 | 38 | 24 | 19 | 18 | 15 | 7  | 0  | 1  |
|                  | POORLY   | 166   | 5                       | 23 | 0  | 44 | 24 | 23 | 21 | 3  | 4  | 13 | 6  |
|                  | WELL     | 60    | 14                      | 24 | 7  | 4  | 2  | 0  | 0  | 7  | 2  | 0  | 0  |

#### 4.2.1 クラスター 1

クラスター 1 はトランスクリプトームデータ上では、PCSK2 の高発現 ( $p < 1.0^{-9}$ ) が見られた。PCSK2 は主に神経内分泌特徴を有する癌、特に小細胞癌において検出される [10, 11, 12, 13]。WIF1 もまた、他クラスと比較して高い発現を示していた ( $p < 1.0^{-9}$ )。分泌型糖タンパク質のウイングレス型 (Wnt) ファミリーは、発達プロセスおよび発癌に広く関与するシグナル伝達分子のグループであり [17, 18]、WIF1 はこの経路の活性を抑制する働きを持っている [16]。

#### 4.2.2 クラスター 2

クラスター 2 は 74% が女性で占められるクラスターである ( $p = 1.2^{-5}$ )。このクラスターでは RPS4Y1 に低い発現が確認され ( $p = 7.3^{-7}$ ) た。RPS4Y1 は Y 染色体上に存在する遺伝子であり、複数の組織の癌で正常な細胞との発現変動が報告されている [26]。また、X 染色体上に存在する XIST には高い発現がみられた ( $p = 5.1^{-10}$ )。XIST は正常細胞においては男女間で発現量に有意な差はみられないが、肺癌を含む複数の癌では発現変動がみられることが報告されている [26]。

このクラスは 11 種類のクラスの中で最も大きいクラスであり、また性別に偏りがみられる。

#### 4.2.3 クラスター 3

このクラスターでは、腫瘍の分化度との相関性がみられる TFF1 の高い発現がみられた ( $p=8.655e-11$ ) [?]。また、OLFM4 が高い発現を示していることも、このクラスターの特徴である ( $p = 3.1^{-7}$ )。OLFM4 は喫煙者の非小細胞肺癌において、その低発現と予後不良との相関が報告されている [28]。PSCA もまた、高い発現を示しており ( $p = 2.0^{-7}$ )、この遺伝子の生成物はキメラ抗原受容体 T 細胞の標的分子として注目されている [29]。さらに、MIA の発現量が高いことも注目すべき点であり ( $p=4.596e-09$ )、この遺伝子は癌の中でも特に扁平上皮細胞癌での発現が高く、診断上及び分子標的として有用であると示唆されている [30]。

上記の特徴から、このクラスは肺腺癌のなかでも扁平上皮癌に似た特徴を持つクラスだと推察する。肺腺癌における PCSA と MIA の発現量に相関がみられれば、今後の創薬において役立つであろうことが予測される。

#### 4.2.4 クラスター 4

このクラスターにおいて特徴的なふるまいは、TFPI2 および XAGE1B がクラスター 3 のそれと対照的な発現を示している点である。TFPI2 の低発現は、非小細胞肺癌の進行に寄与していることが示唆されている [31]。また、非小細胞肺癌における自発的免疫応答を分析結果から、XAGE1B に対する抗体応答は、69 人の IIIB / IV 期の肺腺癌患者の 19% に相当する 13 人で観測されたことが報告されている。

このクラスターは全体の 19% に相当し、また XAGE1B の発現量も高いことから先の研究結果と符合する可能性がある。しかし、このクラスターにおける IIIB / IV 期の肺腺癌患者は全体の 5% にも満たないため、免疫治療の早期診断の観点から XAGE1B は早期の肺腺癌においても抗体応答が確認できるか検証する価値があると考ええる。

#### 4.2.5 クラスター 5

クラスター 5 は ASCL1、CALCA の高い発現がみられた ( $p < 1.0^{-9}$ ,  $p < 1.0^{-9}$ )。ASCL1 は小細胞肺癌の悪性挙動および生存を促進する上で重要な役割を果たすことが報告されている [33]。また、CALCA は胎児及び胎児の成長において重要な役割を担っていると考えられている遺伝子である [34]。さらに、血管緊張 [35]、細胞増殖および分化の役割があるとされている [36]。

このクラスターは喫煙者に多く見られる小細胞肺癌と関連した遺伝子が他のクラスターにない特徴的な発現を示していた。また、このクラスターに属する患者の 83% が喫煙者、あるいは喫煙していたことから、このクラスの肺癌は小細胞肺癌の発生プロセスとのかかわりがあると考えらる。

#### 4.2.6 クラスター 6 と 7

この2つのクラスに関しては他のクラスと比べて特徴的な発現パターンを見出すことはできなかった。これらのクラスに関しては、ほかのクラス全体ではなく BHC において最後のマージとなる 4,5,7,9,10 のグループと、1,2,3,6,8,11 のグループとで比較することが適切だと考える。

#### 4.2.7 クラスター 8

クラスター 8 では CCL20 のダウンレギュレートが確認された ( $p = 3.5^{-8}$ )。CCL20 遺伝子の生産物は炎症細胞の動員に関与し、ERK1 / 2-MAPK および PI3K 経路の活性化を介して肺がん細胞の転移および増殖を促進することが報告されている [37]。また、MMP1 のダウンレギュレートも確認されており ( $p < 1.0^{-9}$ )、この遺伝子の高発現は STAT3 の活性と相関し、非小細胞肺癌の悪性特性を促進することが報告されている [38]。

クラスター 8 は癌の悪性度とその発現とに相関があるとされる遺伝子の発現が低かったことから、このクラスの肺腺癌は悪性度の低いものだと推察される。このクラスの 5 年後生存率は、他のクラスと比べて最も高い 58% であるが、有意な差はない ( $p=0.65$ )。

#### 4.2.8 クラスター 9

クラスター 9 では MUC5B に高い発現がみられた ( $p = 8.1^{-9}$ )。原発性 EGFR 変異体非小細胞肺癌における MUC5B の発現が、患者のより長い生存と関連していることが報告されている [39]。また、AKR ファミリーに属する AKR1C1-3 ( $p = 1.2^{-6}, p = 2.5^{-6}, p = 1.6^{-5}$ ) にも、高い発現がみられた。この遺伝子は、多くの癌で調節不全となる NRF2 パスウェイの過剰活性との関連が報告されている [40]。

クラスター 9 では、突然変異した系との関連を示す遺伝子が高い発現を示していた。癌との関連を示す文献こそ見つからなかったが、FGA および CA12 においても、このクラスターは高い発現を示してた ( $p = 3.3^{-6}, p = 9.2^{-8}$ )。上述した



MUC5B はクラスター 3,6 でも高い発現がみられ ( $p = 3.0^{-7}, p = 1.2^{-8}$ )、AKR1C1 はクラスター 1 においても高い発現がみられた ( $p < 1.0^{-9}$ )。しかし、FGA および CA12 はこれらのクラスターとも異なる発現を示していることから、ここにクラスを特徴づけるメカニズムが存在することが推察される。

#### 4.2.9 クラスター 10

クラスター 10 は癌の分化度が低いサンプルの集まりである ( $p = 0.0052$ )。このクラスは他のクラスと比べて SPINK1 TMPRSS4 と発現量が低いことが特徴としてあげられる ( $p = 6.3^{-7}, p = 2.0^{-7}$ )。SPINK1 は数多くの癌から過剰発現が報告されており、尿や血中から採取されることからバイオマーカーとして注目されている [41]。この遺伝子のノックダウンは腫瘍の成長、肺への血管壁浸潤及び転移を減少させることが報告されている [42]。TMPRSS4 は、非小細胞肺癌の癌細胞において癌幹細胞様特性を誘導することが報告されている [43]。

クラスター 10 においてみられる分化度の低さは、癌の成長、分化に関わる遺伝子である SPINK1 や TMPRSS4 の低さに起因するものと思われる。しかしながら、このクラスの 5 年後生存率は良くはなく、先に挙げた 2 つの遺伝子とは別の遺伝子がかかわる、致死的なメカニズムが動作している可能性がある。そこで、このクラスで発現の高い遺伝子を調べたところ PEG10 と MEGEA6 が該当した ( $p = 0.0047, p = 2.9^{-05}$ )。PEG10 は、癌細胞におけるアポトーシス耐性にとって重要な遺伝的インプリンティング遺伝子として同定されており、PEG10 の発現が臨床的 TNM グレードおよび肺癌における患者の予後と密接に関連していることが報告されている。また、MAGEA6 遺伝子の異常な発現が、様々な腫瘍型において予後の不良と相関することが報告されている [45]。

#### 4.2.10 クラスター 11

クラスター 11 では、TENM1 が低い発現を示した ( $p = 0.0010$ )。Teneurin Transmembrane Protein (TENM) は、形態形成および中枢神経系の発達に参与する高度に保存されたペアルールタンパク質ファミリーであり。成体組織および疾患におけるそれらの機能は、ほとんど知られていない。最近の調査によると、種々の癌細胞株および卵巣腫瘍組織における TENM2 および TRNM4 の発現を調べたところ、TENM4 発現が低い卵巣腫瘍は、より分化した表現型を示さず、これらの患者は平均全生存期間がより短かったことが報告されている [20]。また、低 TENM4 発現が癌細胞に増殖上の利点を提供し、薬剤耐性および臨床的攻撃性の増大を特徴とする未分化状態を示すことも示唆されている [20]。

また、ZNF83 も低い発現を示している。これは前立腺癌の進行において重要な機能を持つことが示唆されており [21]、結腸直腸癌 [23] と肝細胞癌 [22] においても異常な発現が報告されている。これらの癌に共通する特異的なふるまいは、また一部の肺腺癌にもみられる可能性がある。

このクラスターは分化度の低さから、臨床的、組織学的な特徴を発見することが困難なタイプと推察する。しかし、トランスクリプトームデータでは TENM1 や ZNF83 の低発現や DNA の修復に関連する USP34 [25]、びまん性胃がんで異常な発現がみられる NOL8 [24] 等が他のクラスターにないふるまいをしていることから、このクラスの同定は容易であると思われる。複数の癌でみられる特徴的な発現がこのクラスにもみられることから、また、分化が活発でないことからこのタイプの癌は細胞内で普遍的なメカニズムに起因するものであると推察する。

### 4.3 パスウェイ解析

次元削減により抽出した遺伝子群から DAVID を用いてパスウェイ解析を行った。KEGG PATHWAY に記録されているパスウェイから、189 の遺伝子から、2 個以上の遺伝子が存在するパスウェイを抽出した。その結果、18 種のパスウェイが検出された。表 11 に結果を示す。

Metabolism of xenobiotics by cytochrome P450 は癌と関りのあるパスウェイである。CYP（シトクロム P450）は、多環芳香族炭化水素、複素環式アミン、ニトロソアミン、アゾ染料、およびアルキル化剤などの異なる発癌物質を含む、内因性および外因性化合物の代謝および排泄において重要な役割を果たす [46, 47]。この経路で CYP1A1、CYP2A6、CYP2E1、および CYP3A4 / 5 によって産生される化合物の代謝中間体は、しばしば非常に活性であり、腫瘍発生および進行の開始および促進に関連することが報告されている [46, 47, 48]。

| Term                                                  | Count | %        | PValue   | FDR     |
|-------------------------------------------------------|-------|----------|----------|---------|
| hsa00140:Steroid hormone biosynthesis                 | 8     | 4.232804 | 2.82E-06 | 0.00328 |
| hsa00980:Metabolism of xenobiotics by cytochrome P450 | 8     | 4.232804 | 1.47E-05 | 0.01705 |
| hsa00830:Retinol metabolism                           | 7     | 3.703704 | 7.16E-05 | 0.0831  |
| hsa00982:Drug metabolism - cytochrome P450            | 6     | 3.174603 | 8.56E-04 | 0.98982 |
| hsa05204:Chemical carcinogenesis                      | 6     | 3.174603 | 0.001783 | 2.05075 |
| hsa05323:Rheumatoid arthritis                         | 6     | 3.174603 | 0.002714 | 3.1067  |
| hsa04145:Phagosome                                    | 7     | 3.703704 | 0.006527 | 7.32252 |
| hsa00053:Ascorbate and aldarate metabolism            | 3     | 1.587302 | 0.035093 | 33.9577 |
| hsa05164:Influenza A                                  | 6     | 3.174603 | 0.042085 | 39.3062 |
| hsa05310:Asthma                                       | 3     | 1.587302 | 0.042586 | 39.674  |
| hsa05140:Leishmaniasis                                | 4     | 2.116402 | 0.042895 | 39.8993 |
| hsa05332:Graft-versus-host disease                    | 3     | 1.587302 | 0.050617 | 45.2959 |
| hsa00040:Pentose and glucuronate interconversions     | 3     | 1.587302 | 0.059144 | 50.7376 |
| hsa05330:Allograft rejection                          | 3     | 1.587302 | 0.06209  | 52.4992 |
| hsa04940:Type I diabetes mellitus                     | 3     | 1.587302 | 0.077528 | 60.8268 |
| hsa00860:Porphyrin and chlorophyll metabolism         | 3     | 1.587302 | 0.077528 | 60.8268 |
| hsa00983:Drug metabolism - other enzymes              | 3     | 1.587302 | 0.090651 | 66.8317 |
| hsa04672:Intestinal immune network for IgA production | 3     | 1.587302 | 0.094029 | 68.2344 |

図 11 KEGG PATHWAY 上で 2 個以上の遺伝子が記録されていたパスウェイ

#### 4.4 遺伝子のクラスタリング

遺伝子にも BHC によるクラスタリングを行った結果、図 12 に示す 20 個のクラスタに分かれた。さらに、これらクラスターを用いた、サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップを作成し図 18 に示す。列にサンプルクラスターの番号を、行にそれぞれの遺伝子名と対応する遺伝子クラスター番号を示す。また、ヒートマップの色と値との対応を図 13 に示す。

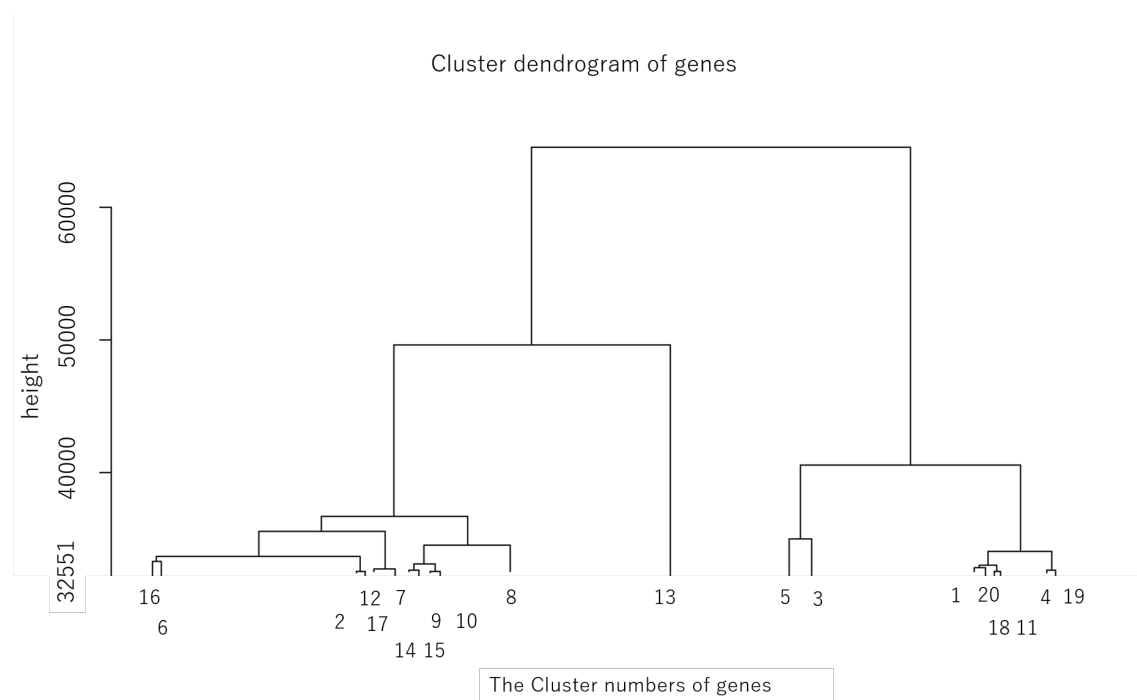


図 12 遺伝子のクラスタリング

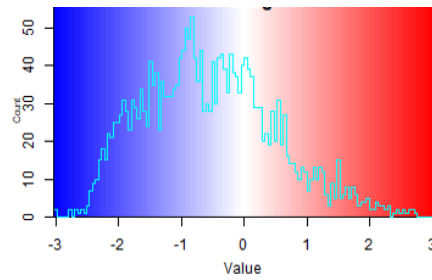


図 13 ヒートマップの色と値との対応

遺伝子に対するBHCによるクラスタリングの結果、CYPファミリーに属する遺伝子は異なるクラスターに属する結果となった。これは、同じ代謝経路上にある遺伝子であっても、肺腺癌患者内でのふるまいに変化があることを示唆している。



図 14 サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ1

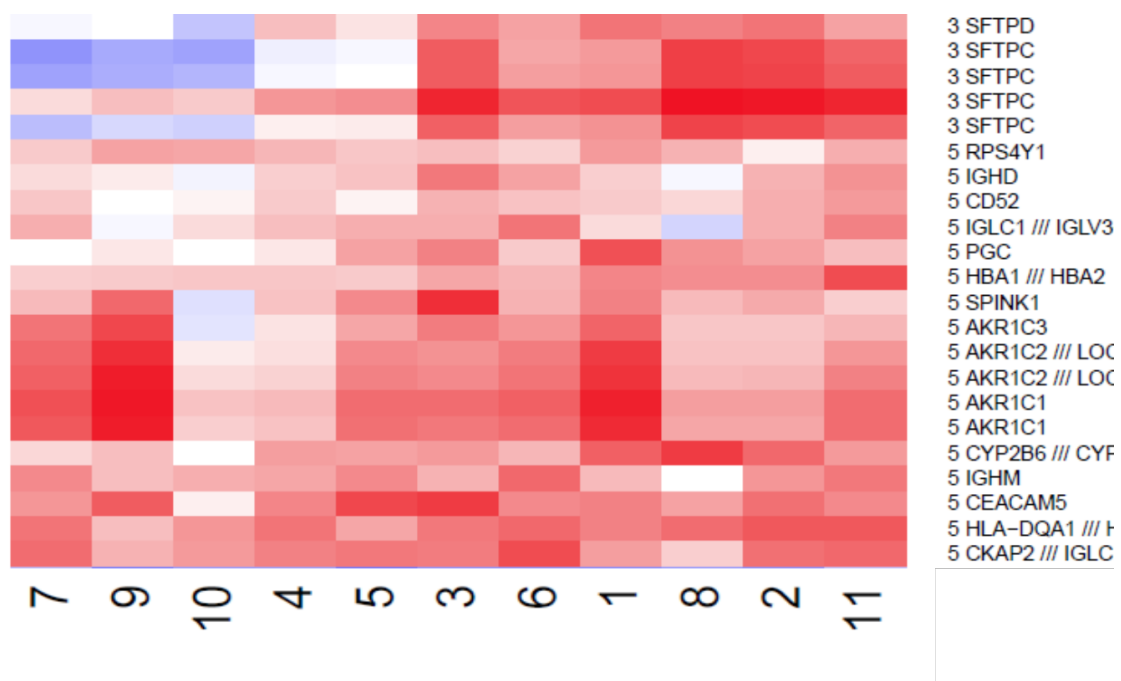


図 15 サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ2

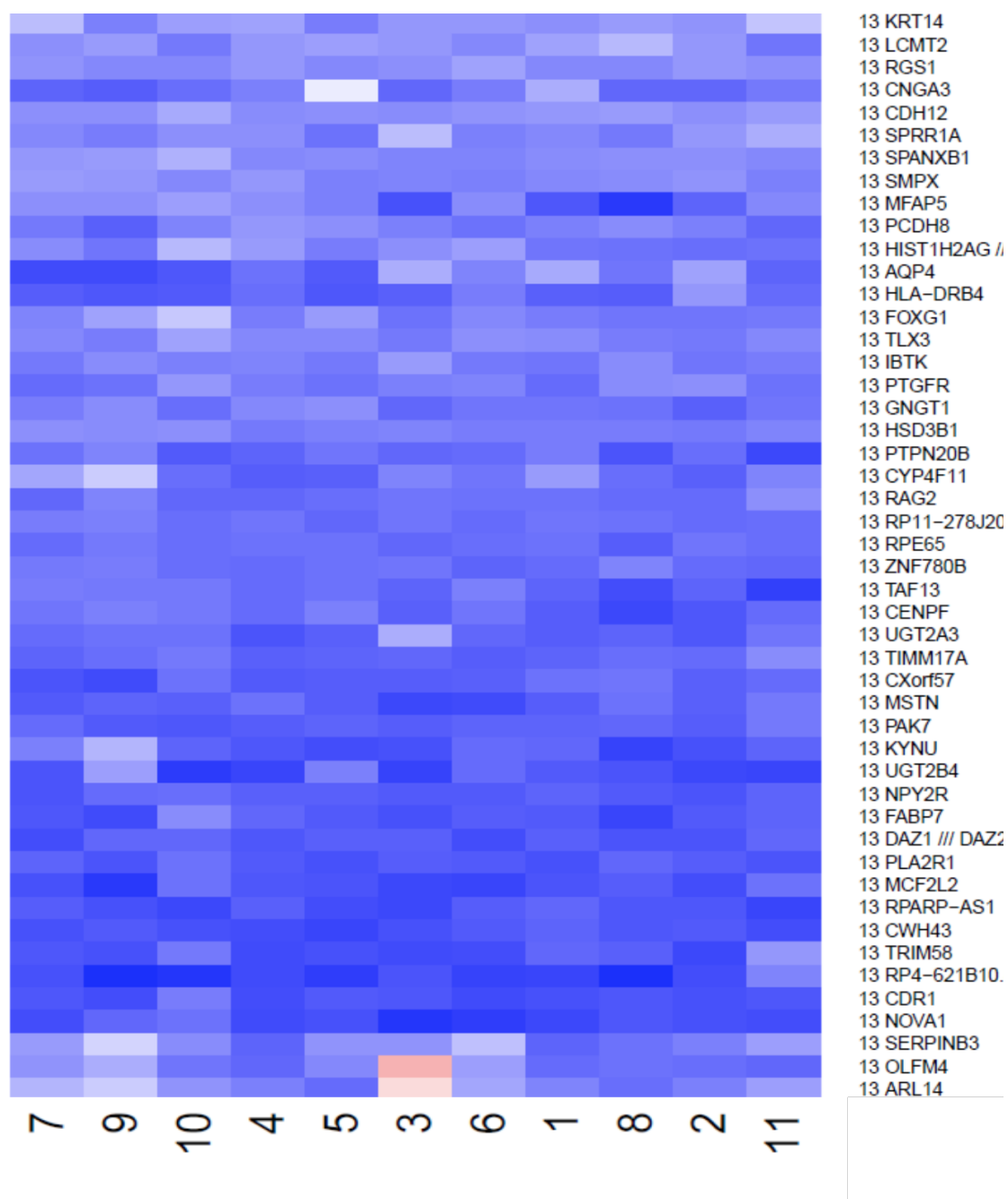


図 16 サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 3



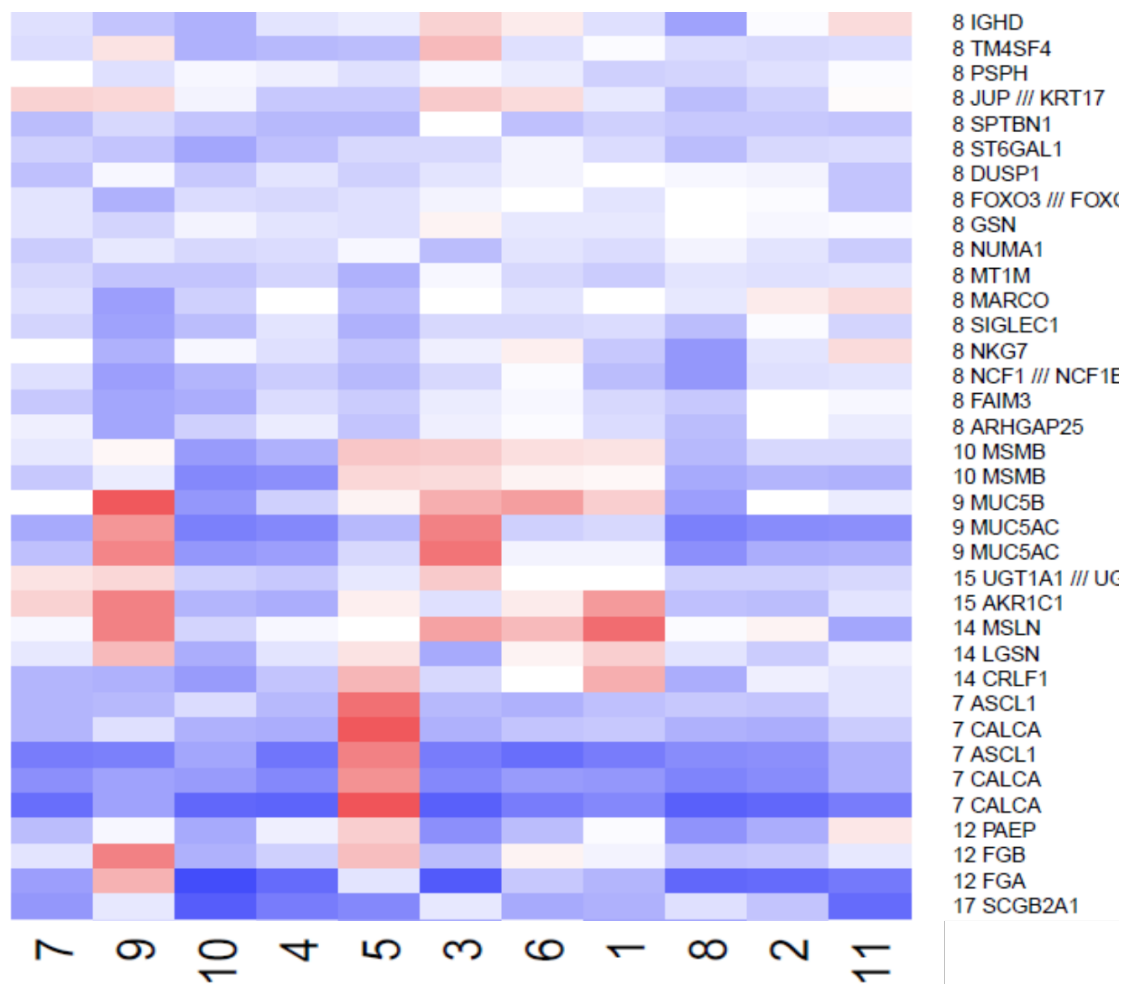


図 17 サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 4

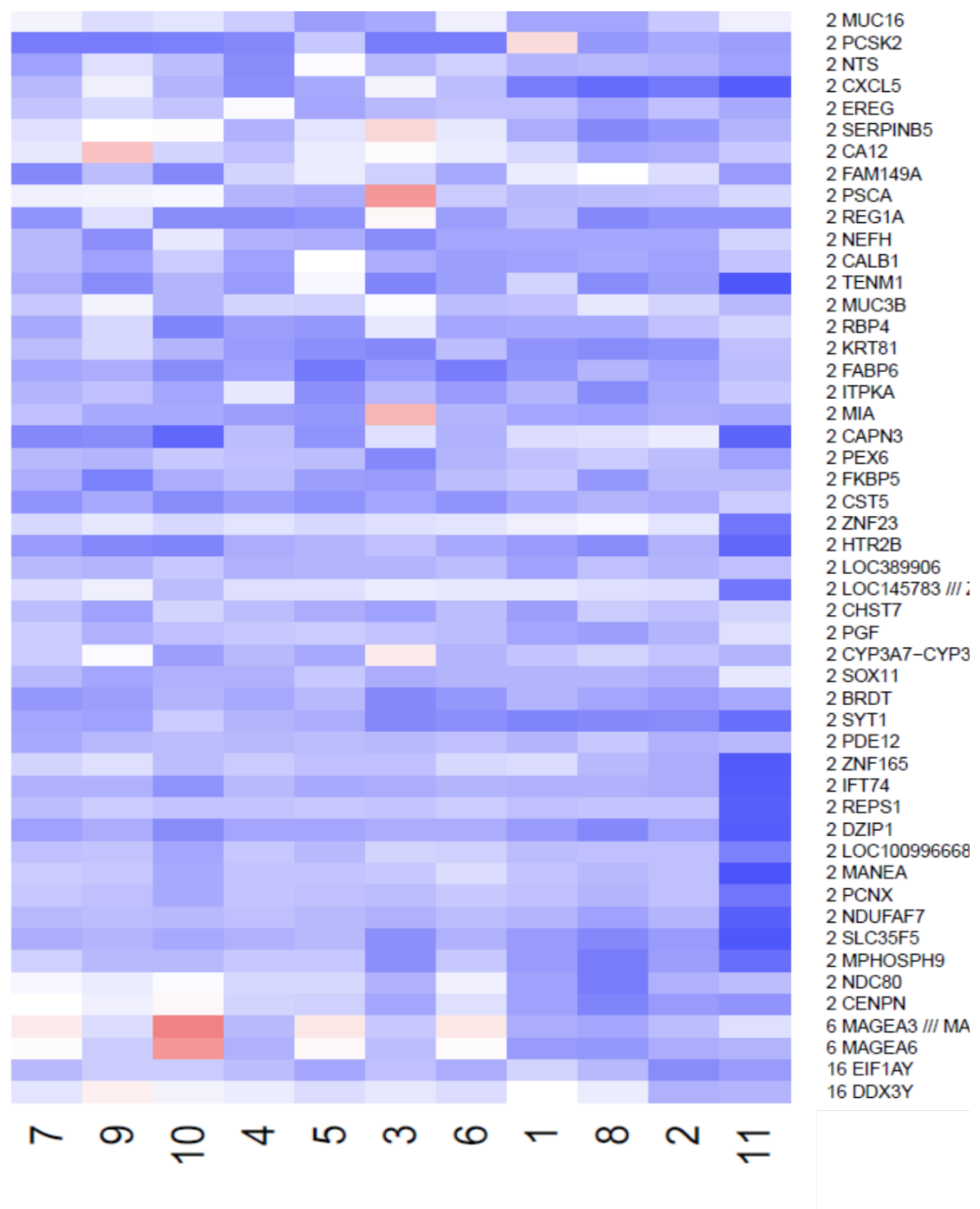


図 18 サンプルクラスターごとに平均化したトランスクリプトームデータのヒートマップ 5

## 5. 結論

BHC を用いたトランスクリプトームデータを用いた肺腺癌患者のクラスターリングにより、11 のクラスを得た。クラスター 3 における MIA や、クラスター 10 における PEG10 等、いくつかのクラスには他のクラスにはない発現を示す遺伝子が見受けられた。また、高い割合で喫煙者を含むクラスター 5 において、喫煙との相関が認められている小細胞癌で有意な発現が示される ASCL1 が高い発現を示していた。その他にも、クラスター 3 における TFF1 と癌の分化など、いくつかのクラスでは臨床データと関連する因子が、1 つのクラスで特徴的な発現を示していた。さらに、クラスター 9 における FGA や CA12 等、あるクラスにおいて特徴的に発現しているが肺腺癌との関りが報告されていない遺伝子も見られた。これらのことから、本研究で用いた手法は GSE68465 データセット内で、トランスクリプトームデータ、臨床データの観点から意味のあるクラスターを抜き出すことに成功した。本来の目的である複数のデータセット間で同様に確認することが可能な肺腺癌の特徴および、関連する遺伝子の探索と、それによる新たな肺腺癌の分類の提唱のためには、また異なるデータセットに対して今回作成したクラスターリングモデルとの対応を調べる必要がある。

## 謝辞

#腰を据えて統計を学ぶにはいい時期だったと思う。ビッグデータの正体を見たり。風のうわさではRがVBAにとって代わる時代が来るとか来ないとか。この2年間、NAISTを存分に楽しんだとも。

```
#Thank.You.R
```

```
repeat
```

```
{
```

```
  if(is.submit(this.paper)) print('thank you !');
```

```
}
```

## 参考文献

- [1] Torre, Lindsey A., et al. "Global cancer statistics, 2012." *CA: a cancer journal for clinicians* 65.2 (2015): 87-108.
- [2] Travis, William D., et al. *Histological typing of lung and pleural tumours*. Springer Science & Business Media, 2012.
- [3] Petrovich, Zbigniew, et al. "Clinical report on the treatment of locally advanced lung cancer." *Cancer* 40.1 (1977): 72-77.
- [4] Morstyn, George, et al. "Small cell lung cancer 1973-1983: early progress and recent obstacles." *International Journal of Radiation Oncology\* Biology\* Physics* 10.4 (1984): 515-539.
- [5] Meyerson M, Hayes DN: Microarray approaches to gene expression analysis, in Tsongalis GJ, Coleman WB (eds): *Molecular Diagnostics: For the Clinical Laboratorian* (ed 2). Totowa, NJ, Humana Press, 2005, pp 121-148
- [6] Shapiro, Samuel Sanford, and Martin B. Wilk. "An analysis of variance test for normality (complete samples)." *Biometrika* 52.3/4 (1965): 591-611.
- [7] Heller, Katherine A., and Zoubin Ghahramani. "Bayesian hierarchical clustering." *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005.
- [8] Hayes, D. Neil, et al. "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts." *Journal of Clinical Oncology* 24.31 (2006): 5079-5090.
- [9] Dennis, Glynn, et al. "DAVID: database for annotation, visualization, and integrated discovery." *Genome biology* 4.9 (2003): R60.

- [10] Demidyuk, Ilya V., et al. "Alterations in gene expression of proprotein convertases in human lung cancer have a limited number of scenarios." *PLoS One* 8.2 (2013): e55752.
- [11] Expression in human lung tumor cells of the proprotein processing enzyme PC1/PC3. Cloning and primary sequence of a 5 kb cDNA. Creemers JW, Roebroek AJ, Van de Ven WJ *FEBS Lett.* 1992 Mar 23; 300(1):82-8.
- [12] Comparative analysis of expression of the proprotein convertases furin, PACE4, PC1 and PC2 in human lung tumours. Mbikay M, Sirois F, Yao J, Seidah NG, Chretien M *Br J Cancer.* 1997; 75(10):1509-14.
- [13] Scopsi L, Gullo M, Rilke F, Martin S, Steiner DF (1995) Proprotein convertases (PC1/PC3 and PC2) in normal and neoplastic human tissues: their use as markers of neuroendocrine differentiation. *J Clin Endocrinol Metab* 80: 294301
- [14] Prohormone convertase and autocrine growth factor mRNAs are coexpressed in small cell lung carcinoma. Rounseville MP, Davis TP *J Mol Endocrinol.* 2000 Aug; 25(1):121-8.
- [15] Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M *Proc Natl Acad Sci U S A.* 2001 Nov 20; 98(24):13790-5.
- [16] Mazieres, Julien, et al. "Wnt inhibitory factor-1 is silenced by promoter hypermethylation in human lung cancer." *Cancer research* 64.14 (2004): 4717-4720.
- [17] Polakis P. Wnt signaling and cancer. *Genes Dev*, 14: 1837-51, 2000.

- [18] Lustig B, Behrens J. The Wnt signaling pathway and its role in tumor development. *J Cancer Res Clin Oncol*, 129: 199-221, 2003.
- [19] Stewart, David J. "Wnt signaling pathway in nonsmall cell lung cancer." *JNCI: Journal of the National Cancer Institute* 106.1 (2014).
- [20] Graumann, Rebecca, et al. "Expression of teneurins is associated with tumor differentiation and patient survival in ovarian cancer." *PloS one* 12.5 (2017): e0177244.
- [21] ZE ZHANG, HE WU. "Identification of potential key genes and high-frequency mutant genes in prostate cancer by using RNA-Seq data" 150001, P.R. China Received April 15, 2016; Accepted June 22, 2017
- [22] Dong H, Ge X, Shen Y, Chen L, Kong Y, Zhang H, Man X, Tang L, Yuan H, Wang H, et al: Gene expression profile analysis of human hepatocellular carcinoma using SAGE and LongSAGE. *BMC Med Genomics* 2: 5, 2009.
- [23] Jovov B, AraujoPerez F, Sigel CS, Stratford JK, McCoy AN, Yeh JJ and Keku T: Differential gene expression between African American and European American colorectal cancer patients. *PLoS One* 7: e30168, 2012.
- [24] Jinawath, Natini, Yoichi Furukawa, and Yusuke Nakamura. "Identification of NOL8, a nucleolar protein containing an RNA recognition motif (RRM), which was overexpressed in diffuse- type gastric cancer." *Cancer science* 95.5 (2004): 430-435.
- [25] Sy, Shirley MH, et al. "The ubiquitin specific protease USP34 promotes ubiquitin signaling at DNA double-strand breaks." *Nucleic acids research* 41.18 (2013): 8572-8580.
- [26] Ma, Jonathan, Sadhika Malladi, and Andrew H. Beck. "Systematic analysis of sex-linked molecular alterations and therapies in cancer." *Scientific reports* 6 (2016): 19119.

- [27] 石和直樹, et al. "原発性肺癌における Thyroid Transcription Factor-1 (TTF-1) 発現の免疫組織化学的検討." 肺癌 41.1 (2001): 45-49.
- [28] Su, Wenmei, et al. "Low expression of olfactomedin 4 correlates with poor prognosis in smoking patients with nonsmall cell lung cancer." Human pathology 46.5 (2015): 732-738.
- [29] Wei, Xinru, et al. "PSCA and MUC1 in non-small-cell lung cancer as targets of chimeric antigen receptor T cells." Oncoimmunology 6.3 (2017): e1284722.
- [30] Sasahira, Tomonori, et al. "A comprehensive expression analysis of the MIA gene family in malignancies: MIA gene family members are novel, useful markers of esophageal, lung, and cervical squamous cell carcinoma." Onco-target 7.21 (2016): 31137.
- [31] Rollin, J., et al. "Expression and methylation status of tissue factor pathway inhibitor-2 gene in non-small-cell lung cancer." British journal of cancer 92.4 (2005): 775.
- [32] Ohue, Yoshihiro, et al. "Spontaneous antibody, and CD4 and CD8 T- cell responses against XAGE- 1b (GAGED2a) in non- small cell lung cancer patients." International journal of cancer 131.5 (2012).
- [33] Borromeo, Mark D., et al. "ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs." Cell reports 16.5 (2016): 1259-1272.
- [34] Calcitonin gene-related peptide (CALCA) is a proangiogenic growth factor in the human placental development. Dong YL, Reddy DM, Green KE, Chauhan MS, Wang HQ, Nagamani M, Hankins GD, Yallampalli C Biol Reprod. 2007 May; 76(5):892-9.
- [35] Calcitonin gene-related peptide and its receptors: molecular genetics, physiology, pathophysiology, and therapeutic potentials. Wimalawansa SJ Endocr



Rev. 1996 Oct; 17(5):533-85.

- [36] Retinoic acid abolishes the calcitonin gene-related peptide autocrine system in F9 teratocarcinoma cells. Segond N, Gerbaud P, Taboulet J, Jullienne A, Moukhtar MS, Evain-Brion D J Cell Biochem. 1997 Mar 1; 64(3):447-57.
- [37] Wang, Beibei, et al. "Production of CCL20 from lung cancer cells induces the cell migration and proliferation through PI3K pathway." Journal of cellular and molecular medicine 20.5 (2016): 920-929.
- [38] Schtz, Alexander, et al. "Lung adenocarcinomas and lung cancer cell lines show association of MMP-1 expression with STAT3 activation." Translational oncology 8.2 (2015): 97-105.
- [39] Wakata, Kouki, et al. "A favourable prognostic marker for EGFR mutant non-small cell lung cancer: immunohistochemical analysis of MUC5B." BMJ open 5.7 (2015): e008366.
- [40] MacLeod, A. Kenneth, et al. "Aldo-keto reductases are biomarkers of NRF2 activity and are co-ordinately overexpressed in non-small cell lung cancer." British journal of cancer 115.12 (2016): 1530.
- [41] Rsnen, Kati, et al. "Emerging roles of SPINK1 in cancer." Clinical chemistry 62.3 (2016): 449-457.
- [42] Tiwari, R., et al. "SPINK1 promotes colorectal cancer progression by down-regulating metallothioneins expression." Oncogenesis 4.8 (2015): e162.
- [43] de Aberasturi, Arrate L., et al. "TMPRSS4 induces cancer stem cell-like properties in lung cancer cells and correlates with ALDH expression in NSCLC patients." Cancer letters 370.2 (2016): 165-176.
- [44] Deng, Xinzhou, et al. "PEG10 plays a crucial role in human lung cancer proliferation, progression, prognosis and metastasis." Oncology reports 32.5 (2014): 2159-2167.

- [45] Pineda, Carlos T., and Patrick Ryan Potts. "Oncogenic MAGEA-TRIM28 ubiquitin ligase downregulates autophagy by ubiquitinating and degrading AMPK in cancer." *Autophagy* 11.5 (2015): 844-846.
- [46] F. J. Gonzalez and H. V. Gelboin: Role of human cytochromes P450 in the metabolic activation of chemical carcinogens and toxins. *Drug Metab Rev*, 26(1-2), 165-83 (1994)
- [47] K. F. Windmill, R. A. McKinnon, X. Zhu, A. Gaedigk, D. M. Grant and M. E. McManus: The role of xenobiotic metabolizing enzymes in arylamine toxicity and carcinogenesis: functional and localization studies. *Mutat Res*, 376(1-2), 153-60 (1997)
- [48] T. Oyama, K. Sugio, H. Uramoto, T. Kawamoto, N. Kagawa, S. Nadaf, D. Carbone and K. Yasumoto: Cytochrome P450 expression (CYP) in non-small cell lung cancer. *Front Biosci*, 12, 2299-308 (2007)
- [49] Futreal, P. Andrew, et al. "A census of human cancer genes." *Nature Reviews Cancer* 4.3 (2004): 177.
- [50] Gibbs, A. R., and F. B. J. M. Thunnissen. "Histological typing of lung and pleural tumours." (2001): 498-499.
- [51] Travis, William D., et al. "The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification." *Journal of thoracic oncology* 10.9 (2015): 1243-1260.
- [52] Wilkerson, Matthew D., et al. "Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation." *PloS one* 7.5 (2012): e36530.
- [53] Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma." *Nature* 511.7511 (2014): 543.

- [54] Hayes, D. Neil, et al. "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts." *Journal of Clinical Oncology* 24.31 (2006): 5079-5090.
- [55] Alizadeh AA, Ross DT, Perou CM, et al: Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol* 195: 41-52, 2001
- [56] Sorlie T, Tibshirani R, Parker J, et al: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 100:8418-8423, 2003
- [57] Cleator S, Ashworth A: Molecular profiling of breast cancer: Clinical implications. *Br J Cancer* 90:1120-1124, 2004
- [58] Takeuchi T, Tomida S, Yatabe Y, et al: Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J Clin Oncol* 24:1679-1688, 2006
- [59] Lynch TJ, Bell DW, Sordella R, et al: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 350:2129-2139, 2004
- [60] Paez JG, Janne PA, Lee JC, et al: EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science* 304:1497- 1500, 2004
- [61] Chemotherapy in non-small cell lung cancer: A meta-analysis using updated data on individual patients from 52 randomised clinical trials: Non- Small Cell Lung Cancer Collaborative Group. *BMJ* 311:899-909, 1995
- [62] Shedden, Kerby, et al. "Gene expressionbased survival prediction in lung adenocarcinoma: a multi-site, blinded validation study." *Nature medicine*

14.8 (2008): 822.

## 付録