Vol.2017-MPS-113 No.41
Vol.2017-BIO-50 No.41
2017/6/25

高速相同性解析ツールGHOSTXを用いた口腔内メタゲノム解析

山澤 まりな[1,2]　伊澤 和輝[1]　大上 雅史[1]　石田 貴士[1,2]　石原 和幸[3]　秋山 泰[1,2,a)]

概要：本研究では，ホールゲノムショットガン法を使用した口腔内メタゲノムデータの大規模解析を行うことで歯周病に関連した細菌の存在度や分布，遺伝子の種類や頻度を調査し，その結果を歯周病のリスク予測へ応用することを目指している．はじめに口腔内メタゲノムデータに対して本研究室で開発された配列相同性検索ツールである GHOSTX を使用した配列相同性検索を行った．続いて，配列相同性検索の結果を用いて，検出される細菌の多様性を健常者と歯周病罹患者で比較した．この結果，これまでに歯周病への関連度が高いと言われていた細菌種と逆の性質を持つかもしれない 2 つの細菌種を見つけた．

キーワード：メタゲノム解析，WGS，歯周病

Oral metagenomic analysis using high-speed homology searching tool GHOSTX

Marina Yamasawa[1,2]　Kazuki Izawa[1]　Masahito Ohue[1]　Takashi Ishida[1,2]　Kazuyuki Ishihara[3]
Yutaka Akiyama[1,2,a)]

Abstract: In this study, we studied the abundance and distribution of bacteria related to periodontal disease by carrying out a large-scale analysis of oral metagenomic data using the whole genome shotgun method, aiming to apply it to the risk prediction of periodontal disease. Firstly, we applied a sequence homology search tool, GHOSTX, which is developed in our laboratory to oral cavity metagenomic data. Secondly, using the results of sequence homology search, we compared differences of the diversity of bacteria between healthy volunteers and periodontal patients. As a result, we found two bacterial species that may have opposite properties to those which had been said to be highly relevant to periodontal disease.

Keywords: Metagenomic analysis, Whole genome shotgun analysis, Periodontal disease

1. 導入

1.1 メタゲノム解析

次世代 DNA シーケンサーの急速な発達により，近年 DNA 塩基配列を読み取るコストが大きく減少している [1]．このコストの低下により，従来のゲノム解析のように単一生物のゲノムのみでなく，微生物集団のゲノムをまとめて読み取るメタゲノム解析が行われるようになった．また，土壌や海中，生物の体内などに生息する多種多様な微生物は，環境中でお互いに相互作用し合っており，そのほとんどは実験室内での培養は困難とされている [2]．この本来の環境の再現という点からも，環境中から多種多様な微生物が混在したままのサンプルを採取し，サンプル中のすべての微生物のゲノム配列をシーケンスするメタゲノム解析と呼ばれる手法が取られるようになった．メタゲノム解

[1] 東京工業大学 情報理工学院 情報工学系，
Department of Computer Science, School of Computing, Tokyo Institute of Technology
[2] 東京工業大学 情報生命博士教育院，
Education Academy of Computational Life Sciences, Tokyo Institute of Technology
[3] 東京歯科大学 大学院歯学研究科 歯学専攻 微生物学講座，
Microbiology, Department of Basic Science, Tokyo Dental College
[a)] akiyama@c.titech.ac.jp

析を用いることにより，環境中の既知の微生物の構成や遺伝的機能の特徴の同定に加えて，未知の微生物や遺伝子が解明できる可能性がある．また，メタゲノム解析の一手法として系統分類解析のみでなく遺伝子機能に関する解析を可能とするホールゲノムショットガン（Whole Genome Shotgun）解析があり，これはゲノム配列全体をすべて読み取る手法である．

1.2 配列相同性検索

ある環境から採取したメタゲノムデータ内のリードをデータベースに対して検索をかける手法として，配列相同性検索がある．DNA などの核酸配列やタンパク質のアミノ酸配列を変異や挿入，欠失なども考慮しながら大規模データベース内の配列と比較し，類似性の高い配列の探索をする．これを行うツールとして，BLAST [3] が広く利用されているが，次世代 DNA シーケンサーが読み取る膨大なリードの検索により特化した高速なツールである GHOSTX [4] や GHOSTZ [5] なども存在する．

1.3 口腔内細菌叢と歯周病

ヒトの口腔内とは身体の中でも最も重要視されている部位の 1 つであり，極めて複雑かつ高度に組織化されたバイオフィルム [6, 7] である歯垢からは 600 種類以上もの菌種が確認されている [8, 9]．歯周病はこれらの菌の内いくつかによって引き起こされると言われている歯周組織に発生する慢性疾患の総称である．平成 23 年度の厚生労働省による調査 [10] では，45 歳以上の日本人において歯周疾患の目安となる歯周ポケットの深さが 4 mm 以上存在している割合が半数に達していると報告された．また，歯周病は放置すると心血管疾患や糖尿病，肺疾患および肥満などを悪化させる可能性があるとされている [11–13]．

1.4 本研究の目的

本研究では，健常者と歯周病罹患者の口腔内メタゲノムデータに対して WGS 解析を行う．配列相同性検索では GHOSTX を使用した．その結果に対して系統分類と遺伝子機能の 2 つの観点から健常者と歯周病罹患者の 2 群間比較を行い，歯周病に関するマーカーとなるような細菌種や遺伝子を見つけることを目的とする．2 群間比較を行う対象データとして今回独自にサンプリングを行ったデータの他に，公開されているメタゲノムデータも加え，考察を行った．

2. 手法

2.1 使用データ概要

2.1.1 東京歯科大学データ

東京歯科大学にて健常者 10 名，歯周病罹患者 3 名から歯垢サンプルの採取を行った．健常者（Healthy）から 11 サンプル（S_H，うち S_H7, S_H8 が同一人物），歯周病罹患者（Patient）は患部と健常部に関して歯垢サンプルの採取を行い，患部（S_P）から 6 サンプル，健常部（S_PH）から 5 サンプル得た．歯周病罹患者 3 名を順に罹患者 A，B，C とし，病名は A は限局性慢性歯周炎，B，C はどちらも慢性歯周炎であった．各歯垢サンプルから DNA を抽出した後に Nextera XT DNA Library Prep Kit を用いてライブラリ調整を行った．調整したライブラリは Illumina 社の MiSeq によりシーケンシングを行った．リード長は 151 bp である．**表 1**–**表 2** に各データのリード数を示す．

2.1.2 A. Duran-Pinedo（2014）データ

A. Duran-Pinedo ら [14] の研究で使用されている，健常者 6 名（サンプル ID：H），歯周病罹患者 7 名（サンプル ID：P）の口腔内メタゲノムデータを使用した．これらは Illumina 社の MiSeq でシーケンスされており，リード長は 250 bp である．**表 3**–**表 4** に各データのリード数を示す．

2.2 クオリティチェック（QC）

上記の各口腔内メタゲノムデータに対し，以下のクオリティチェックを行った．表 1–4 にクオリティチェック（QC）後のリード数を示す．

2.2.1 クオリティフィルタリング

FASTX-Toolkit [15] 内の `fastq_quality_filter` を使用し，クオリティスコアが 20 以上の塩基が 80%未満であるようなリードを除去した．ここでクオリティスコアとは，$Q = -10\log_{10}P_{\mathrm{error}}$ で求められる，シーケンサーで読み取られたデータの信頼性を表すスコアである（P_{error}：シーケンシングの際にエラーが生じる確率）．ただし，A. Duran-Pinedo（2014）データに関しては論文 [14] 内で同様の操作を行ったデータを公開しているため，本研究ではクオリティフィルタリングを行っていない．

2.2.2 ヒト由来ゲノム除去

口腔内メタゲノムデータである各サンプル内には，サンプリングの際にヒト由来であるゲノムが混入している場合がある．ヒト由来のゲノムが残存したまま配列相同性検索を行ってしまうと，ヒトが元々持つ遺伝子機能と微生物によって活発化している遺伝子機能との区別がつかなくなる．よって，配列相同性検索前にヒトゲノムを除いておく必要がある．本研究では，Bowtie2 [16] を使用し，UCSC [17] にて公開されているヒトリファレンスゲノム（h19）[18] に対して各サンプルのマッピングを行い，マップされたリードを除去した．

2.3 配列相同性検索

クオリティチェック後，配列相同性検索ツールである GHOSTX [4] を使用して配列相同性検索を行った．この際，シーケンサーで読み取った各リードそれぞれの最良スコアのヒットのみを使用した．

Vol.2017-MPS-113 No.41
Vol.2017-BIO-50 No.41
2017/6/25

表 1 東京歯科大学データ（健常者）

サンプル ID	S_H1	S_H2	S_H3	S_H4	S_H5	S_H6
リード数	2,887,601	1,915,633	2,711,070	2,433,209	2,123,728	2,132,532
QC 後リード数	820,569	129,853	596,645	599,840	592,002	374,923
サンプル ID	S_H7	S_H8	S_H9	S_H10	S_H11	
リード数	42,618	152,142	192,350	648,426	1,454,752	
QC 後リード数	2,555	9,992	3,843	14,962	78,585	

表 2 東京歯科大学データ（歯周病罹患者 A，B，C）

罹患者 A/サンプル ID	S_P1	S_P2	S_PH1	S_PH2
リード数	1,359,230	1,292,864	84,302	2,624,566
QC 後リード数	49,997	30,185	33,711	312,230
罹患者 B/サンプル ID	S_P3	S_P4	S_PH3	S_PH4
リード数	93,072	2,398,604	1,205,826	2,697,604
QC 後リード数	18,362	1,045,154	383,209	968,288
罹患者 C/サンプル ID	S_P5	S_P6	S_PH5	
リード数	327,016	309,300	352,898	
QC 後リード数	132,882	133,531	165,085	

表 3 A. Duran-Pinedo（2014）データ（健常者）

サンプル ID	H1	H2	H3	H4	H5	H6
リード数	13,409,154	7,750,679	8,022,900	432,952	897,382	410,109
QC 後リード数	7,770,793	5,192,508	1,581,420	431,268	896,628	72,766

表 4 A. Duran-Pinedo（2014）データ（歯周病罹患者）

サンプル ID	P1	P2	P3	P4	P5	P6	P7
リード数	9,521,815	7,554,209	12,263,433	17,239,936	396,999	2,030,156	563,603
QC 後リード数	5,624,904	5,996,315	12,263,167	17,055,563	387,384	2,028,379	563,488

Vol.2017-MPS-113 No.41
Vol.2017-BIO-50 No.41
2017/6/25

2.4 使用データベース

配列相同性検索時には，京都大学が公開している KEGG GENES [19] にて 2016 年 11 月に公開されたデータベースのうち原核生物のみのデータベースを使用した．データベース内のアミノ酸配列数は 13,440,746 本，約 4.9 GB である．

2.5 相対存在度の計算

配列相同性検索で出力されたデータに対して後述の解析スクリプトを実行し，KO 番号を基にした各データ内の遺伝子発現量と生物種名，それぞれの相対存在度の取得を行った．この解析スクリプトの出力内容としては，

- KO 番号，相対存在度
- 生物分類名，相対存在度

の 2 つがある．KO 番号とは，遺伝子の機能的類似性に関する情報を集積した KEGG ORTHOLOGY と呼ばれるデータベース上で使用される識別子である．以下に，この解析スクリプトの内容を示す．

解析用スクリプトの内容

(1) 相同性検索の際に出力されたファイル内には，得られた配列それぞれに対してアラインメントスコアや配列一致度が計算されている．

(2) この 2 つの値に対して閾値を設定し，その閾値以上の配列をヒット配列として定義する．
本研究では閾値をアラインメントスコア：40，配列一致度：70 としている．

(3) ヒット配列と定義された配列に対応する KO 番号，生物名を取得し，対応する遺伝子の出現頻度をカウントする．その際に，ヒット配列の配列長でヒットしたリード数の正規化を行う．

2.6 健常者および歯周病罹患者の 2 群間比較

上記で得た KO 番号と生物分類名の相対存在度を用いて，以下の解析を行った．

(1) Socransky's pyramid に関する比較

(2) PCA 解析

2.6.1 Socransky's pyramid に関する比較

Socransky's pyramid とは，歯周病に関連する菌を統計を用いて初めて表したとされている研究 [20, 21] で提示された歯周病関連菌の分類指標である（**図 1**）．歯周病への関連度が最も高いと考えられる菌の 3 種である "*Porphyromonas gingivalis*"，"*Treponema denticola*"，"*Tannerella forsythensis*" を Red Complex としてピラミッドの頂点に分類し，その下に Orange Complex，Yellow Complex などと続く．ここでは，Socransky's pyramid に分類される菌に着目して解析を行う．

2.6.2 PCA 解析

主成分分析（Principal Component Analysis）とは，多次元のデータを低次元に縮約する手法の 1 つである．これを行うことにより様々な種からなる高次元の口腔内メタゲノムデータを比較的低次元に落とした解析が可能になる．

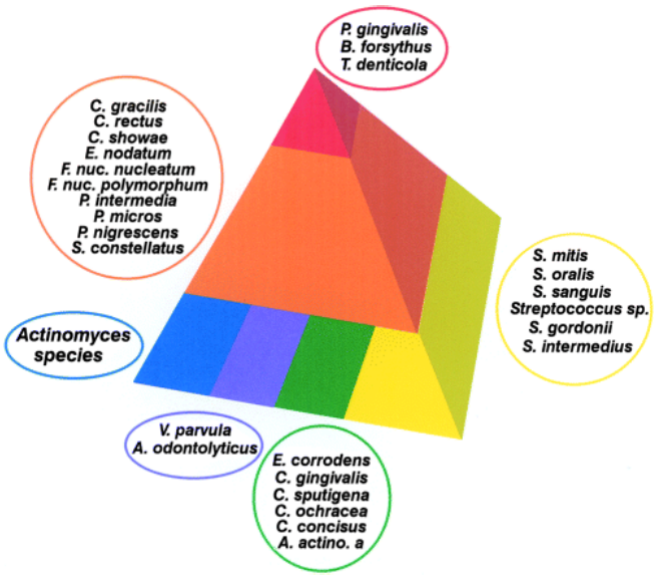

図 1 Socransky's pyramid [20, 21]

3. 結果

3.1 Socransky's pyramid に関する解析

今回使用したデータそれぞれに対して，Socransky's pyramid に分類される種のみの相対存在比率を現したグラフを**図 2**–**図 4** に示す．図 2 は全て健常者のサンプルであるが，歯周病に大きく関連するとされている *Treponema denticola* や，*Fusobacterium nucleatum* などが存在している．

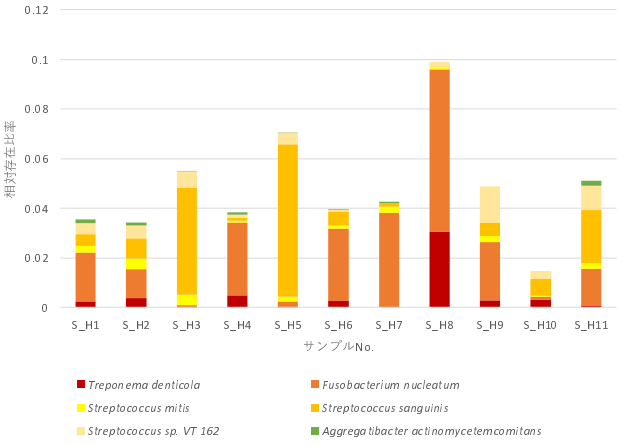

図 2 東京歯科大学データの Socransky 分類（健常者）

Vol.2017-MPS-113 No.41
Vol.2017-BIO-50 No.41
2017/6/25

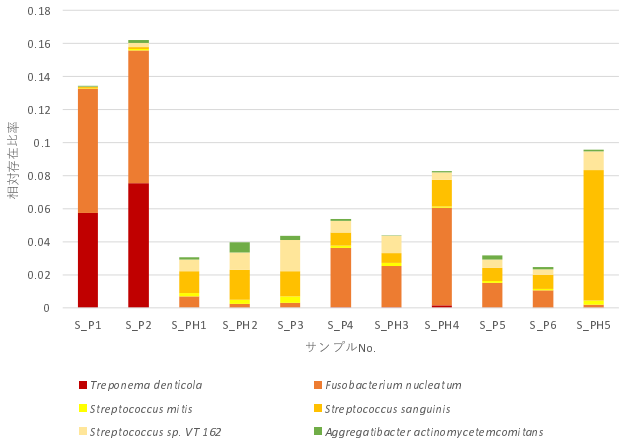

図 3 東京歯科大学データの Socransky 分類（歯周病罹患者患部・健常部）

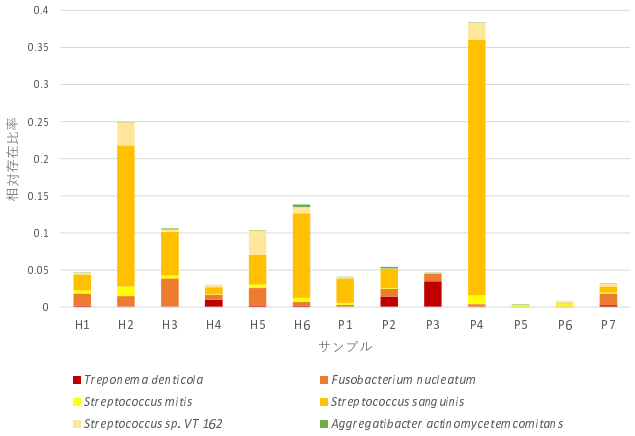

図 4 A. Duran-Pinedo（2014）データの Socransky 分類

3.2 PCA 解析

3.2.1 種階層 PCA 解析

今回使用した東京歯科大学データの全サンプルを用いた PCA 解析の結果と，歯周病罹患者の患部 · 健常部のみの PCA 解析の結果を**図 5**–**図 6** に示す．PCA 解析の入力は種名とその存在比率であり，全サンプルを用いた解析におけるベクトルの次元は 622 であり，歯周病罹患者のみの解析では 612 である．図 5 では，サンプル S_H8，S_P1，S_P2，サンプル S_H9，S_H10 がそれぞれ似通った特徴を持っており，図 6 ではサンプル S_P1，S_P2，S_PH4，サンプル S_PH1，S_P3，S_P4 が似通った特徴を持っている．A. Duran-Pinedo（2014）データの全サンプルを用いた PCA 解析の結果を**図 7** に示す．ベクトルの次元は 636 である．ここでは，サンプル P4 と P6 がその他のサンプルと離れたところにあるものの，健常者サンプルがおおむね *Streptococcus Sanguinis* のベクトルに沿って右側に分類され，歯周病罹患者のサンプルが左側に固まって分類された．

3.2.2 KO 番号に関する PCA 解析

各遺伝子配列に対して割り振られている KO 番号を用いて PCA 解析を行った．入力は KO 番号とその相対存在比

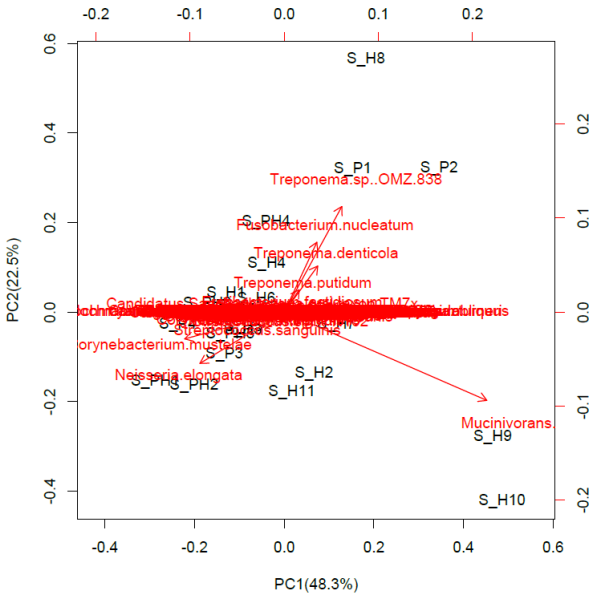

図 5 東京歯科大学データの PCA（全データ使用）

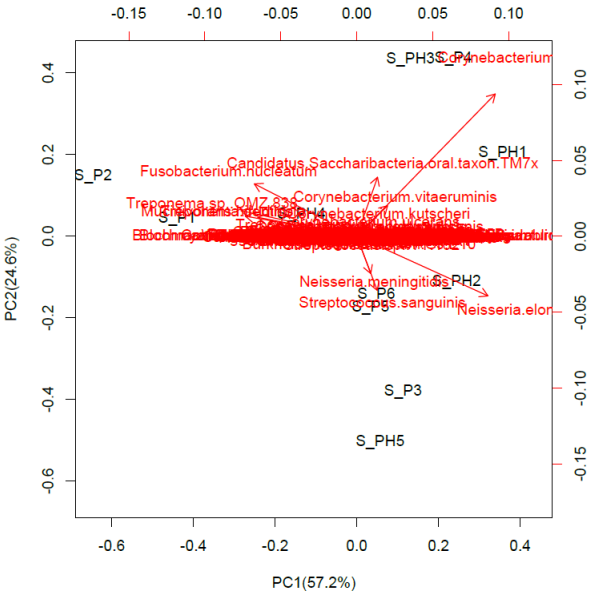

図 6 東京歯科大学データの PCA（歯周病罹患者患部，健常部のみ使用）

率である．東京歯科大学データの PCA 結果を**図 8** に，A. Duran-Pinedo（2014）データの PCA 結果を**図 9** に示す．それぞれ使用したベクトルの次元は，3987 と 6091 である．どちらも 2,3 個の離れたデータが存在し，その他は全てのサンプルが 1 か所に固まる結果になった．

4. 考察

4.1 Socransky's pyramid に関する解析

東京歯科大学データに関して，歯周病罹患者 A のサンプルは同一患者の中でも患部と健常部で細菌叢が分かれていたが，歯周病罹患者 B，C に関してはそうはならなかった．このことから，病状によって細菌叢の変化がもたら

Vol.2017-MPS-113 No.41
Vol.2017-BIO-50 No.41
2017/6/25

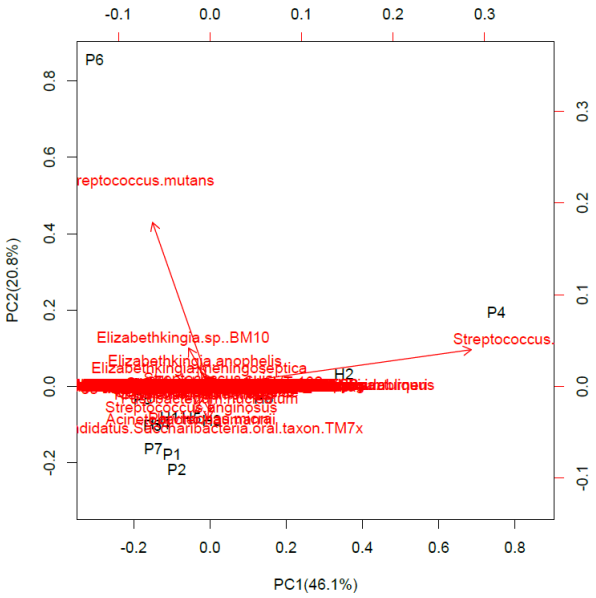

図 7　A. Duran-Pinedo（2014）データの PCA

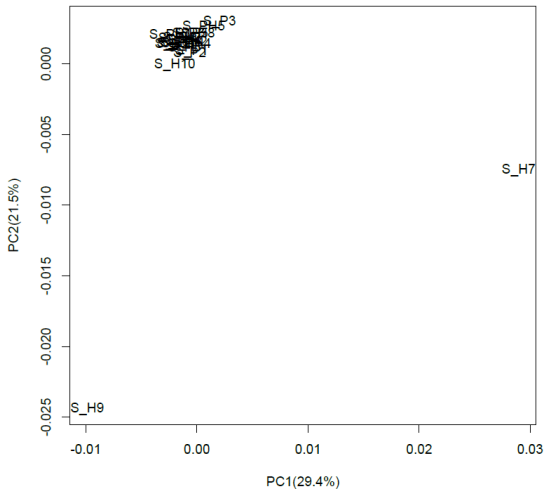

図 8　東京歯科大学データの PCA（KO 番号）

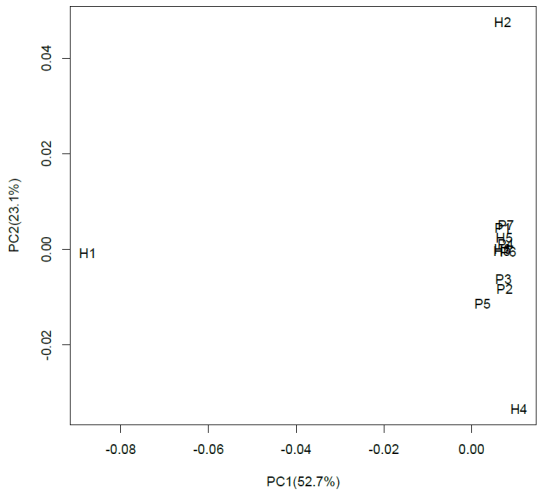

図 9　A. Duran-Pinedo（2014）データの PCA（KO 番号）

されるのではないかと考えられる．限局性慢性歯周炎とは，一部において歯周病の進行が極端に進むものであるため，慢性歯周炎よりも大きな細菌叢および口腔内環境の変化が起きていると考えられる．*Treponema denticola* や *Fusobacterium nucleatum* はこの口腔内環境の変化を反映しているのではないかと考えられる．また，東京歯科大学データと A. Duran-Pinedo（2014）データを比較すると，東京歯科大学データにより多くのこれら 2 種の菌が多く見受けられる．また，東京歯科大学データの健常者の中にも Red complex や Orange complex が比較的高い割合を占めるサンプルがあることから，日本人の口腔内にはもともと比較的多くこれらの菌が存在するのかもしれない．

4.2　PCA 考察

4.2.1　種階層解析

図 5 から，Socransky's pyramid において，歯周病に大きく関連しているとされていた *Treponema denticola* と *Fusobacterium nucleatum* が同じベクトルの向きを持っていることが分かる．これは特定のサンプルにこの 2 種が比較的多く存在することを示す．このことから，これら 2 種が口腔内環境において相互作用している可能性があげられる．実際にこの *Fusobacterium nucleatum* は，*Treponema denticola* などの菌を凝集する菌と言われている．また，そのベクトルに特化したサンプルとしては S_H8, S_P1, S_P2, サンプル S_H9，S_H10 があげられ，S_H8 に関しては歯周病のリスクが高いのかもしれない．*Treponema denticola* や *Fusobacterium nucleatum* のベクトルと同じ向きをもつ種としては，*Treponema sp. OMZ 838* や，*Treponema putidum* などがあり，逆の向きに近いベクトルを持つ種としては，*Neisseria eloganta* や，*Corynebacterium mustelae* がある．今後の解析としては，主にこれらの種の歯周病への関連の知見を集めたい．

4.2.2　遺伝子機能解析

図 8 や図 9 を見ると，飛び地になっているサンプルが見受けられる．これらは，遺伝子 1 つ 1 つに関して見ているために個人差が大きく出てしまったものと考えられる．ここから，データ内に存在するすべての遺伝子を用いる解析は 2 群間比較には向いておらず，ある程度の存在比率以上のものなどの条件を課してデータの一部を使用した方が効果的であると考えられる．

5.　結論

本研究では，健常者と歯周病罹患者の口腔内メタゲノムデータに対して，WGS を用いた 2 群間比較解析を行った．現状ではそれぞれのサンプルに存在するすべての種を解析に使用しているためサンプルの分類に効果的でない種も混在しているはずだが，PCA を行うことによりそれらの菌と効果的に分類できる菌を大まかに目をつけることができ

ると考えた．また，KO 番号を用いた解析に関しては，今回の様に全てを用いると分類ができないため，より効果的に 2 群を比較出来るような種や，KO 番号をいくつか選び出す手法を見つけることが今後の課題である．

謝辞 本研究の一部は，JST CREST「EBD：次世代の年ヨッタバイト処理に向けたエクストリームビッグデータの基盤技術」（課題番号 JPMJCR1303）の支援を受けて行われた．

参考文献

[1] Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, Zhang J, Weinstock GM, Isaacs F, Rozowsky J, Gerstein, M. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* 17(1): 1.

[2] 鎌形 洋一. (2007). 難培養微生物とは何か？. *Journal of Environmental Biotechnology*, 7(2): 69–73.

[3] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215.3: 403–410.

[4] Suzuki S, Kakuta M, Ishida T, Akiyama Y. (2014). GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS ONE* 9.8: e103833.

[5] Suzuki S, Kakuta M, Ishida T, Akiyama Y. (2014). Faster sequence homology searches by clustering subsequences. *Bioinformatics*: btu780.

[6] Kolenbrander PE. (2000). Oral microbial communities: biofilms, interactions, and genetic systems. *Annual Review of Microbiology* 54: 413–437.

[7] Kolenbrander PE, Andersen RN, Blehert DS, Egland PG, Foster JS, Palmer RJJ. (2002). Communication among oral bacteria. *Microbiology and Molecular Biology Reviews* 66: 486–505.

[8] Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, Sahasrabudhe A, Dewhirst FE. (2001). Bacterial diversity in human subgingival plaque. *Journal of Bacteriology* 183: 3770–3783.

[9] Dewhirst FE, Chen T, Izard J, Paster BJ, TannerACR, YuW-H, Lakshmanan A, Wade WG. (2010). The human oral microbiome. *Journal of Bacteriology* 192: 5002–5017.

[10] 厚生労働省 平成 23 年歯科疾患実態調査，http://www.mhlw.go.jp/toukei/list/62-23.html

[11] Teles R, Wang CY. (2011). Mechanisms involved in the association between peridontal diseases and cardiovascular disease. *Oral Diseases*, 17.5: 450–461.

[12] Ali J, Pramod K, Tahir MA, Ansari SH. (2011). Autoimmune responses in periodontal diseases. *Autoimmunity Reviews*, 10.7: 426–431.

[13] Bascones-Martinez A, Matesanz-Perez P, Escribano-Bermejo M, Gonzlez-Moles M, Bascones-Ilundain J, Meurman JH. (2011). Periodontal disease and diabetes-Review of the Literature. *Medicina Oral Patologia Oral y Cirugia Bucal*, 16.6: e722–729.

[14] Duran-Pinedo AE, Chen T, Teles R, Starr JR, Wang X, Krishnan K, Frias-Lopez J. (2014). Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *ISME Journal* 8.8: 1659–1672.

[15] FASTX-Toolkit, http://hannonlab.cshl.edu/fastx_toolkit/

[16] Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9.4: 357–359.

[17] UCSC, https://genome.ucsc.edu/index.html

[18] UCSC Genome Bioinformatics, http://hgdownload.cse.ucsc.edu/downloads.html

[19] Kanehisa M, Goto S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28: 27–30.

[20] Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL. (1998). Microbial complexes in subgingival plaque. *Journal of Clinical Periodontology* 25.2: 134–144.

[21] Socransky SS, Haffajee AD. (2002). Dental biofilms: difficult therapeutic targets. *Periodontology* 28.1: 12–55.