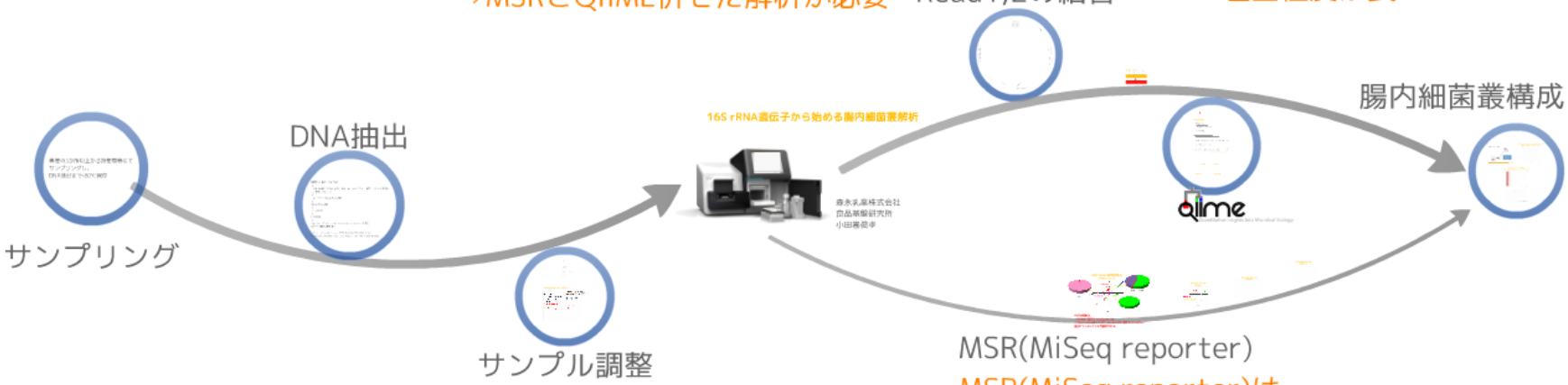
16S rRNA遺伝子から始める腸内細菌叢解析

森永乳業株式会社
食品基盤研究所
小田巻俊孝

Workflow

Read1,2の結合は
・PhiXやクオリティスコアの低い配列を除去
・リードの同定率向上(Unclassifiedが減少)
する上で有効だが、
・設定が厳しいとActinobacteriaを低く見積もる可能性がある。
⇒MSRとQIIME併せた解析が必要

カスタムプライマー
・V4領域はBacteroidetesを低く見積もってしまう
⇒250bp×2に合った増幅部位の検証
・12塩基のIndex配列は>Q30リードが激減
⇒６塩基程度が良い？



サンプルの濃度測定は最重要

サンプルライブラリーは推奨値よりも高め
・PCR product 16pM, PhiX 6pM
（MiSeq v2はこの1.5倍程度？）

MSR(MiSeq reporter)は
PhiXのリードが数%混入してしまう。

サンプリング

DNA抽出

糞便をPBS等で１０倍希釈
↓
１０倍希釈液２００ｍｇにTE飽和フェノールと0.1mm径グラスビーズを混合し、破砕(Fastprep)
↓
フェノール・クロロホルム抽出
↓
クロロホルム抽出
↓
イソプロ沈
↓
TEに溶解
↓
High Pure PCR Template Purification Kit (Roche社製)
にてPCR阻害物質を除去

（Appl Environ Microbiol. 2008 Nov;74(21):6814-7. doi: 10.1128/AEM.01106-08. Epub 2008 Sep 12. PMID: 18791010）

糞便をPBS等で１０倍希釈
↓
１０倍希釈液２００ｍｇにTE飽和フェノールと0.1mm径グラスビーズを混合し、破砕(Fastprep)
↓
フェノール・クロロホルム抽出
↓
クロロホルム抽出
↓
イソプロ沈
↓
TEに溶解
↓
High Pure PCR Template Purification Kit (Roche社製)
にてPCR阻害物質を除去

（Appl Environ Microbiol. 2008 Nov;74(21):6814-7. doi: 10.1128/AEM.01106-08. Epub 2008 Sep 12. PMID: 18791010）

16S rRNA Amplification Protocol

PCRを3連で実施
↓
電気泳動して増幅を確認
↓
カラムを用いたキットにてPCR産物を精製
↓
Picogreenで全サンプルを濃度測定
↓
一緒に解析するサンプルを等量混合
↓
電気泳動にてバンドを切り出す
（primer dimerが100bp以上になりカラムで精製しきれないため）
↓
精製した混合PCR産物をPicogreenで濃度測定し、2nMまで希釈
（DNA濃度を正確に測定することが最重要）
↓
マニュアルに従いPCR産物とPhiXを別々に変性・希釈し20pM溶液を作成
↓
16pMのPCR産物と6pMのPhiX溶液を混合し解析サンプルとする
（イルミナではPhiXの30-50%混合を推奨）
↓
Read1,index,Read2用のCustom primerをそれぞれ12,13,14番のポートに0.5μMずつ混合
（PhiX用のプライマーが必要なためカスタムポートである18,19,20番は使用しない）

> TaKaRaEx Taq HS　（PrimeSTAR® HS ダイレクト PCR キット）

Thermocycler condition

Temp	Time	
94℃	3 min	
94℃	45 sec	┐
50℃	1 min	├ 20(-25) cycle
72℃	1.5 min	┘
72℃	10 min	
4℃	hold	

サンプル調整

illumina

マルチプレックスの16Sメタゲノムを MiSeq™システムで高速解析

はじめに

複雑な微生物群集のプロファイリング解析において、次世代シーケンサーは強力なツールとなりつつあります。MiSeqシステムは簡単操作のシーケンスワークフローを可能にし、サンプルからデータまで迅速な解析が行えます。実績のある1塩基合成反応（SBS: Sequence by Synthesis）テクノロジーを採用したMiSeqは、次世代シーケンサー上位機種であるHiSeq™ 2000システムと同等の高品質なデータを、ランあたり1Gb以上産出します。本アプリケーションノートでは、24サンプルのPCR産物をインデックスを用いた150塩基のペアエンドでシーケンスを行い、末端部分を重ねたリードを使って16Sメタゲノム解析を行いました。

宿主関連と自由生活微生物群集間の既知の違いや、シーケンスデータから得られる生物学的な結論は、どのシーケンスシステムでも高い再現性をもたらすはずです[1,2]。シーケンス後の微生物群集解析は、オープンソースのソフトウェアパッケージとして複数の標準群集解析ツールを統合したQIIME（Quantitative Insights Into MicrobialEcology）パイプライン[3]を用いて行いました。これらの結果は、これまでイルミナのGenome Analyzer™システムで行われてきた高いスループットの微生物群集シーケンス解析が、MiSeqシステムでも同様に行え、これまで以上に低コストで簡単なワークフローで実施できることを示しています。

手法と結果

16S V4領域のシーケンス

16Sリボゾーム RNA（rRNA）遺伝子の配列多型は、微生物群集の分類学的多様性の特徴付けを行う際に広く使われています。16S rRNA遺伝子の配列は9つの超可変領域と、その間に存在する保存配列領域から構成されています。16S rRNA遺伝子の配列と超可変領域は、数多くの生物種で既に決定されており、Greengenes[4]やRibosomal Database Project[5]など複数のデータベースで入手可能です。分類においては、16S rRNA遺伝子の全長の代わりに、超可変領域のみをシーケンスすることで十分な情報が得られます[6,7]。多くの微生物では、16S rRNA遺伝子の4番目の超可変領域（V4）は約254塩基配列からなり、配列長の違いはわずか数塩基です。

V4領域増幅のながれ

16S rRNA遺伝子のV4領域をシーケンスするために、プライマーはV4領域周辺の保存性の高い領域をもとにデザインしました[2]。MiSeqシステムでは150塩基のペアエンド解析ができ、ふたつのリードの末端がオーバーラップすることで非常に高品質な全V4領域の情報を得る事ができます（図1）。これらのプライマーの末端にはMiSeqシステムで解析が可能なように、対応アダプターとインデックス用のバーコードが組み込まれています。この実験では、24種類の識別が可能なインデックス対応のプライマーを使い、24サンプルからV4領域を増幅しました。MiSeqシステムのランには、これらの24サンプルを混合（プール）し、1回のランで同時に解析を行いました。

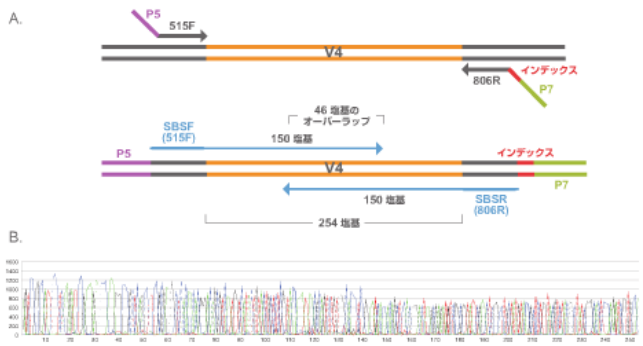
MiSeqシステムでのシーケンスラン

異なるインデックスをつけた24サンプルは混合後、MiSeqシステム用の試薬カートリッジにアプライし、フローセルと一緒にMiSeqシステムにセットしました。クラスター形成を行い、その後13サイクルのインデックス配列およびV4可変領域のシーケンスを行いました。これらのステップはすべて自動化されており、約28時間で終了しました。

データ解析

1次解析（画像解析およびベースコール）はMiSeqシステム上で行いました。クオリティフィルターしたqseqファイルは、QIIME[8]を用いてオフラインで解析しました。QIIMEはシーケンスの生データから論文投稿に必要な図を作成することができる

図1：V4領域増幅のながれとペアエンドリード



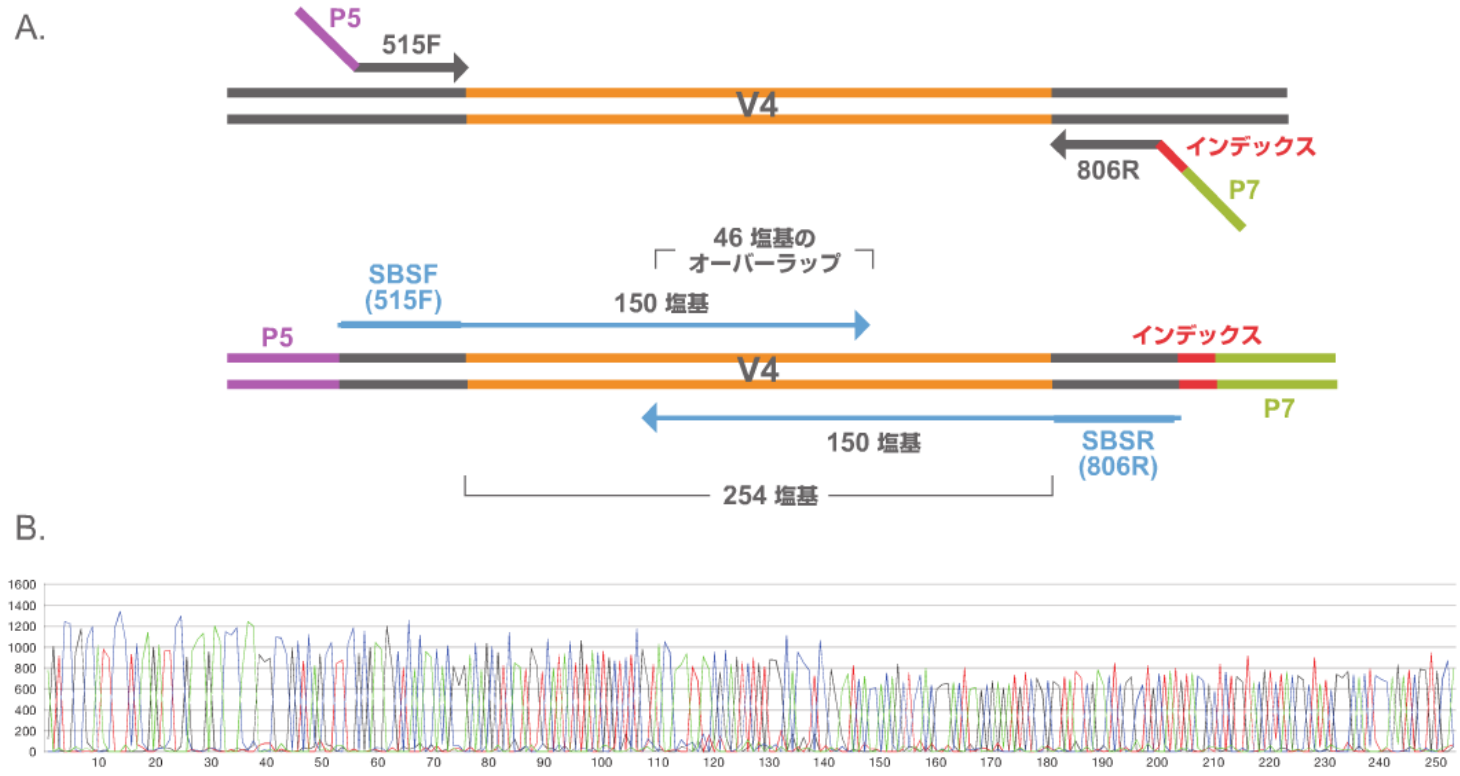
A. 各サンプルのV4領域は515Fと806Rのプライマーを使い増幅しました。プライマーにはP5とP7の配列も組み込まれています。46塩基のオーバーラップをもたらすペアエンドの150塩基のシーケンスランにより、V4領域の全254塩基情報を得ることができます。

B. V4領域の完全な254塩基配列をペアエンドリードで読み取った例。（マトリックスとフェージングを修正後の）生強度を示しています。

の分類学的多様性の特徴付けを行う際に広く使われています。16S rRNA遺伝子の配列は9つの超可変領域と、その間に存在する保存配列領域から構成されています。16S rRN遺伝子の配列と超可変領域は、数多くの生物種で既に決定されており、

1次解析（画像解析およびベースコール）はMiSeqシステム上で行いました。クオリティフィルターしたqseqファイルは、QIIME[8]を用いてオフラインで解析しました。QIIMEはシーケンスの生データから論文投稿に必要な図を作成することができる

図1：V4領域増幅のながれとペアエンドリード



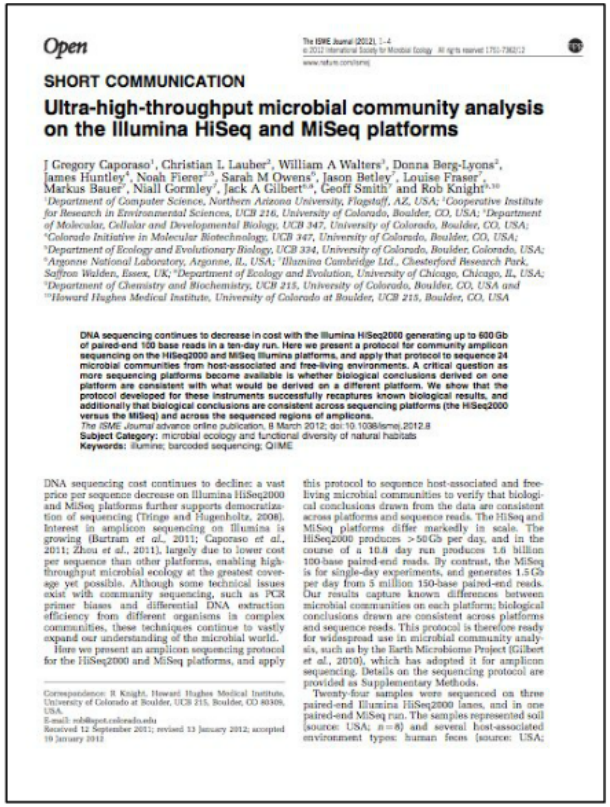
A. 各サンプルのV4領域は515Fと806Rのプライマーを使い増幅しました。プライマーにはP5とP7の配列も組み込まれています。46塩基のオーバーラップをもたらすペアエンドの150塩基のシーケンスランにより、V4領域の全254塩基情報を得ることができます。

B. V4領域の完全な254塩基配列をペアエンドリードで読み取った例。（マトリックスとフェージングを修正後の）生強度を示しています。

ACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTT
ACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAAC
ACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACG
TACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAAC
ACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTT

MiSeq利用文献例: 16Sのメタゲノム解析プロトコル

- 16S V4 領域をカスタムプライマーで増幅
- インデックス配列は 2168 種類記載
- HiSeq, MiSeq に対応
- カスタムプライマーでシーケンス（別途合成が必要）
- MiSeq では PhiX コントロールを 約50% (この時は47%)混ぜてラン
- 実際のサンプルシートの作成方法やランの仕方までsupplementに 記載されている。
- データ解析はMiSeq Reporterでは無く、著者らが開発したQIIMEというソフトウェアを使用
- Amazon Web Services からも資金を獲得



Open

SHORT COMMUNICATION

Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms

J Gregory Caporaso[1], Christian L Lauber[2], William A Walters[3], Donna Berg-Lyons[2], James Huntley[4], Noah Fierer[2,5], Sarah M Owens[6], Jason Betley[7], Louise Fraser[7], Markus Bauer[8], Niall Gormley[7], Jack A Gilbert[6,9], Geoff Smith[7] and Rob Knight[1,10]

[1]Department of Computer Science, Northern Arizona University, Flagstaff, AZ, USA; [2]Cooperative Institute for Research in Environmental Sciences, UCB 216, University of Colorado, Boulder, CO, USA; [3]Department of Molecular, Cellular and Developmental Biology, UCB 347, University of Colorado, Boulder, CO, USA; [4]Colorado Initiative in Molecular Biotechnology, UCB 347, University of Colorado, Boulder, CO, USA; [5]Department of Ecology and Evolutionary Biology, UCB 334, University of Colorado, Boulder, Colorado, USA; [6]Argonne National Laboratory, Argonne, IL, USA; [7]Illumina Cambridge Ltd., Chesterford Research Park, Saffron Walden, Essex, UK; [8]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA; [9]Department of Chemistry and Biochemistry, UCB 215, University of Colorado, Boulder, CO, USA and [10]Howard Hughes Medical Institute, University of Colorado at Boulder, UCB 215, Boulder, CO, USA

DNA sequencing continues to decrease in cost with the Illumina HiSeq2000 generating up to 600 Gb of paired-end 100 base reads in a ten-day run. Here we present a protocol for community amplicon sequencing on the HiSeq2000 and MiSeq Illumina platforms, and apply that protocol to sequence 24 microbial communities from host-associated and free-living environments. A critical question as more sequencing platforms become available is whether biological conclusions derived on one platform are consistent with what would be derived on a different platform. We show that the protocol developed for these instruments successfully recaptures known biological results, and additionally that biological conclusions are consistent across sequencing platforms (the HiSeq2000 versus the MiSeq) and across the sequenced regions of amplicons.

The ISME Journal advance online publication, 8 March 2012; doi:10.1038/ismej.2012.8

Subject Category: microbial ecology and functional diversity of natural habitats

Keywords: Illumina; barcoded sequencing; QIIME

DNA sequencing cost continues to decline; a vast price per sequence decrease on Illumina HiSeq2000 and MiSeq platforms further supports democratization of sequencing (Tringe and Hugenholtz, 2008). Interest in amplicon sequencing on Illumina is growing (Bartram *et al.*, 2011; Caporaso *et al.*, 2011; Zhou *et al.*, 2011), largely due to lower cost per sequence than other platforms, enabling high-throughput microbial ecology at the greatest coverage yet possible. Although some technical issues exist with community sequencing, such as PCR primer biases and differential DNA extraction efficiency from different organisms in complex communities, these techniques continue to vastly expand our understanding of the microbial world.

Here we present an amplicon sequencing protocol for the HiSeq2000 and MiSeq platforms, and apply this protocol to sequence host-associated and free-living microbial communities to verify that biological conclusions drawn from the data are consistent across platforms and sequence reads. The HiSeq and MiSeq platforms differ markedly in scale. The HiSeq2000 produces >50 Gb per day, and in the course of a 10.8 day run produces 1.6 billion 100-base paired-end reads. By contrast, the MiSeq is for single-day experiments, and generates 1.5 Gb per day from 5 million 150-base paired-end reads. Our results capture known differences between microbial communities on each platform; biological conclusions drawn are consistent across platforms and sequence reads. This protocol is therefore ready for widespread use in microbial community analysis, such as by the Earth Microbiome Project (Gilbert *et al.*, 2010), which has adopted it for amplicon sequencing. Details on the sequencing protocol are provided as Supplementary Methods.

Twenty-four samples were sequenced on three paired-end Illumina HiSeq2000 lanes, and in one paired-end MiSeq run. The samples represented soil (source: USA; $n = 8$) and several host-associated environment types: human feces (source: USA;

Correspondence: R Knight, Howard Hughes Medical Institute, University of Colorado at Boulder, UCB 215, Boulder, CO 80309, USA.
E-mail: robknight@colorado.edu

Received 12 September 2011; revised 13 January 2012; accepted 19 January 2012

Caporaso JG, et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ***ISME J***. 2012 Mar 8.

illumina®

Primers for paired-end 16s community sequencing on the
 Illumina HiSeq platform using bacteria/archaeal primer
 515F/806R.

515F (forward primer) PCR primer sequence:

Field number (space-delimited), description:
1, 5' Illumina adapter
2, Forward primer pad
3, Forward primer linker
4, Forward primer

AATGATACGGCGACCACCGAGATCTACAC TATGGTAATT GT GTGCCAGCMGCCGCGGTAA

806R (reverse primer) PCR primer sequence (each sequence contains different barcode):

2168 GoLay barcoded reverse PCR primers. Each primer is followed by a barcode identifier
generated specifically for this set of primers.

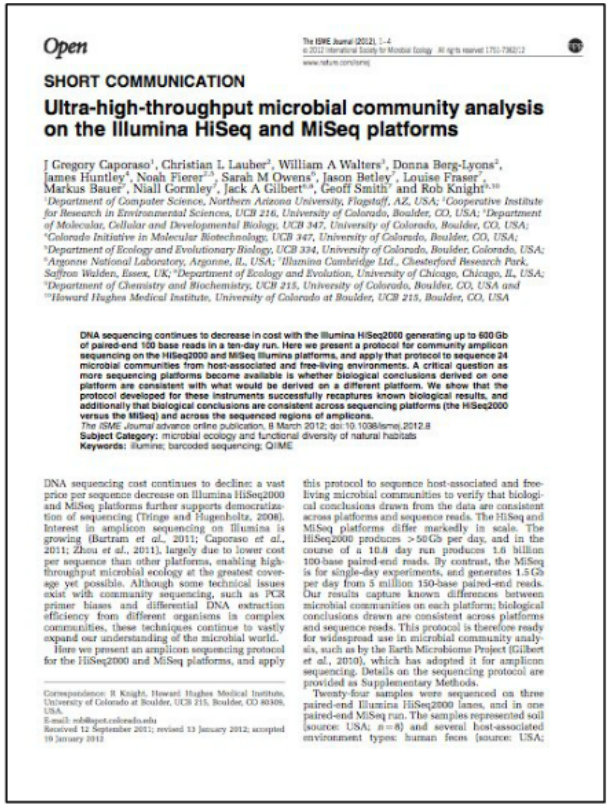
Field number (space-delimited), description:
1, Reverse complement of 3' Illumina adapter
2, Golay barcode
3, Reverse primer pad
4, Reverse primer linker
5, Reverse primer

CAAGCAGAAGACGGCATACGAGAT TCCCTTGTCTCC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc0
CAAGCAGAAGACGGCATACGAGAT ACGAGACTGATT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc1
CAAGCAGAAGACGGCATACGAGAT GCTGTACGGATT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2
CAAGCAGAAGACGGCATACGAGAT ATCACCAGGTGT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc3
CAAGCAGAAGACGGCATACGAGAT TGGTCAACGATA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc4
CAAGCAGAAGACGGCATACGAGAT ATCGCACAGTAA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc5
CAAGCAGAAGACGGCATACGAGAT GTCGTGTAGCCT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc6
CAAGCAGAAGACGGCATACGAGAT AGCGGAGGTTAG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc7
CAAGCAGAAGACGGCATACGAGAT ATCCTTTGGTTC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc8
CAAGCAGAAGACGGCATACGAGAT TACAGCGCATAC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc9
CAAGCAGAAGACGGCATACGAGAT ACCGGTATGTAC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc10
CAAGCAGAAGACGGCATACGAGAT AATTGTGTCGGA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc11
CAAGCAGAAGACGGCATACGAGAT TGCATACACTGG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc12
⋮ ⋮ ⋮
CAAGCAGAAGACGGCATACGAGAT AGCGTCTGAACT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2157
CAAGCAGAAGACGGCATACGAGAT ATCGCGACTGCT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2158
CAAGCAGAAGACGGCATACGAGAT TGGAGGTTCTCA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2159
CAAGCAGAAGACGGCATACGAGAT TGCTTGTAGGCA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2160
CAAGCAGAAGACGGCATACGAGAT CTTAAATGGGCA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2161
CAAGCAGAAGACGGCATACGAGAT GGTATCACCCTG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2162
CAAGCAGAAGACGGCATACGAGAT CGCCTTCATAAC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT 806rcbc2163

```
CAAGCAGAAGACGGCATACGAGAT AGCGGAGGTTAG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc7
CAAGCAGAAGACGGCATACGAGAT ATCCTTTGGTTC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc8
CAAGCAGAAGACGGCATACGAGAT TACAGCGCATAC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc9
CAAGCAGAAGACGGCATACGAGAT ACCGGTATGTAC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc10
CAAGCAGAAGACGGCATACGAGAT AATTGTGTCGGA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc11
CAAGCAGAAGACGGCATACGAGAT TGCATACACTGG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc12
   :                                     :                                        :
   :                                     :                                        :
CAAGCAGAAGACGGCATACGAGAT AGCGTCTGAACT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2157
CAAGCAGAAGACGGCATACGAGAT ATCGCGACTGCT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2158
CAAGCAGAAGACGGCATACGAGAT TGGAGGTTCTCA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2159
CAAGCAGAAGACGGCATACGAGAT TGCTTGTAGGCA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2160
CAAGCAGAAGACGGCATACGAGAT CTTAAATGGGCA AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2161
CAAGCAGAAGACGGCATACGAGAT GGTATCACCCTG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2162
CAAGCAGAAGACGGCATACGAGAT CGCCTTGATAAG AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2163
CAAGCAGAAGACGGCATACGAGAT CGTTTATCCGTT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2164
CAAGCAGAAGACGGCATACGAGAT TTGTACTCACTC AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2165
CAAGCAGAAGACGGCATACGAGAT TTCCCACCCATT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2166
CAAGCAGAAGACGGCATACGAGAT GCCGCATTCGAT AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT        806rcbc2167


-------------------------
Read 1 sequencing primer:
-------------------------
Field number (space-delimited), description:
1, Forward primer pad
2, Forward primer linker
3, Forward primer

TATGGTAATT GT GTGCCAGCMGCCGCGGTAA

-------------------------
Read 2 sequencing primer:
-------------------------
Field number (space-delimited), description:
1, Reverse primer pad
2, Reverse primer linker
3, Reverse primer

AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT

----------------------
Index sequence primer:
----------------------
Field number (space-delimited), description:
1, RC of reverse primer
1, RC of reverse primer linker
1, RC of reverse primer pad

ATTAGAWACCCBDGTAGTCC GG CTGACTGACT
```

MiSeq利用文献例: 16Sのメタゲノム解析プロトコル

- 16S V4 領域をカスタムプライマーで増幅
- インデックス配列は 2168 種類記載
- HiSeq, MiSeq に対応
- カスタムプライマーでシーケンス（別途合成が必要）
- MiSeq では PhiX コントロールを 約50% (この時は47%)混ぜてラン
- 実際のサンプルシートの作成方法やランの仕方までsupplementに 記載されている。
- データ解析はMiSeq Reporterでは無く、著者らが開発したQIIMEというソフトウェアを使用
- Amazon Web Services からも資金を獲得



Open

SHORT COMMUNICATION

Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms

J Gregory Caporaso[1], Christian L Lauber[2], William A Walters[3], Donna Berg-Lyons[2], James Huntley[4], Noah Fierer[2,5], Sarah M Owens[6], Jason Betley[7], Louise Fraser[7], Markus Bauer[8], Niall Gormley[7], Jack A Gilbert[6,9], Geoff Smith[7] and Rob Knight[1,10]

[1]Department of Computer Science, Northern Arizona University, Flagstaff, AZ, USA; [2]Cooperative Institute for Research in Environmental Sciences, UCB 216, University of Colorado, Boulder, CO, USA; [3]Department of Molecular, Cellular and Developmental Biology, UCB 347, University of Colorado, Boulder, CO, USA; [4]Colorado Initiative in Molecular Biotechnology, UCB 347, University of Colorado, Boulder, CO, USA; [5]Department of Ecology and Evolutionary Biology, UCB 334, University of Colorado, Boulder, Colorado, USA; [6]Argonne National Laboratory, Argonne, IL, USA; [7]Illumina Cambridge Ltd., Chesterford Research Park, Saffron Walden, Essex, UK; [8]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA; [9]Department of Chemistry and Biochemistry, UCB 215, University of Colorado, Boulder, CO, USA and [10]Howard Hughes Medical Institute, University of Colorado at Boulder, UCB 215, Boulder, CO, USA

DNA sequencing continues to decrease in cost with the Illumina HiSeq2000 generating up to 600 Gb of paired-end 100 base reads in a ten-day run. Here we present a protocol for community amplicon sequencing on the HiSeq2000 and MiSeq Illumina platforms, and apply that protocol to sequence 24 microbial communities from host-associated and free-living environments. A critical question as more sequencing platforms become available is whether biological conclusions derived on one platform are consistent with what would be derived on a different platform. We show that the protocol developed for these instruments successfully recaptures known biological results, and additionally that biological conclusions are consistent across sequencing platforms (the HiSeq2000 versus the MiSeq) and across the sequenced regions of amplicons.

The ISME Journal advance online publication, 8 March 2012; doi:10.1038/ismej.2012.8

Subject Category: microbial ecology and functional diversity of natural habitats

Keywords: Illumina; barcoded sequencing; QIIME

DNA sequencing cost continues to decline; a vast price per sequence decrease on Illumina HiSeq2000 and MiSeq platforms further supports democratization of sequencing (Tringe and Hugenholtz, 2008). Interest in amplicon sequencing on Illumina is growing (Bartram *et al.*, 2011; Caporaso *et al.*, 2011; Zhou *et al.*, 2011), largely due to lower cost per sequence than other platforms, enabling high-throughput microbial ecology at the greatest coverage yet possible. Although some technical issues exist with community sequencing, such as PCR primer biases and differential DNA extraction efficiency from different organisms in complex communities, these techniques continue to vastly expand our understanding of the microbial world.

Here we present an amplicon sequencing protocol for the HiSeq2000 and MiSeq platforms, and apply this protocol to sequence host-associated and free-living microbial communities to verify that biological conclusions drawn from the data are consistent across platforms and sequence reads. The HiSeq and MiSeq platforms differ markedly in scale. The HiSeq2000 produces >50 Gb per day, and in the course of a 10.8 day run produces 1.6 billion 100-base paired-end reads. By contrast, the MiSeq is for single-day experiments, and generates 1.5 Gb per day from 5 million 150-base paired-end reads. Our results capture known differences between microbial communities on each platform; biological conclusions drawn are consistent across platforms and sequence reads. This protocol is therefore ready for widespread use in microbial community analysis, such as by the Earth Microbiome Project (Gilbert *et al.*, 2010), which has adopted it for amplicon sequencing. Details on the sequencing protocol are provided as Supplementary Methods.

Twenty-four samples were sequenced on three paired-end Illumina HiSeq2000 lanes, and in one paired-end MiSeq run. The samples represented soil (source: USA; $n = 8$) and several host-associated environment types: human feces (source: USA;

Correspondence: R Knight, Howard Hughes Medical Institute, University of Colorado at Boulder, UCB 215, Boulder, CO 80309, USA.
E-mail: robknight@colorado.edu

Received 12 September 2011; revised 13 January 2012; accepted 19 January 2012

Caporaso JG, et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ***ISME J***. 2012 Mar 8.

illumina®

16S rRNA Amplification Protocol

PCRを3連で実施 ⟶ TaKaRa Ex Taq HS （PrimeSTAR® HS などHigh Fidelityなタイプはアダプターの付属した長いプライマーと反応性が低い）

Thermocycler condition

Temp	Time	
94℃	3 min	
94℃	45 sec	20 (~25) cycle
50℃	1 min	
72℃	1.5 min	
72℃	10 min	
4℃	hold	

↓

電気泳動にて増幅を確認

↓

カラムを用いたキットにてPCR産物を精製

↓

Picogreenで全サンプルを濃度測定

↓

一度に解析するサンプルを等量混合

↓

電気泳動にてバンドを切り出す
（primer dimerが100bp以上になりカラムで精製しきれないため）

↓

精製した混合PCR産物をPicogreenで濃度測定し、2nMまで希釈
（DNA濃度を正確に測定することが最重要）

↓

マニュアルに従いPCR産物とPhiXを別々に変性・希釈し20pM溶液を作成

↓

16pMのPCR産物と6pMのPhiX溶液を混合し解析サンプルとする
（イルミナではPhiXの30~50%混合を推奨）

↓

Read1,index,Read2用のCustom primerをそれぞれ12,13,14番のポートに0.5μMずつ混合
（PhiX用のプライマーが必要なためカスタムポートである18,19,20番は使用しない）

森
食
小

Illumina Sequencing Analysis Viewer 1.8.20 - 120525_M00700_0008_A000000000-A1065

Sequencing Analysis Viewer

Run Folder: F:\121030MiSeqdata-backup\MiSeqAnalysis\120525_M00700_0008_A000000000-A1065 Browse Refresh

Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

Run Summary

Level	Yield Total (G)	Projected Total Yield (G)	Yield Perfect (G)	Yield <=3 errors (G)	Aligned (%)	% Perfect [Num Cycles]	% <=3 errors [Num Cycles]	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 20	% >= Q30
Read 1	0.5	0.5	0.2	0.3	50.19	68.1 [150]	97.2 [150]	0.58	892	98.9	82.0
Read 2 (I)	0.0	0.0	0.0	0.0	0.00	0.0 [11]	0.0 [11]	0.00	635	0.0	32.8
Read 3	0.5	0.5	0.2	0.3	49.02	64.4 [150]	96.3 [150]	0.73	723	77.3	84.2
Total	1.1	1.1	0.4	0.5	49.60	66.2	96.8	0.66	750	88.1	81.2

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
1	12	459 +/- 6	93.59 +/- 0.40	0.123 / 0.438	3.89	3.64	82.0	0.5	150	50.2 +/- 0.4	0.58 +/- 0.04	0.25 +/- 0.06	0.23 +/- 0.03	0.28 +/- 0.03	892 +/- 22	98.9 +/- 2.5

Read 2 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
1	12	459 +/- 6	93.59 +/- 0.40	0.074 / 0.136	3.89	3.64	32.8	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	635 +/- 21	0.0 +/- 0.0

Read 3

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
1	12	459 +/- 6	93.59 +/- 0.40	0.285 / 0.323	3.89	3.64	84.2	0.5	150	49.0 +/- 0.5	0.73 +/- 0.04	0.44 +/- 0.08	0.45 +/- 0.05	0.51 +/- 0.04	723 +/- 13	77.3 +/- 1.2

Copy to Clipboard... Generate IVC Plots...

15:49
2013/01/21

Illumina Sequencing Analysis Viewer 1.8.20 - 120525_M00700_0008_A000000000-A1065

Sequencing Analysis Viewer

Run Folder: F:\121030MiSeqdata-backup\MiSeqAnalysis\120525_M00700_0008_A000000000-A1065 Browse Refresh

Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

Run Summary

Level	Yield Total (G)	Projected Total Yield (G)	Yield Perfect (G)	Yield <=3 errors (G)	Aligned (%)	% Perfect [Num Cycles]	% <=3 errors [Num Cycles]	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 20	% >= Q30
Read 1	0.5	0.5	0.2	0.3	50.19	68.1 [150]	97.2 [150]	0.58	892	98.9	82.0
Read 2 (I)	0.0	0.0	0.0	0.0	0.00	0.0 [11]	0.0 [11]	0.00	635	0.0	32.8
Read 3	0.5	0.5	0.2	0.3	49.02	64.4 [150]	96.3 [150]	0.73	723	77.3	84.2
Total	1.1	1.1	0.4	0.5	49.60	66.2	96.8	0.66	750	88.1	81.2

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
1	12	459 +/- 6	93.59 +/- 0.40	0.123 / 0.438	3.89	3.64	82.0	0.5	150	50.2 +/- 0.4	0.58 +/- 0.04	0.25 +/- 0.06	0.23 +/- 0.03	0.28 +/- 0.03	892 +/- 22	98.9 +/- 2.5

Read 2 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
1	12	459 +/- 6	93.59 +/- 0.40	0.074 / 0.136	3.89	3.64	32.8	0.0	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	635 +/- 21	0.0 +/- 0.0

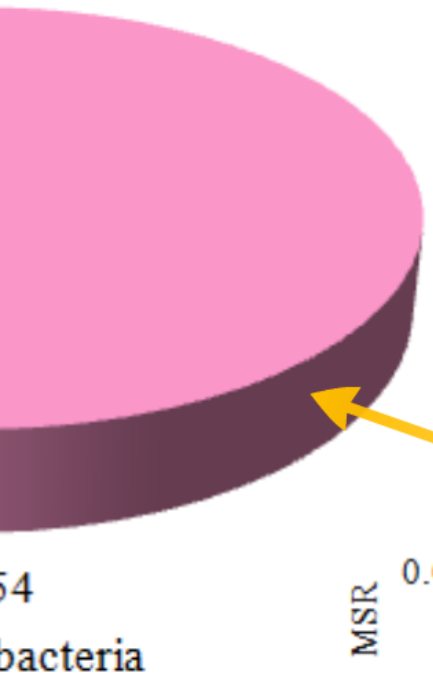
Read 3

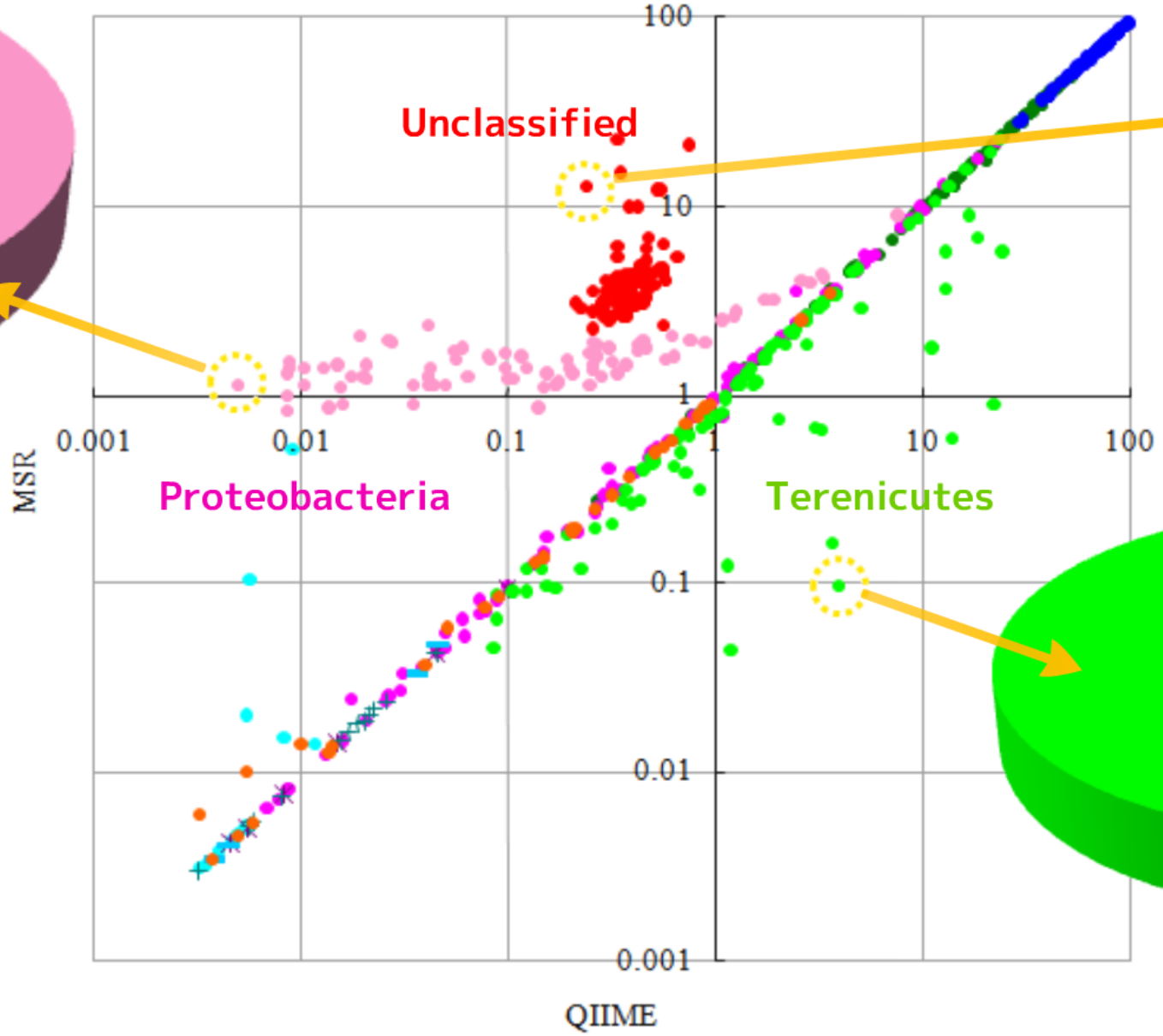
Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
1	12	459 +/- 6	93.59 +/- 0.40	0.285 / 0.323	3.89	3.64	84.2	0.5	150	49.0 +/- 0.5	0.73 +/- 0.04	0.44 +/- 0.08	0.45 +/- 0.05	0.51 +/- 0.04	723 +/- 13	77.3 +/- 1.2

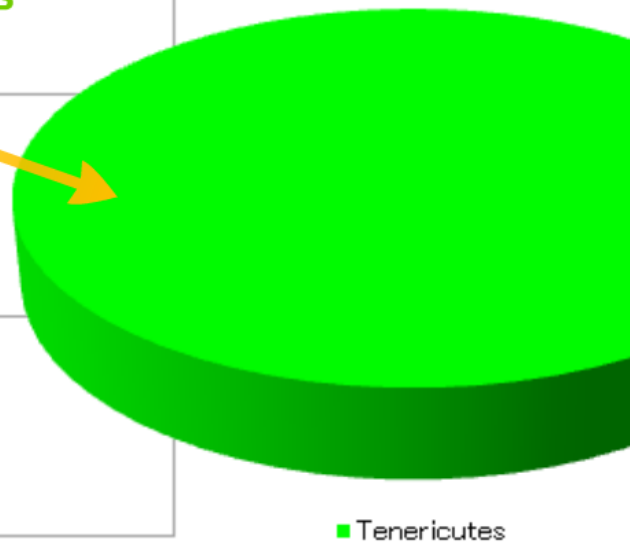
Copy to Clipboard... Generate IVC Plots...

15:49
2013/01/21

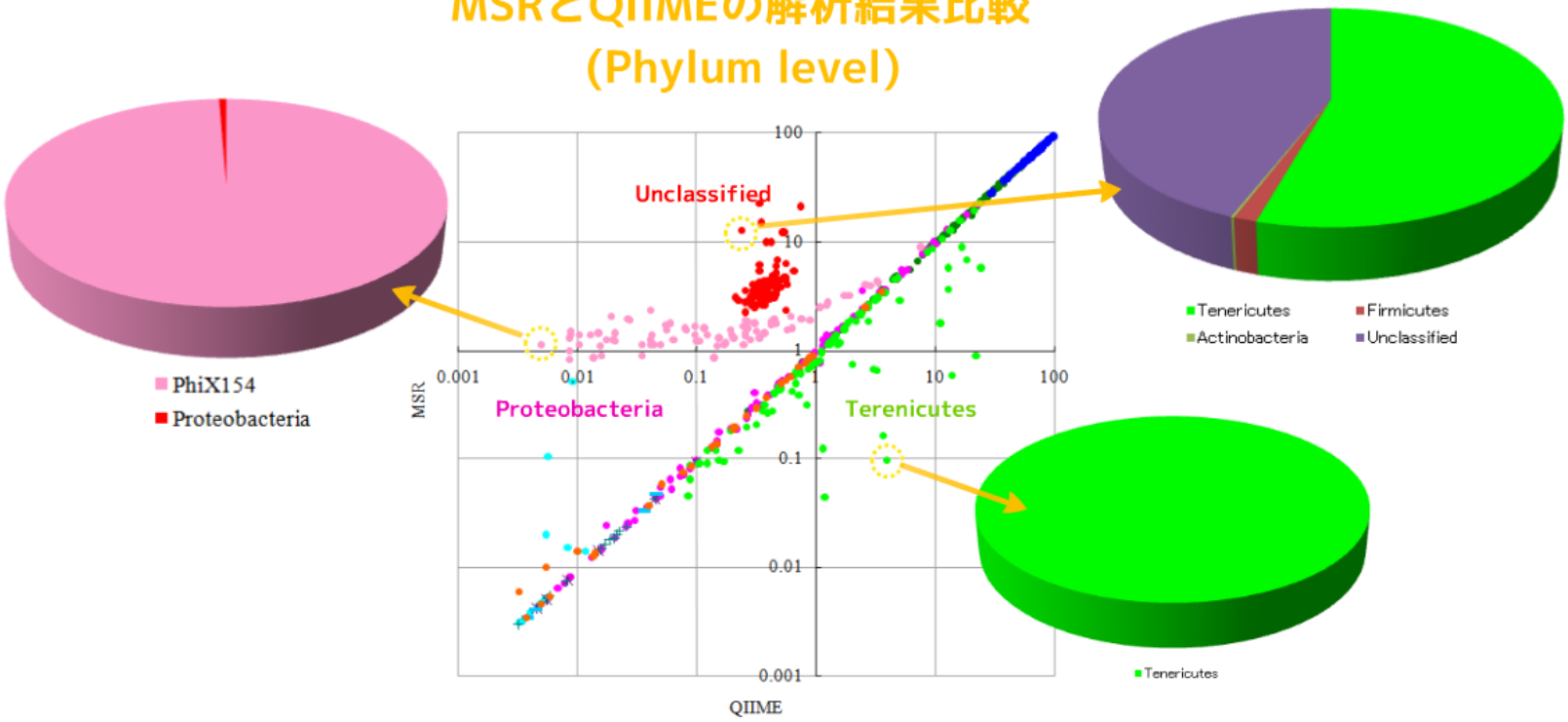
MSRとQIIMEの解析結果比較
(Phylum level)







MSRとQIIMEの解析結果比較
(Phylum level)



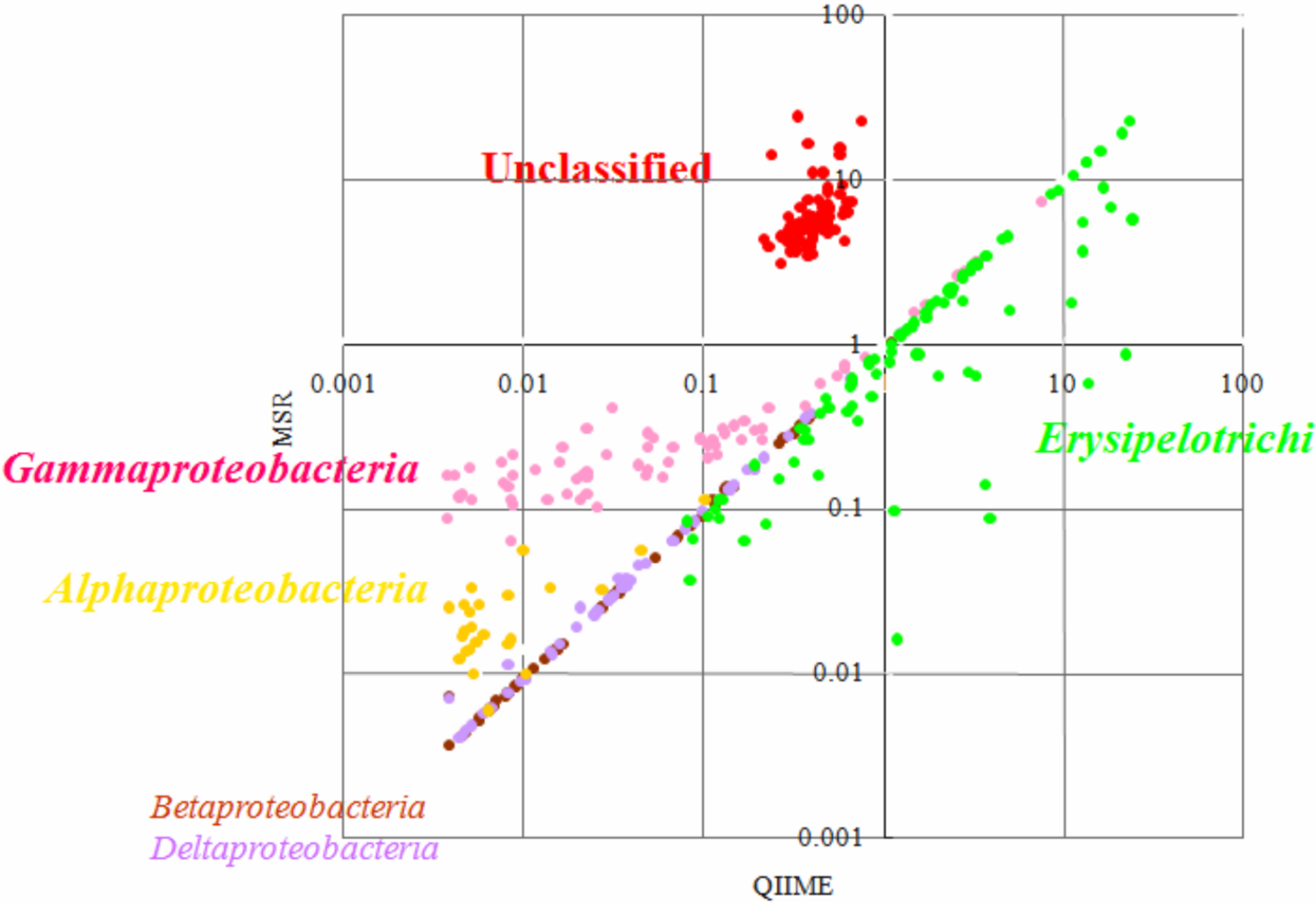
MSRの結果は

・PhiXの一部をProteobacteriaに

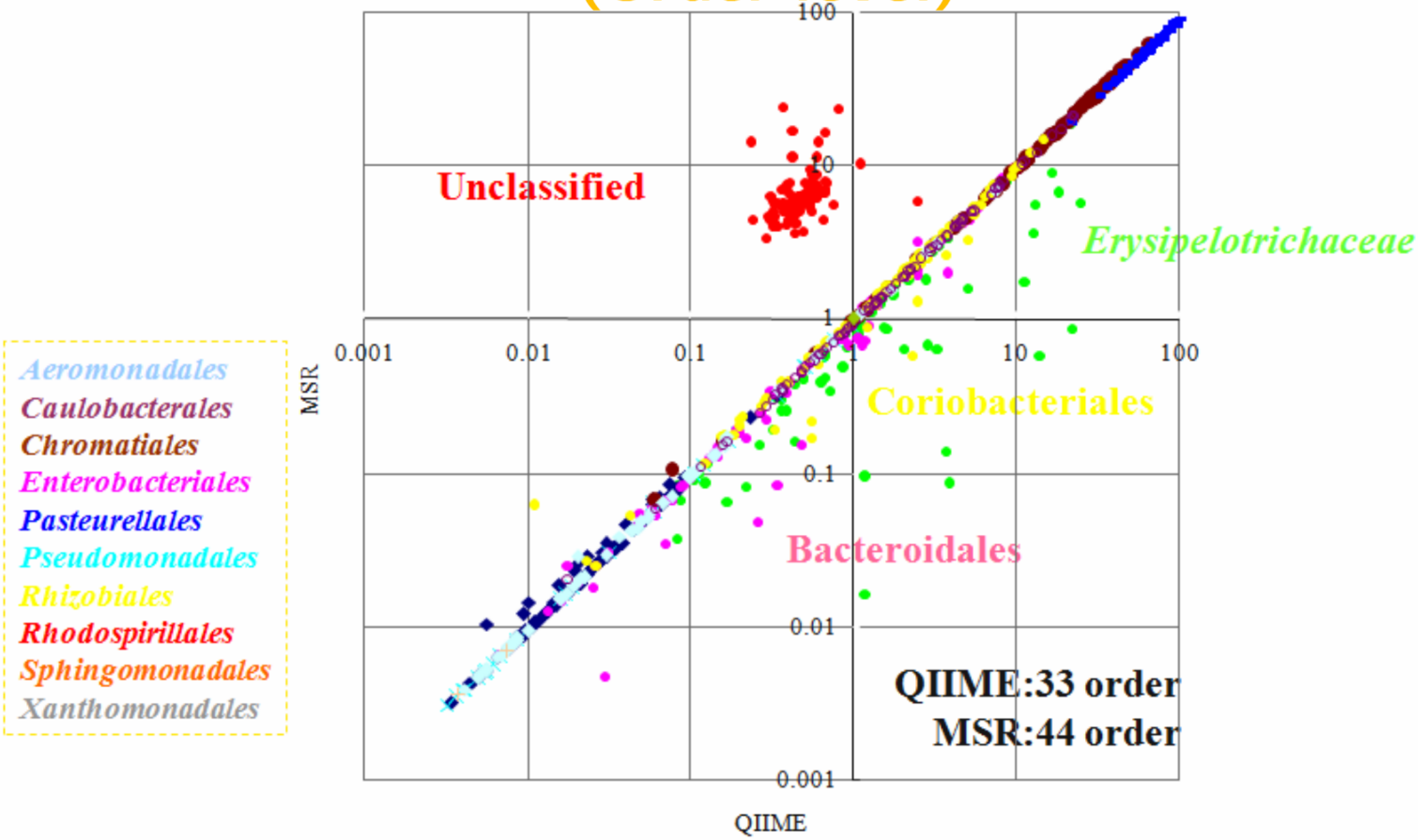
・Tenericutes(正しくはFirmicutes)の一部をUnclassified

振分けてしまっている可能性がある。

(Class level)


Unclassified
Gammaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Deltaproteobacteria
Erysipelotrichi
100
10
1
0.1
0.01
0.001
0.001
0.01
0.1
1
10
100
MSR
QIIME

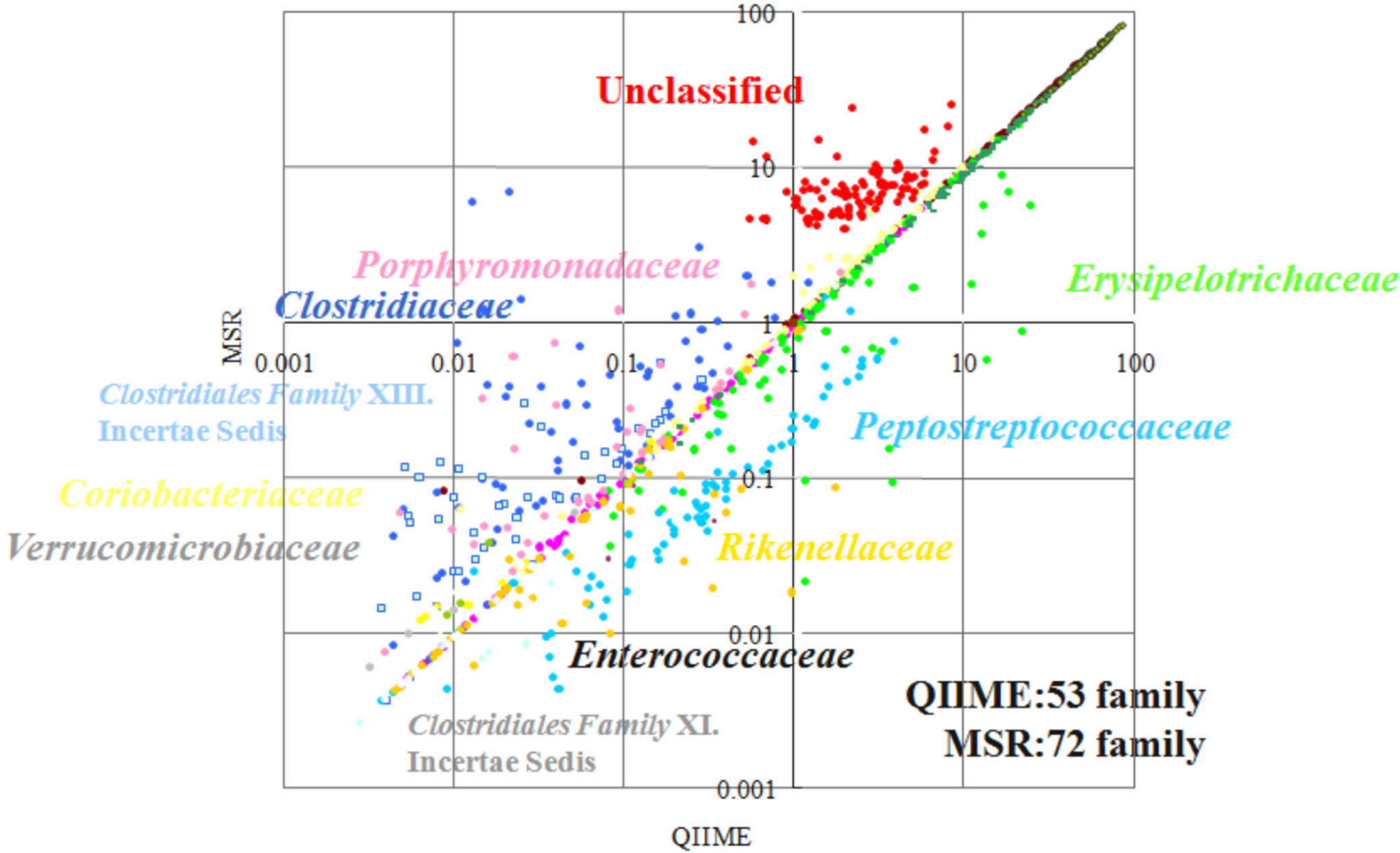
MSRとQIIMEの解析結果比較 (Order level)



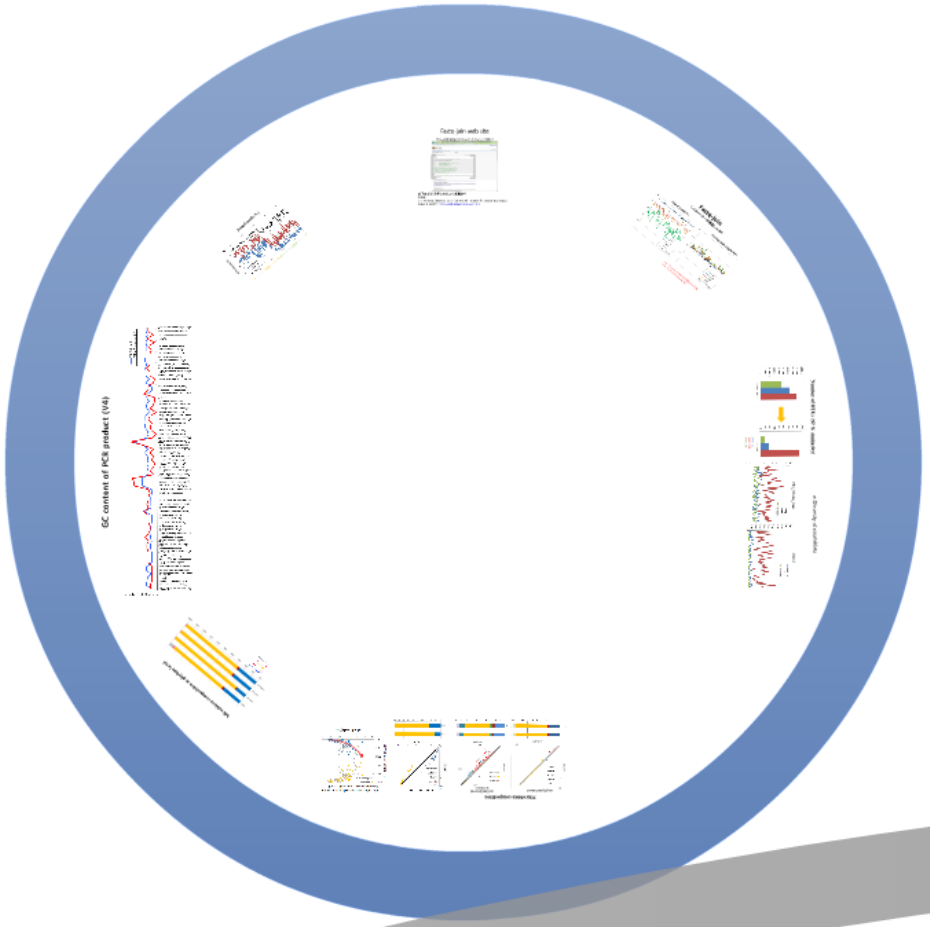
Order levelになると、phiX配列がUnclassifiedに移行

MSRとQIIMEの解析結果比較

(Family level)



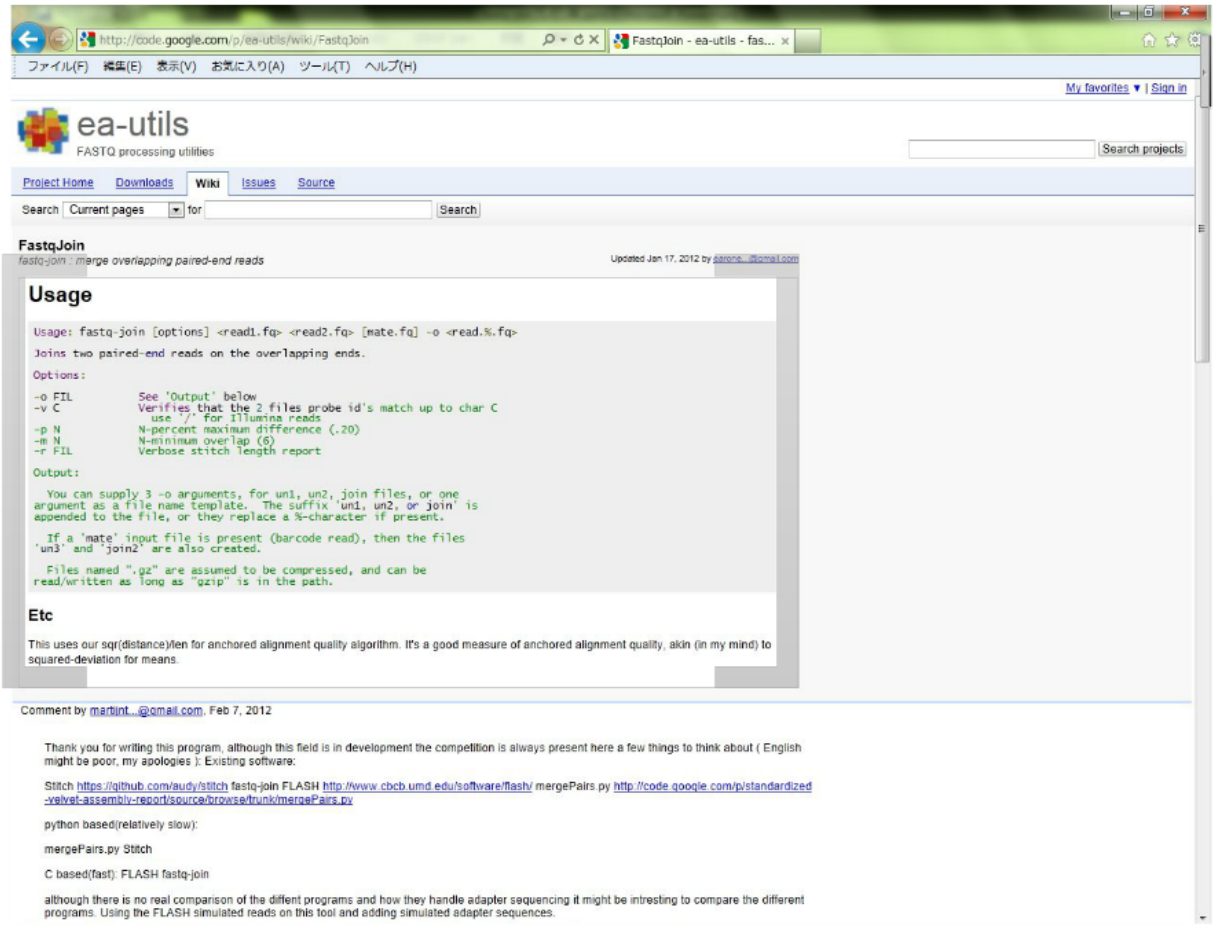
Read1,2の結合



析

Fastq-join web site

http://code.google.com/p/ea-utils/wiki/FastqJoin



以下のように引用してほしいと記載あり

Citing:

Erik Aronesty (2011). ea-utils : "Command-line tools for processing biological sequencing data"; http://code.google.com/p/ea-utils

FastqJoin

fastq-join : merge overlapping paired-end reads

Updated Jan 17, 2012 by earone...@gmail.com

Usage

```
Usage: fastq-join [options] <read1.fq> <read2.fq> [mate.fq] -o <read.%.fq>

Joins two paired-end reads on the overlapping ends.

Options:

-o FIL          See 'Output' below
-v C            Verifies that the 2 files probe id's match up to char C
                  use '/' for Illumina reads
-p N            N-percent maximum difference (.20)
-m N            N-minimum overlap (6)
-r FIL          Verbose stitch length report

Output:

  You can supply 3 -o arguments, for un1, un2, join files, or one
argument as a file name template.  The suffix 'un1, un2, or join' is
appended to the file, or they replace a %-character if present.

  If a 'mate' input file is present (barcode read), then the files
'un3' and 'join2' are also created.

  Files named ".gz" are assumed to be compressed, and can be
read/written as long as "gzip" is in the path.
```

Etc

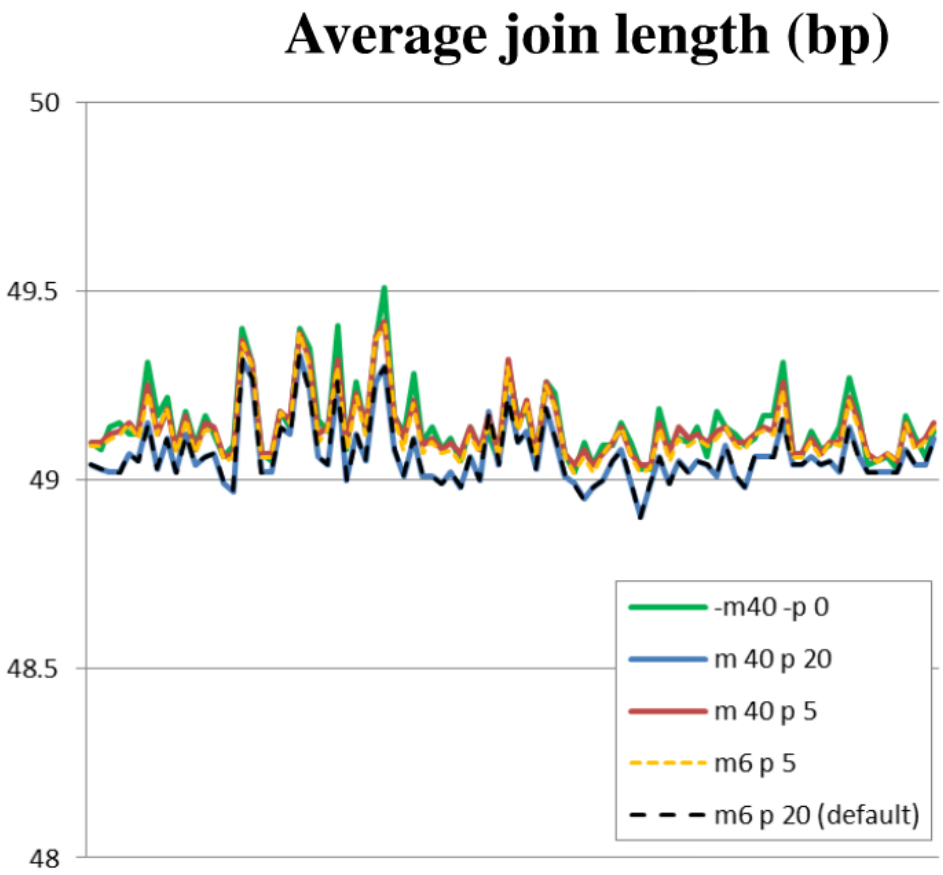
This uses our sqr(distance)/len for anchored alignment quality algorithm. It's a good measure of anchored alignment quality, akin (in my mind) to squared-deviation for means.

Comment by martijnt...@gmail.com, Feb 7, 2012

Thank you for writing this program, although this field is in development the competition is always present here a few things to think about (English might be poor, my apologies): Existing software:

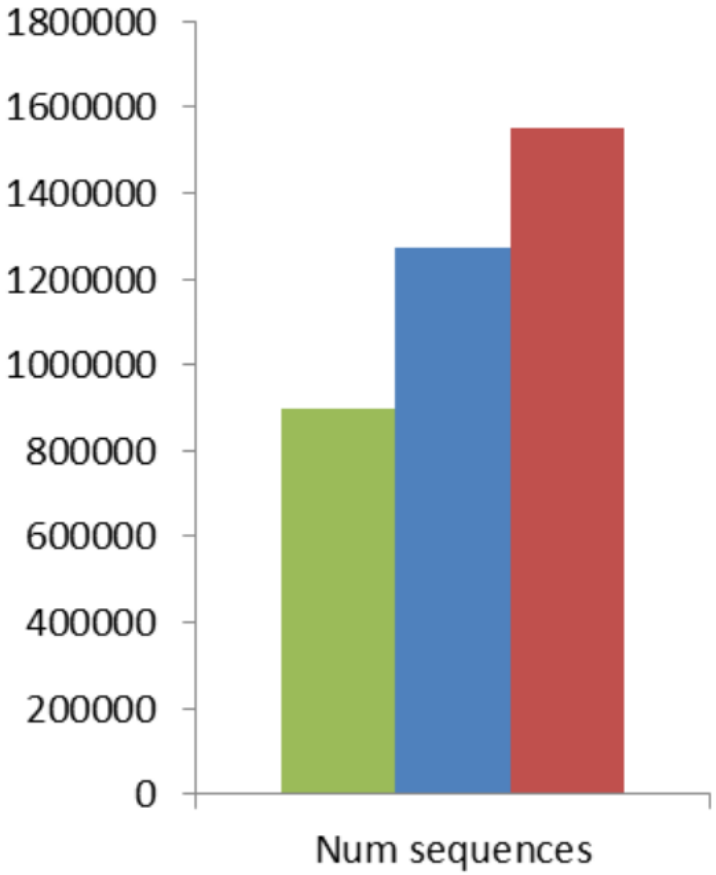
Fastq-join

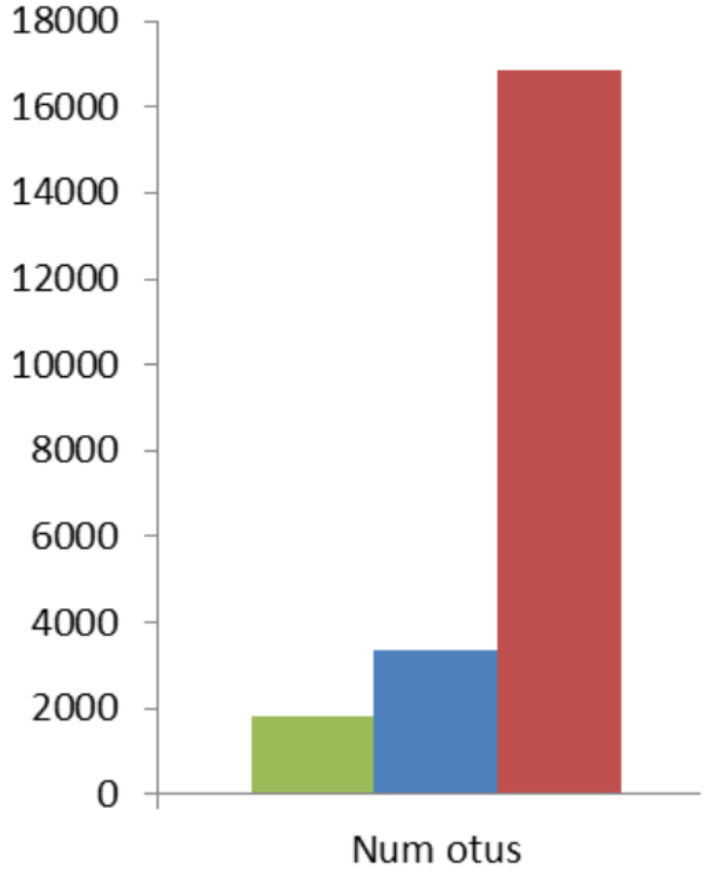
（パラメーターの検証、n=90）

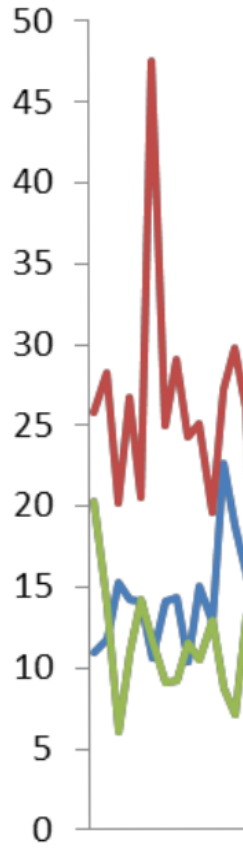


-p N　N-percent maximum difference (.20)
-m N　N-minimum overlap (6)

Number of OTUs (97 % similarity)



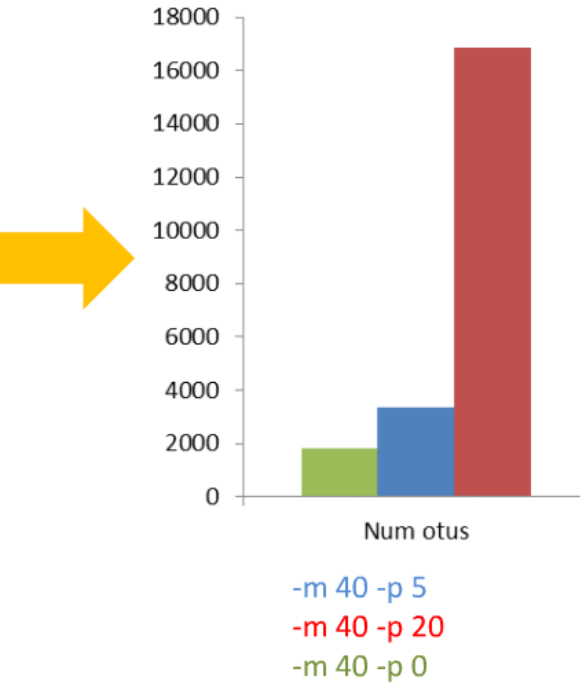


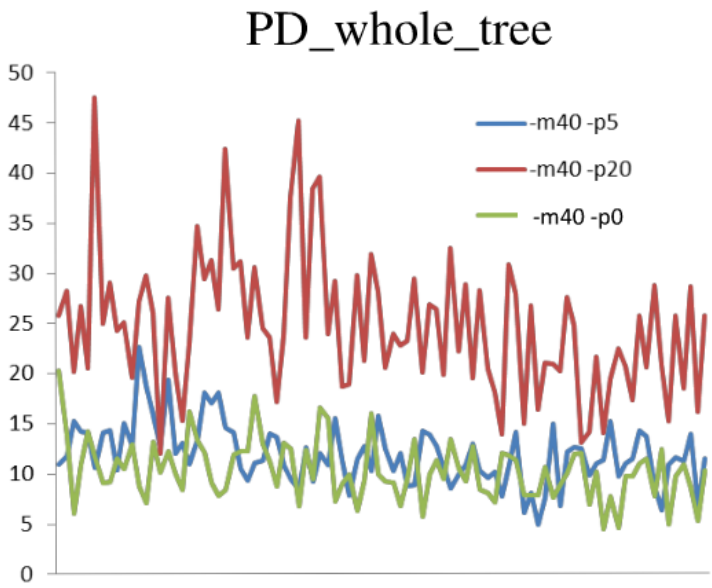


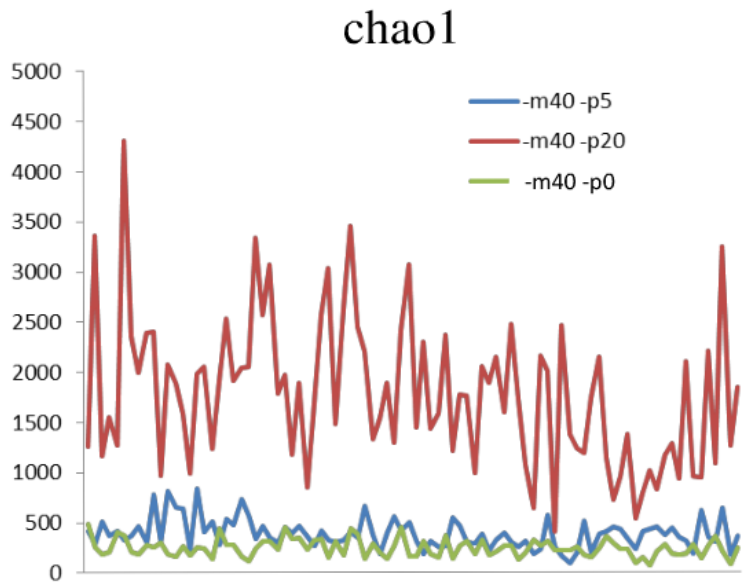
-m 40 -p 5
-m 40 -p 20
-m 40 -p 0

TUs (97 % similarity)

α-Diversity of microbiota

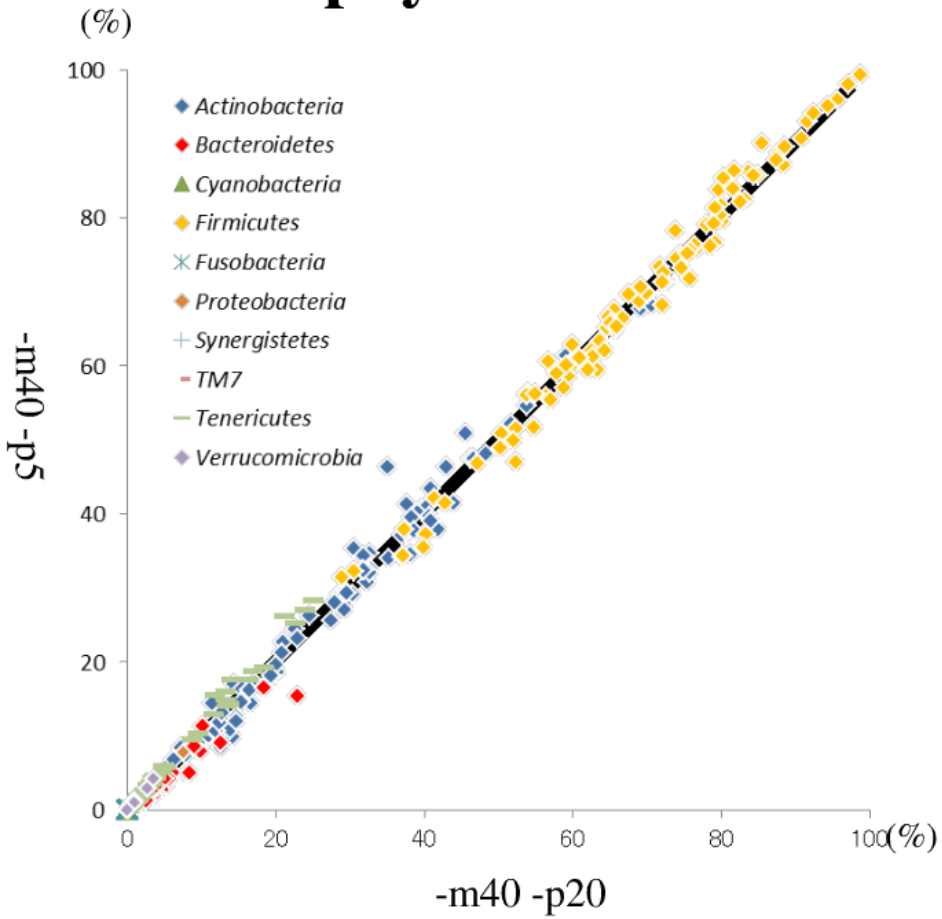


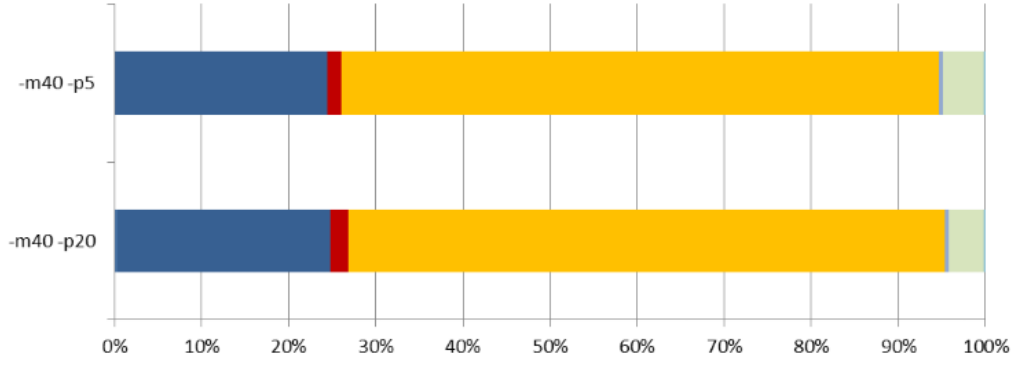




Microbiota composition

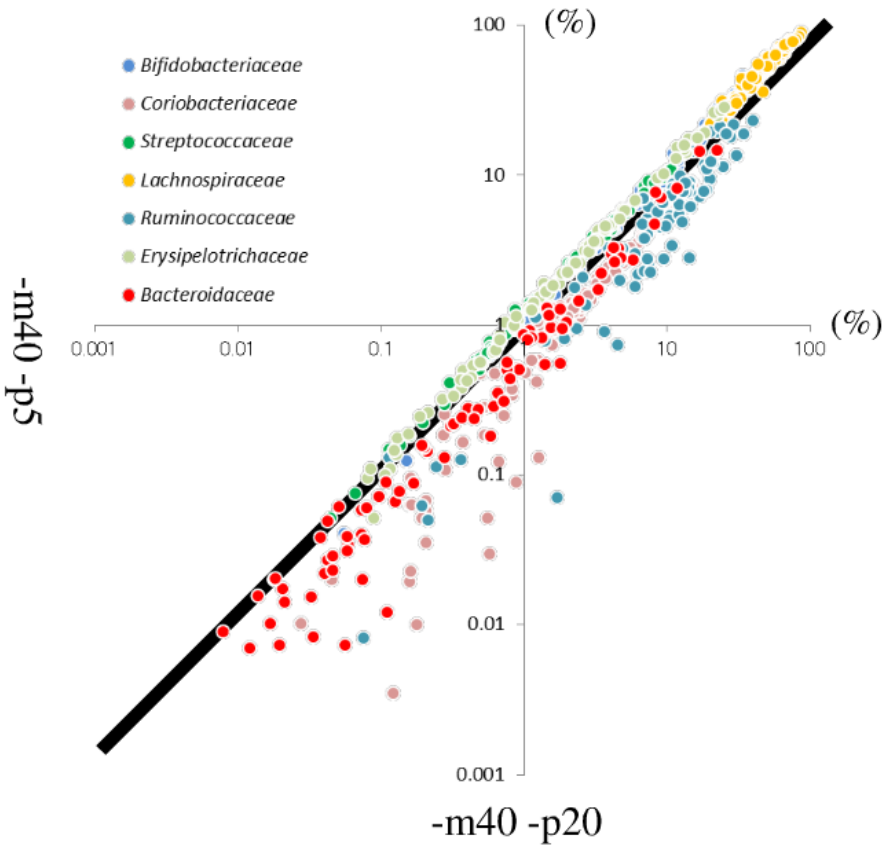
at phylum level

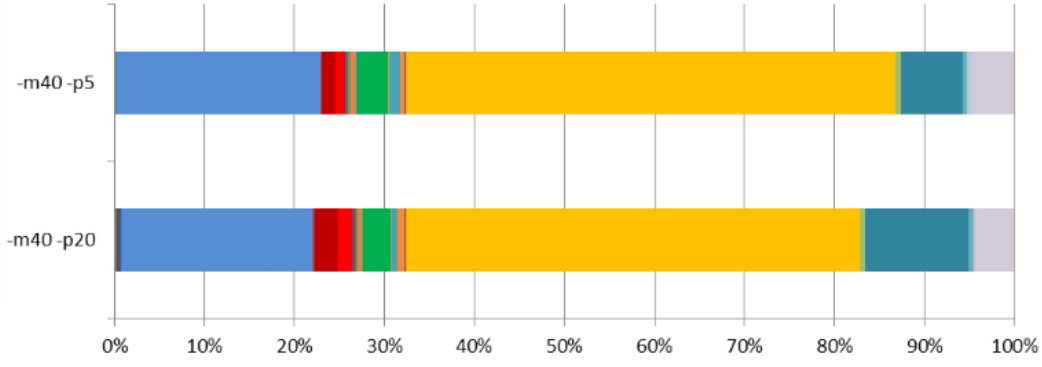




at family level

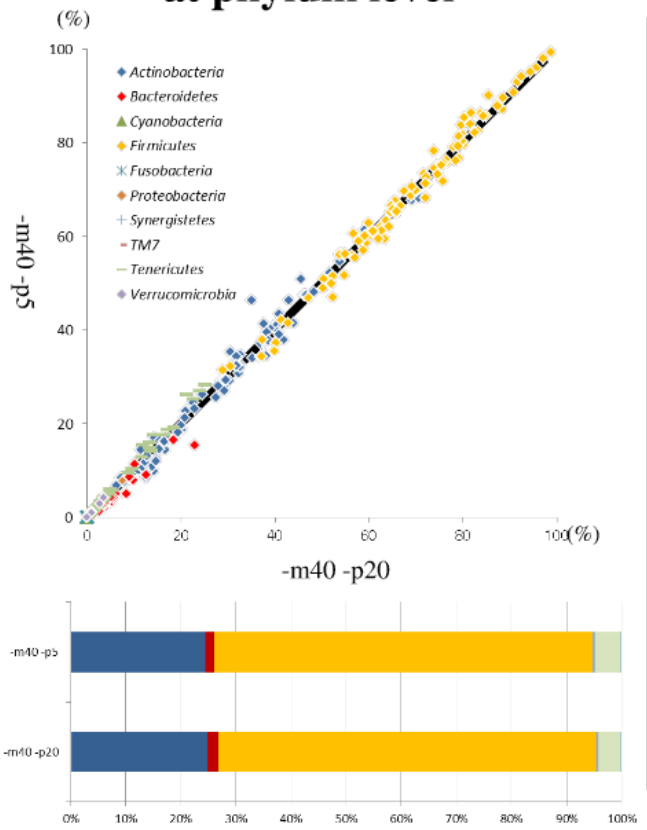
(predominant)





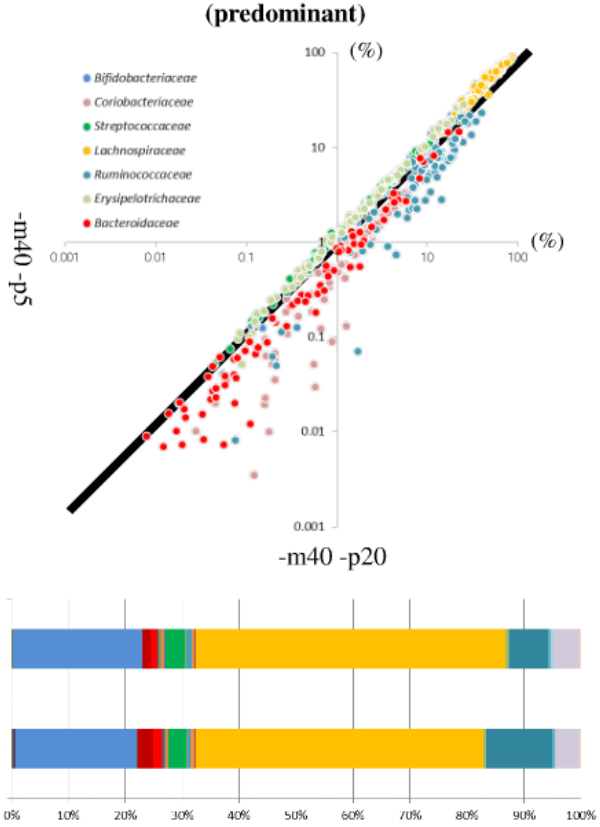
Microbiota composition

at phylum level

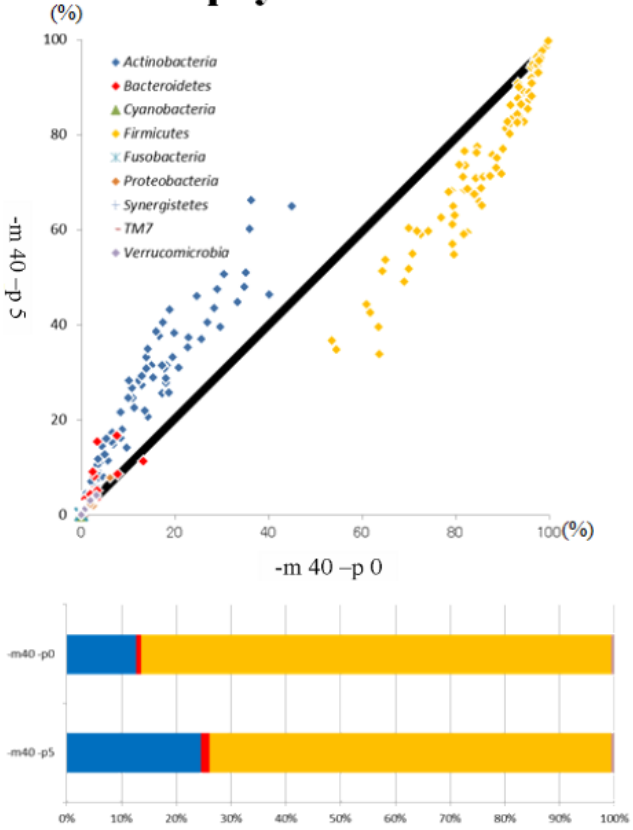


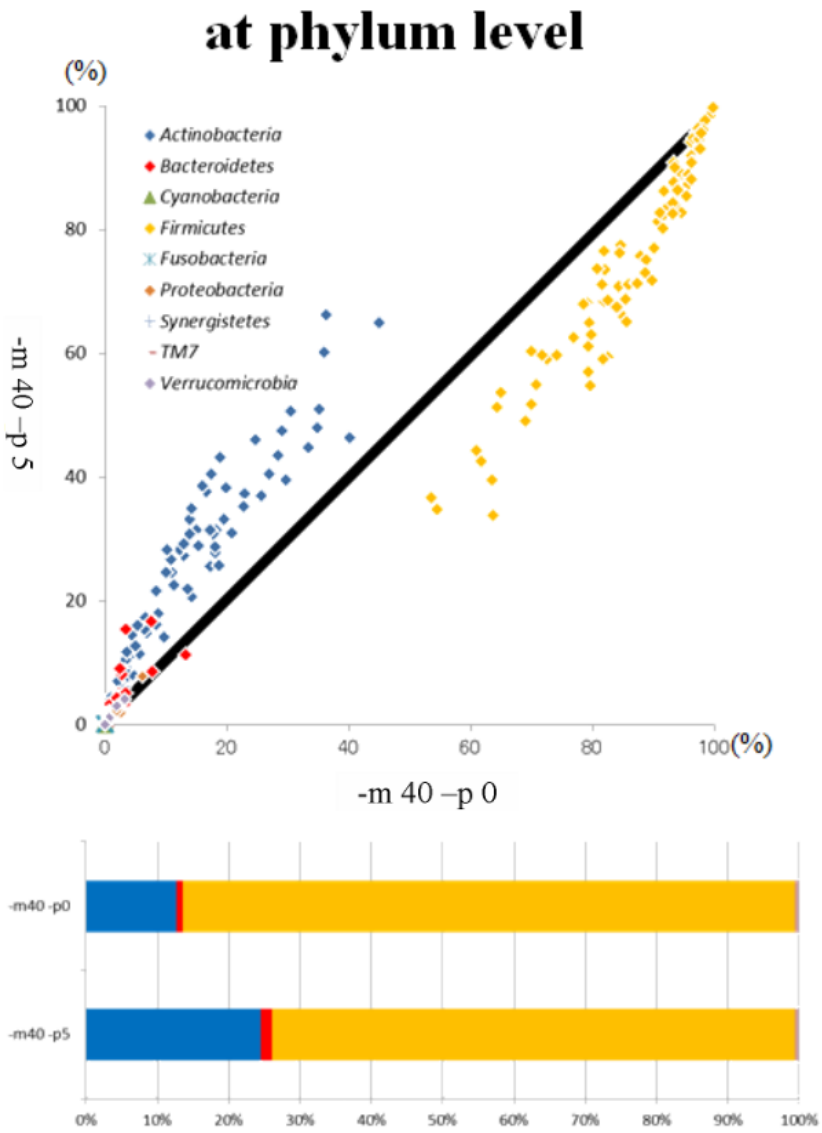
at family level

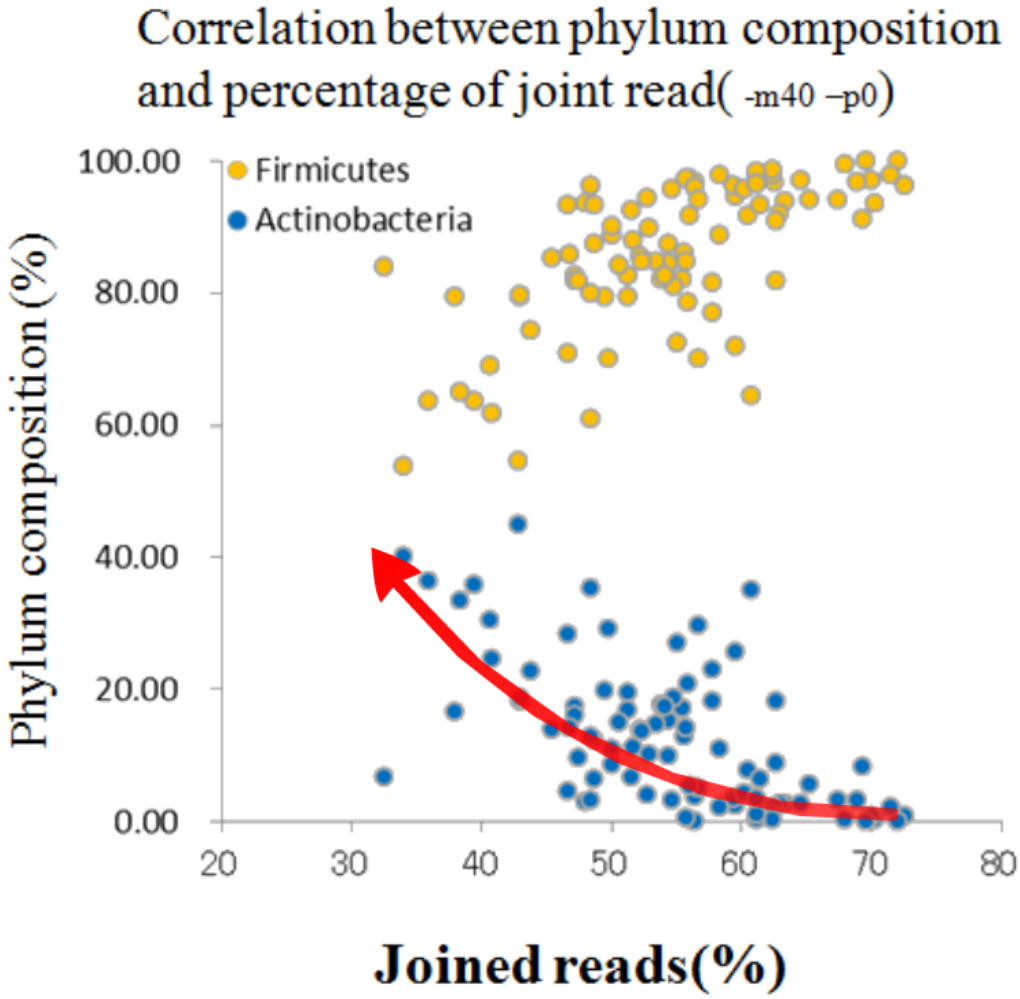
(predominant)



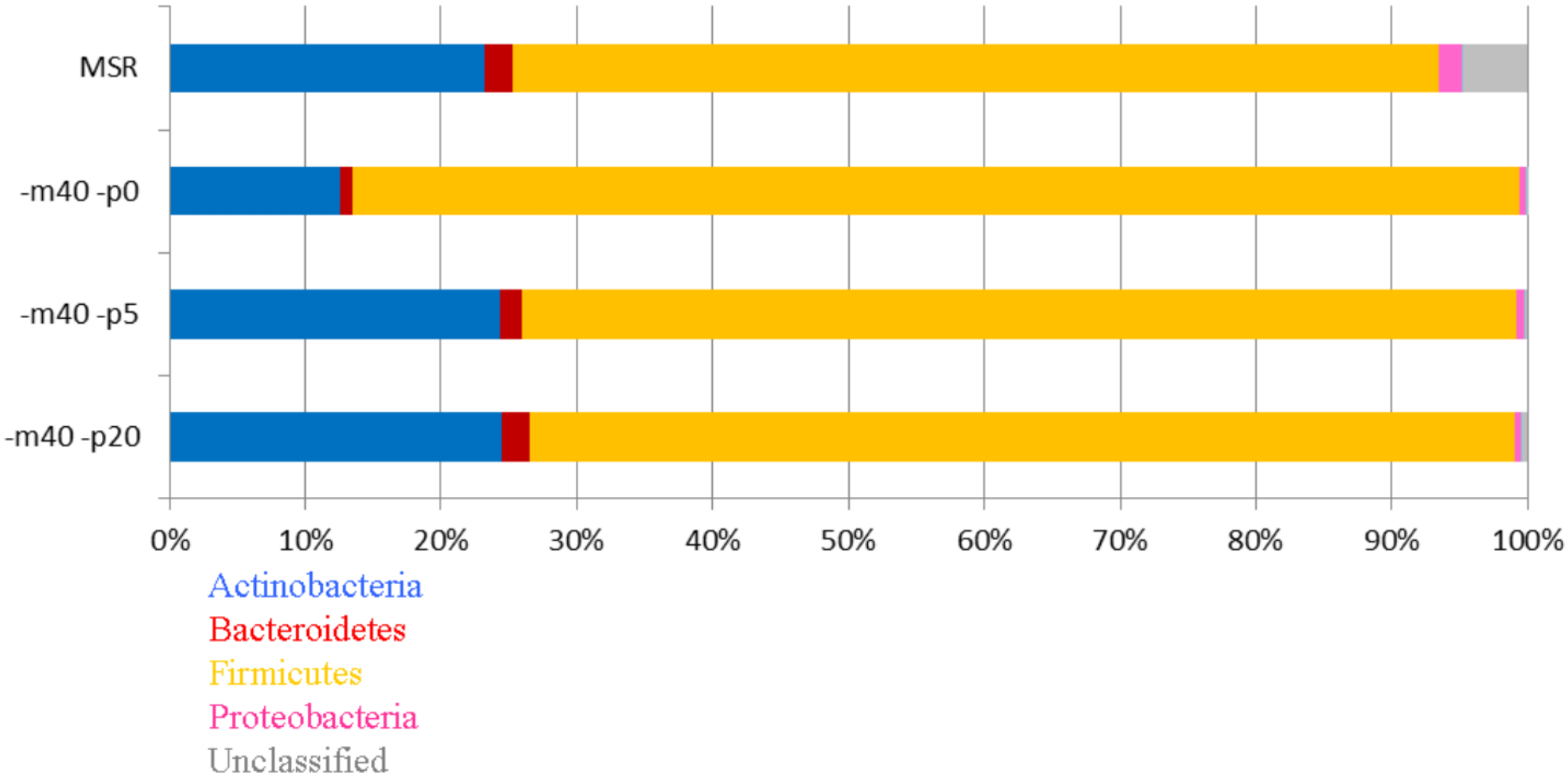
at phylum level



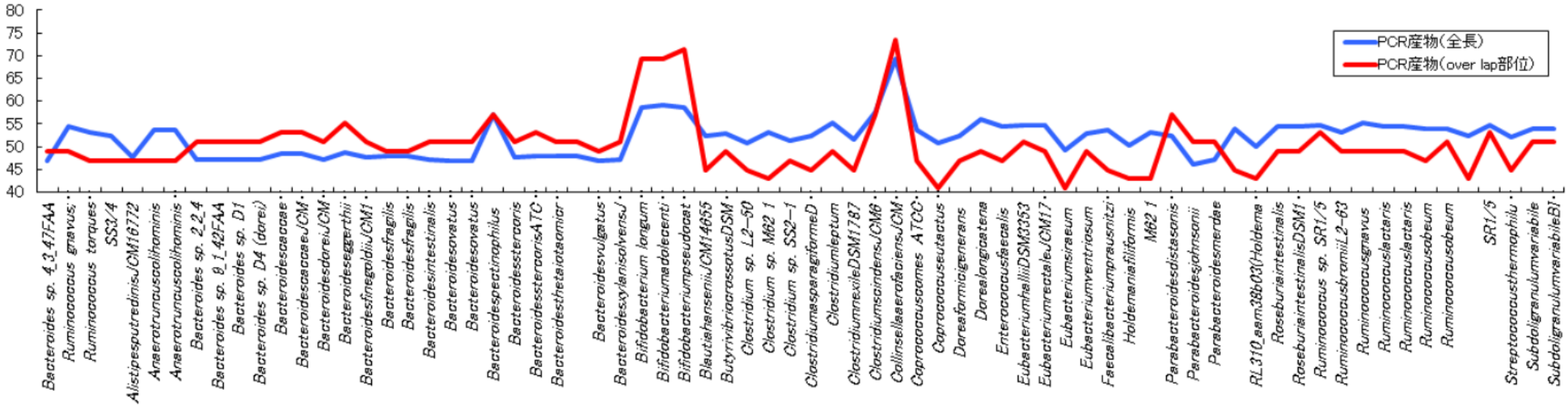

at phylum level
(%)
100
80
60
40
20
0
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Fusobacteria
Proteobacteria
Synergistetes
TM7
Verrucomicrobia
-m 40 –p 5
0
20
40
60
80
100(%)
-m 40 –p 0
-m40 -p0
-m40 -p5
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


Correlation between phylum composition and percentage of joint read(-m40 –p0)
Firmicutes
Actinobacteria
Phylum composition (%)
100.00
80.00
60.00
40.00
20.00
0.00
20
30
40
50
60
70
80
Joined reads(%)

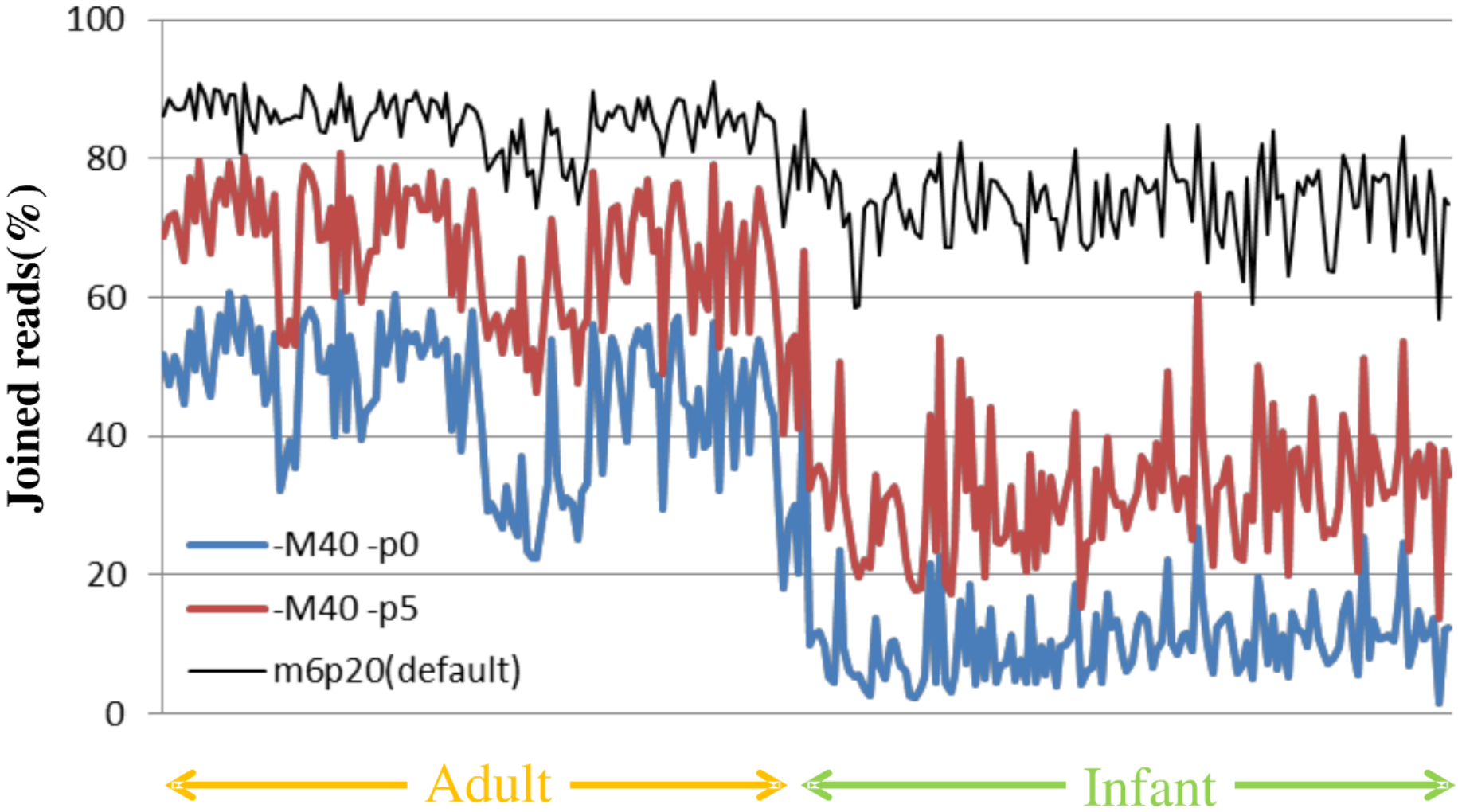
Microbiota composition at phylum level



GC content of PCR product (V4)

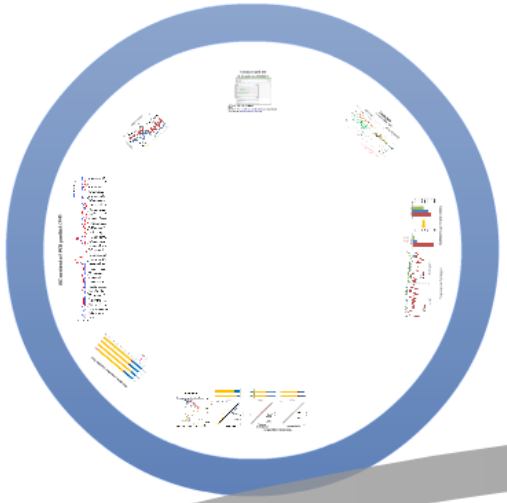


Joined reads(%)



Read1,2の結合

⇒６塩基程



解析

qiime
Quantitative Insights Into Microbial Ecology

QIIME用ファイルへの変換

1．fastqからfastaへファイル形式を変換

2．fastaファイル内のシーケンス名を変換

Miseqからのoutput(fasta変換後）

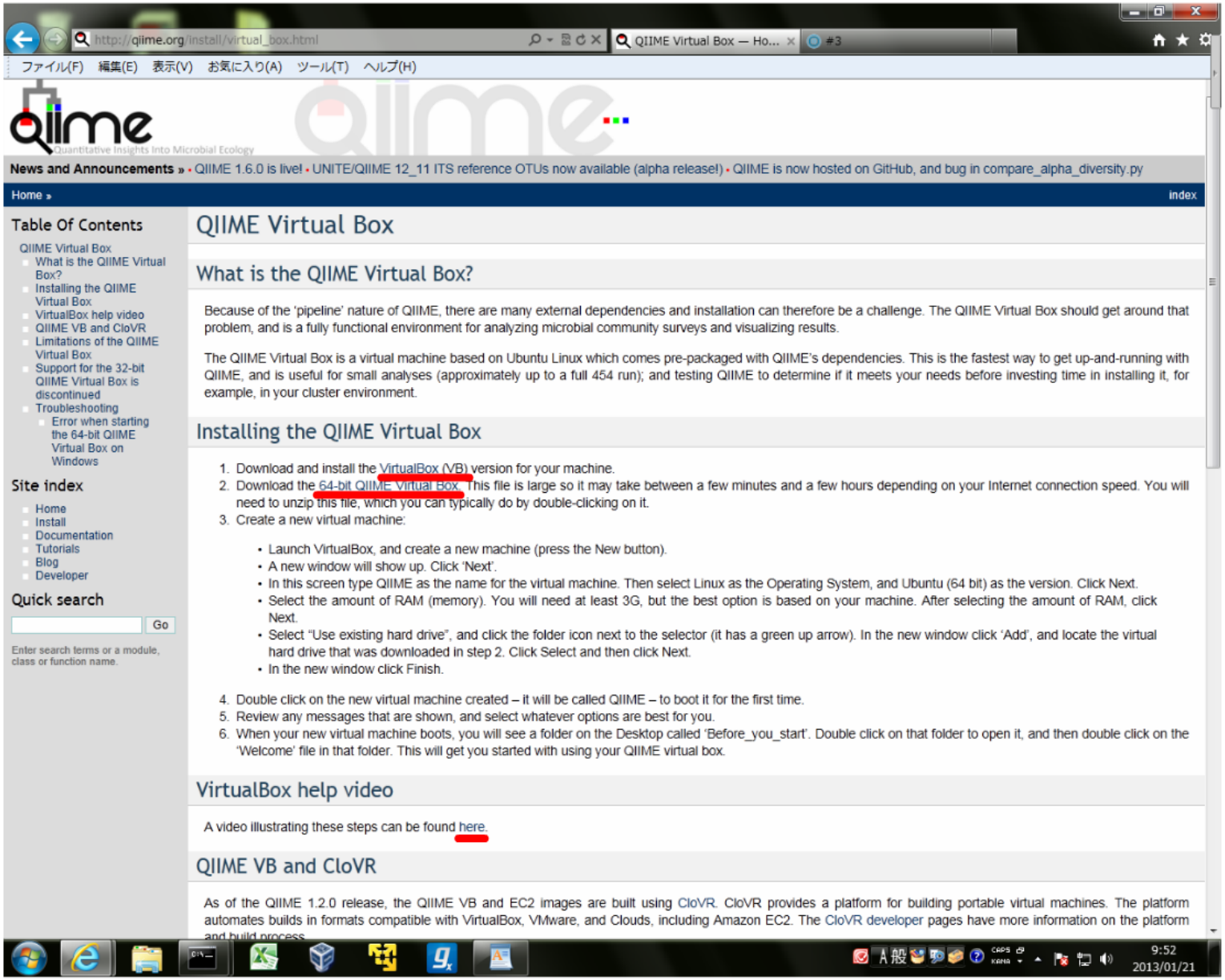
```
>M00752:12:000000000-A2F5C:1:1106:21667:6702 1:N:0:1
TGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCTGGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGTCCCTAGTTTT
>M00752:12:000000000-A2F5C:1:1106:5074:10091 1:N:0:1
ATCATGGCGACCATCCAAAGGATAAACATCATAGGCAGTCGGGAGGGTAGTCGGAACCGAAGAAGACTCAAAGCGAACCAAACAGGCAAAAAATTTAGGGTCGGCATCAAAAGCAATATCA
>M00752:12:000000000-A2F5C:1:1109:7833:13253 1:N:0:1
CACTCCGTGGACAGATTTGTCATTGTGAGCATTTTCATCCCGAAGTTGCGGCTCATTCTGATTCTGAACAGCTTCTTGGGAAGTAGCGACAGCTTGGTTTTTAGTGAGTTGTTCCATTCTT
```

QIIME解析用のfastaファイル

```
>101_1
TGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCTGGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGTCCCTAGTTTT
>101_2
ATCATGGCGACCATCCAAAGGATAAACATCATAGGCAGTCGGGAGGGTAGTCGGAACCGAAGAAGACTCAAAGCGAACCAAACAGGCAAAAAATTTAGGGTCGGCATCAAAAGCAATATCA
>101_3
CACTCCGTGGACAGATTTGTCATTGTGAGCATTTTCATCCCGAAGTTGCGGCTCATTCTGATTCTGAACAGCTTCTTGGGAAGTAGCGACAGCTTGGTTTTTAGTGAGTTGTTCCATTCTT
```

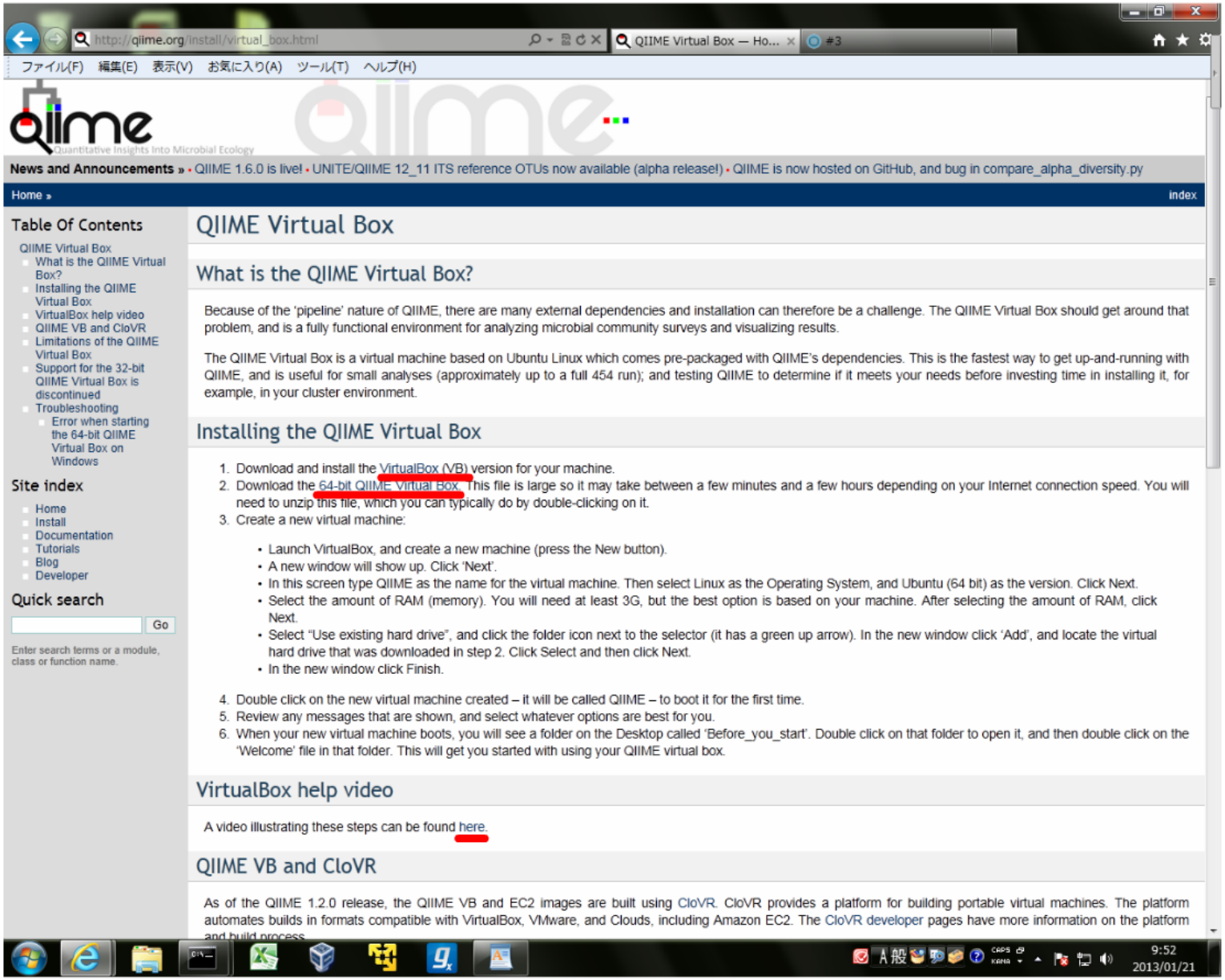
Quantitative Insights Into Microbial Ecology

QIIMEインストール方法

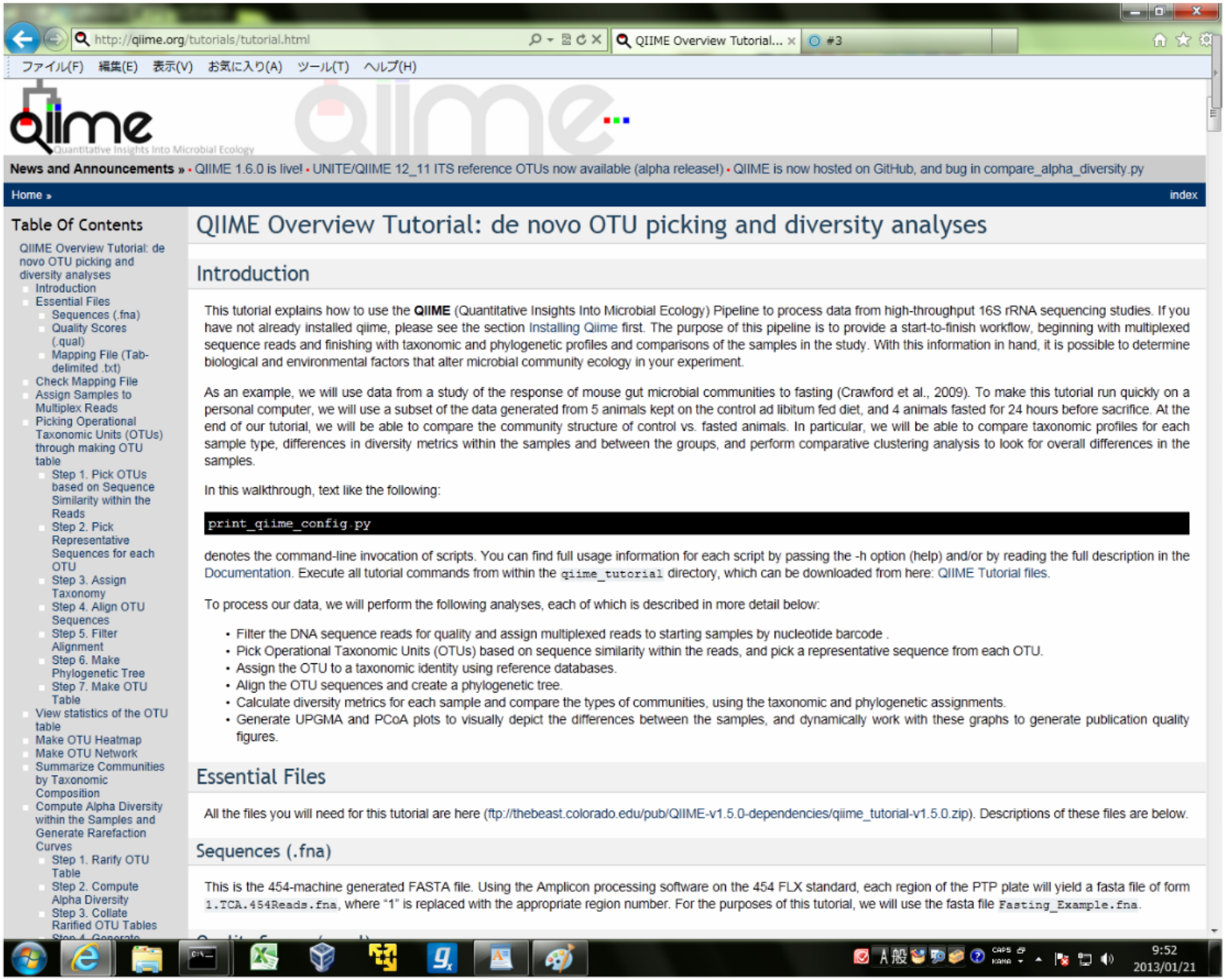



http://qiime.org/install/virtual_box.html
(F)
編集(E)
表示(V)
お気に入り(A)
ツール
me

QIIMEインストール方法



QIIME Tutorial



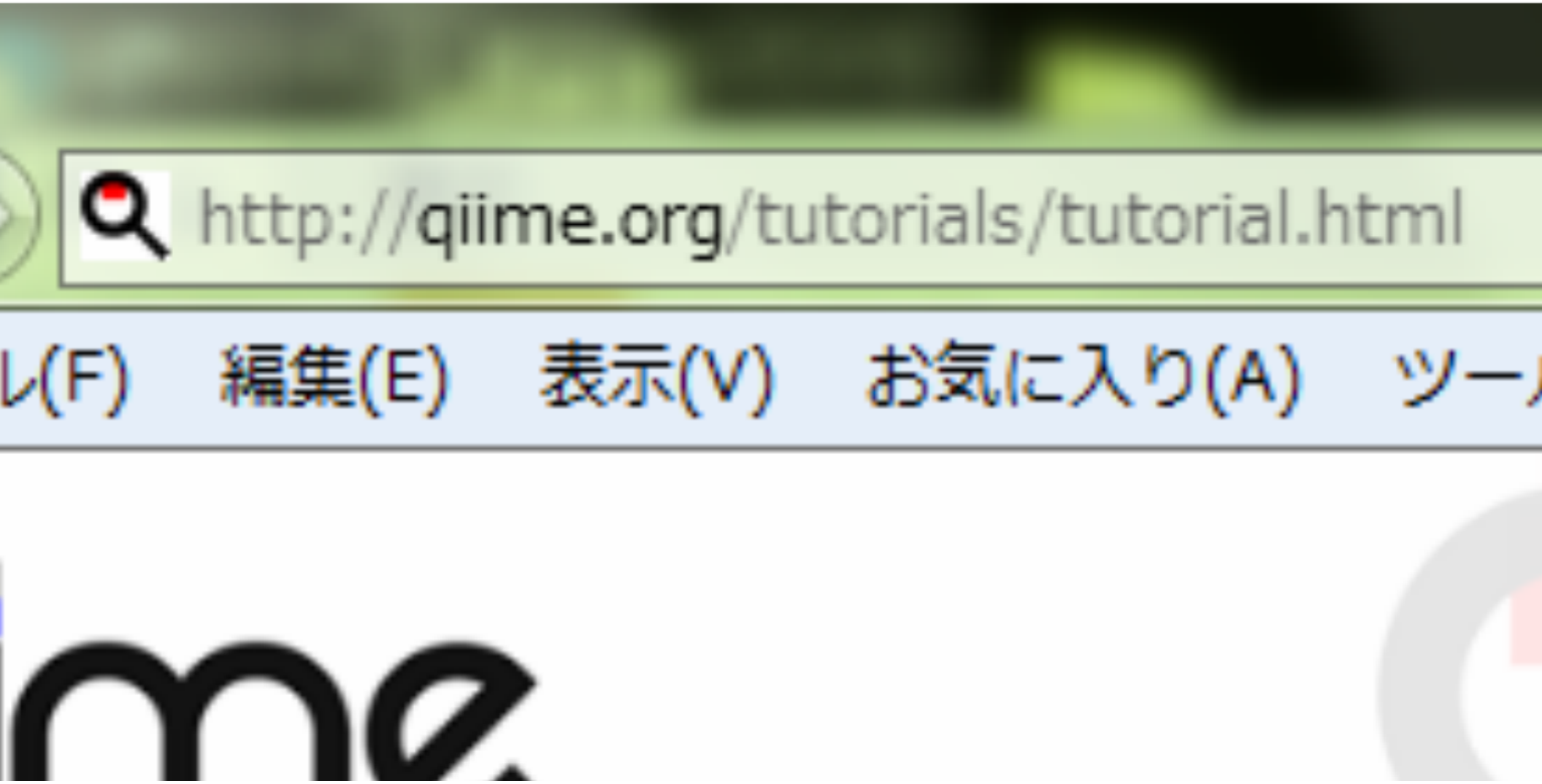

http://qiime.org/tutorials/tutorial.html
ル(F)　編集(E)　表示(V)　お気に入り(A)　ツー

Table Of Contents

- QIIME Overview Tutorial: de novo OTU picking and diversity analyses
  - Introduction
  - Essential Files
    - Sequences (.fna)
    - Quality Scores (.qual)
    - Mapping File (Tab-delimited .txt)
  - Check Mapping File
  - Assign Samples to Multiplex Reads
  - Picking Operational Taxonomic Units (OTUs) through making OTU table
    - Step 1. Pick OTUs based on Sequence Similarity within the Reads
    - Step 2. Pick Representative Sequences for each OTU
    - Step 3. Assign Taxonomy
    - Step 4. Align OTU Sequences
    - Step 5. Filter Alignment
    - Step 6. Make Phylogenetic Tree
    - Step 7. Make OTU Table
  - View statistics of the OTU table
  - Make OTU Heatmap
  - Make OTU Network
  - Summarize Communities by Taxonomic Composition
  - Compute Alpha Diversity within the Samples and Generate Rarefaction Curves
    - Step 1. Rarify OTU Table
    - Step 2. Compute Alpha Diversity
    - Step 3. Collate Rarified OTU Tables

QIIME Overview Tutorial: de novo OTU picking and diversity analyses

Introduction

This tutorial explains how to use the **QIIME** (Quantitative Insights Into Microbial Ecology) Pipeline to process data from high-throughput 16S rRNA sequencing studies. If you have not already installed qiime, please see the section Installing Qiime first. The purpose of this pipeline is to provide a start-to-finish workflow, beginning with multiplexed sequence reads and finishing with taxonomic and phylogenetic profiles and comparisons of the samples in the study. With this information in hand, it is possible to determine biological and environmental factors that alter microbial community ecology in your experiment.

As an example, we will use data from a study of the response of mouse gut microbial communities to fasting (Crawford et al., 2009). To make this tutorial run quickly on a personal computer, we will use a subset of the data generated from 5 animals kept on the control ad libitum fed diet, and 4 animals fasted for 24 hours before sacrifice. At the end of our tutorial, we will be able to compare the community structure of control vs. fasted animals. In particular, we will be able to compare taxonomic profiles for each sample type, differences in diversity metrics within the samples and between the groups, and perform comparative clustering analysis to look for overall differences in the samples.

In this walkthrough, text like the following:

```
print_qiime_config.py
```

denotes the command-line invocation of scripts. You can find full usage information for each script by passing the -h option (help) and/or by reading the full description in the Documentation. Execute all tutorial commands from within the `qiime_tutorial` directory, which can be downloaded from here: QIIME Tutorial files.

To process our data, we will perform the following analyses, each of which is described in more detail below:

- Filter the DNA sequence reads for quality and assign multiplexed reads to starting samples by nucleotide barcode .
- Pick Operational Taxonomic Units (OTUs) based on sequence similarity within the reads, and pick a representative sequence from each OTU.
- Assign the OTU to a taxonomic identity using reference databases.
- Align the OTU sequences and create a phylogenetic tree.
- Calculate diversity metrics for each sample and compare the types of communities, using the taxonomic and phylogenetic assignments.
- Generate UPGMA and PCoA plots to visually depict the differences between the samples, and dynamically work with these graphs to generate publication quality figures.

Essential Files

All the files you will need for this tutorial are here (ftp://thebeast.colorado.edu/pub/QIIME-v1.5.0-dependencies/qiime_tutorial-v1.5.0.zip). Descriptions of these files are below.

Sequences (.fna)

This is the 454-machine generated FASTA file. Using the Amplicon processing software on the 454 FLX standard, each region of the PTP plate will yield a fasta file of form `1.TCA.454Reads.fna`, where "1" is replaced with the appropriate region number. For the purposes of this tutorial, we will use the fasta file `Fasting_Example.fna`.

samples.

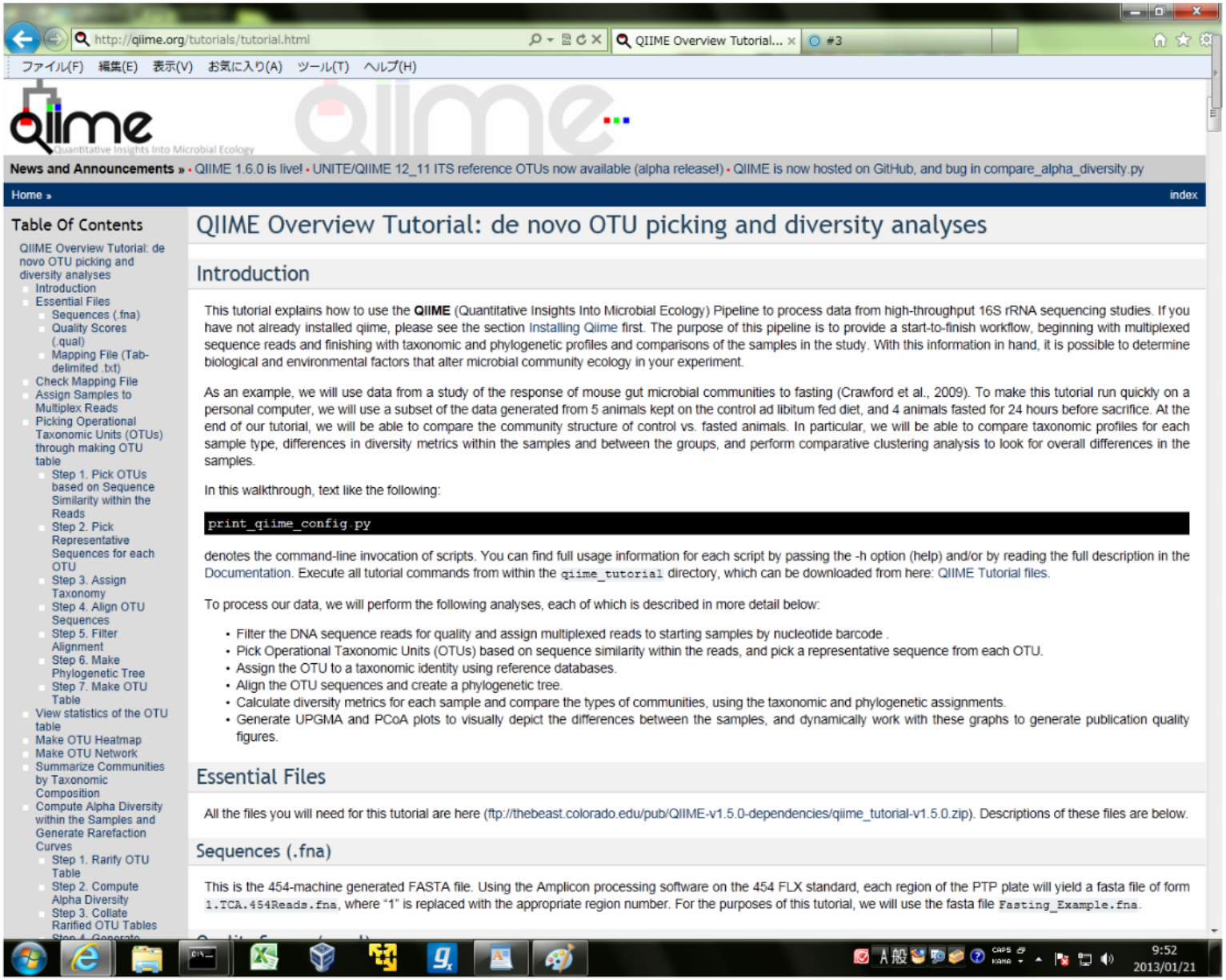
In this walkthrough, text like the

```
print_qiime_config.py
```

denotes the command-line invoc

Documentation. Execute all tuto

QIIME Tutorial



QIIME解析の流れ

1．準備するファイル

・Sequences (.fna)

・[redacted]

・Mapping File (Tab-delimited .txt)

2．Check Mapping File

```
check_id_map.py -m Mapファイル名 -o mapping_output
```

[redacted]

4．Picking Operational Taxonomic Units (OTUs) through making OTU table

```
pick_otus_through_otu_table.py -i split_library_output/seqs.fna -o otus
```

5．View statistics of the OTU table

```
per_library_stats.py -i otus/otu_table.biom
```

6．Summarize Communities by Taxonomic Composition

```
summarize_taxa_through_plots.py -i otus/otu_table.biom -o wf_taxa_summary -m Mapファイル名
```

otu_table_L2.txt⇒Phylum
otu_table_L3.txt⇒Class
otu_table_L4.txt⇒Order
otu_table_L5.txt⇒Family
otu_table_L6.txt⇒Genus

pick_otus_through_otu_table

Picking Operational Taxonomic Units (OTUs) through making OTU table

Here we will be running the pick_otus_through_otu_table.py workflow, which performs a series of small steps by calling a series of other scripts automatically. This workflow consists of the following steps:

1. Picking OTUs (for more information, refer to pick_otus.py)
2. Picking a representative sequence set, one sequence from each OTU (for more information, refer to pick_rep_set.py)
3. Aligning the representative sequence set (for more information, refer to align_seqs.py)
4. Assigning taxonomy to the representative sequence set (for more information, refer to assign_taxonomy.py)
5. Filtering the alignment prior to tree building - removing positions which are all gaps, or not useful for phylogenetic inference (for more information, refer to filter_alignment.py)
6. Building a phylogenetic tree (for more information, refer to make_phylogeny.py)
7. Building an OTU table (for more information, refer to make_otu_table.py)

Using the output from split_libraries.py (the seqs.fna file), run the following command:

```
pick_otus_through_otu_table.py -i split_library_output/seqs.fna -o otus
```

QIIME解析の流れ

1．準備するファイル

・Sequences (.fna)

・[redacted]

・Mapping File (Tab-delimited .txt)

2．Check Mapping File

```
check_id_map.py -m Mapファイル名 -o mapping_output
```

[redacted]

4．Picking Operational Taxonomic Units (OTUs) through making OTU table

```
pick_otus_through_otu_table.py -i split_library_output/seqs.fna -o otus
```

5．View statistics of the OTU table

```
per_library_stats.py -i otus/otu_table.biom
```

6．Summarize Communities by Taxonomic Composition

```
summarize_taxa_through_plots.py -i otus/otu_table.biom -o wf_taxa_summary -m Mapファイル名
```

otu_table_L2.txt⇒Phylum
otu_table_L3.txt⇒Class
otu_table_L4.txt⇒Order
otu_table_L5.txt⇒Family
otu_table_L6.txt⇒Genus

per_library_stats.py -

```
Num samples: 9

Seqs/sample summary:
   Min: 146
   Max: 150
   Median: 148.0
   Mean: 148.111111111
   Std. dev.: 1.4487116456
   Median Absolute Deviation: 1.0
   Default even sampling depth in
      core_qiime_analyses.py (just a suggestion): 146

Seqs/sample detail:
   PC.355: 146
   PC.481: 146
   PC.636: 147
   PC.354: 148
   PC.635: 148
   PC.593: 149
   PC.607: 149
   PC.356: 150
   PC.634: 150
```

Summarize C

QIIME解析の流れ

1．準備するファイル

・Sequences (.fna)

・[redacted]

・Mapping File (Tab-delimited .txt)

2．Check Mapping File

```
check_id_map.py -m Mapファイル名 -o mapping_output
```

[redacted]

4．Picking Operational Taxonomic Units (OTUs) through making OTU table

```
pick_otus_through_otu_table.py -i split_library_output/seqs.fna -o otus
```

5．View statistics of the OTU table

```
per_library_stats.py -i otus/otu_table.biom
```

6．Summarize Communities by Taxonomic Composition

```
summarize_taxa_through_plots.py -i otus/otu_table.biom -o wf_taxa_summary -m Mapファイル名
```

otu_table_L2.txt⇒Phylum
otu_table_L3.txt⇒Class
otu_table_L4.txt⇒Order
otu_table_L5.txt⇒Family
otu_table_L6.txt⇒Genus

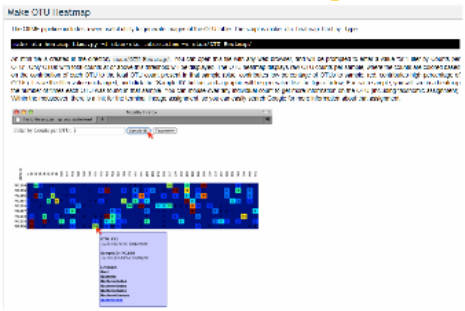
e

onomic Composition

-o wf_taxa_summary -m Mapファイル名

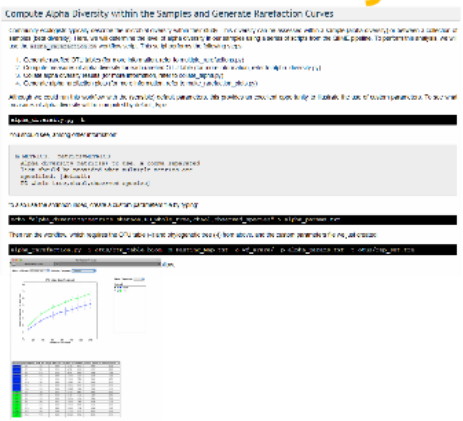
otu_table_L2.txt⇒Phylum
otu_table_L3.txt⇒Class
otu_table_L4.txt⇒Order
otu_table_L5.txt⇒Family
otu_table_L6.txt⇒Genus

Heatmap

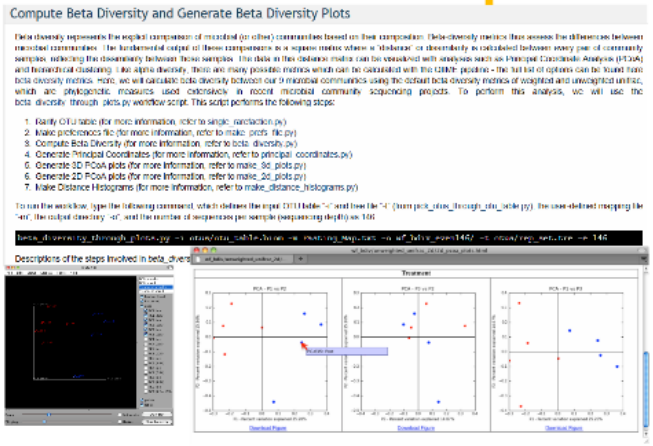


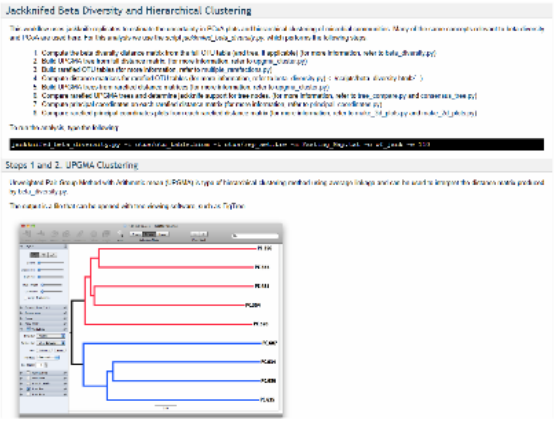
Network

α-diversity



β-diversity



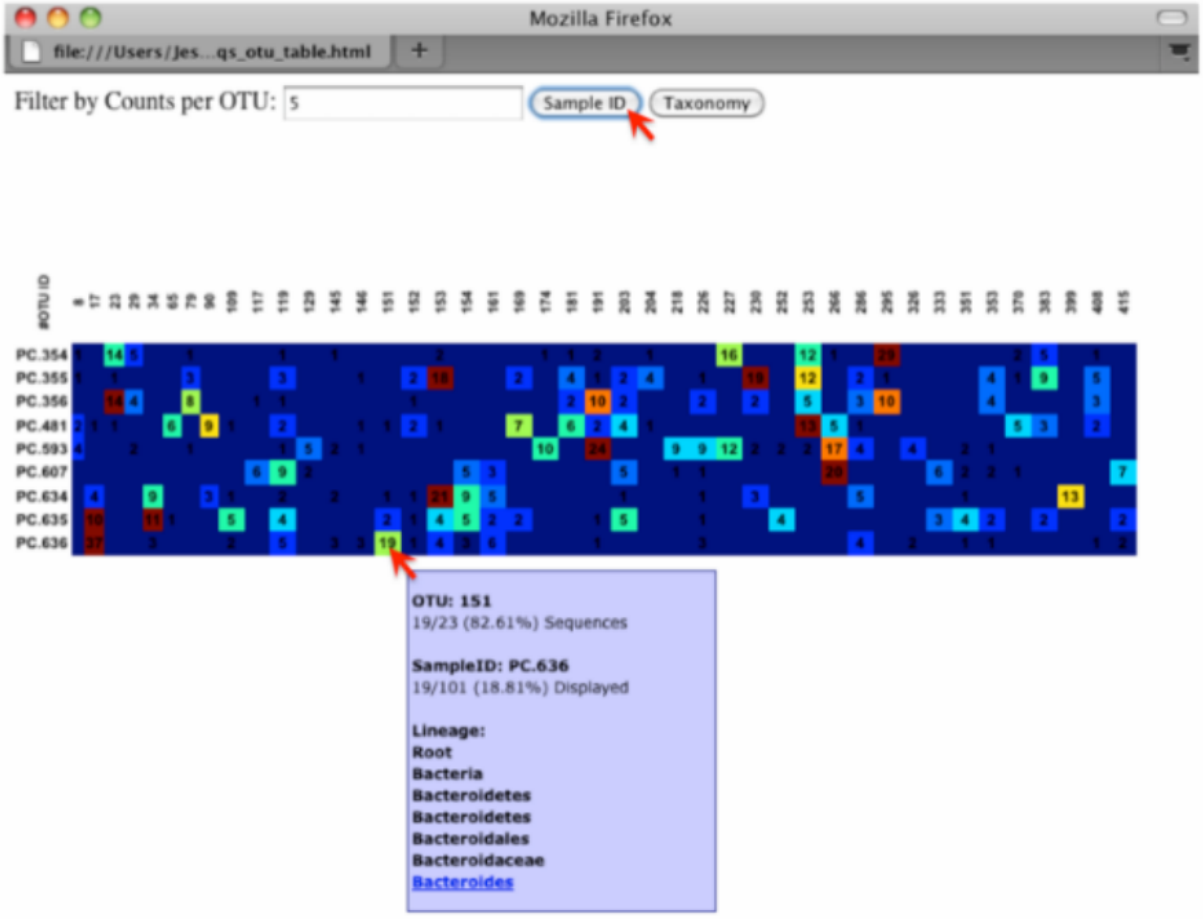


Make OTU Heatmap

The QIIME pipeline includes a very useful utility to generate images of the OTU table. The script is make_otu_heatmap_html.py. Type:

```
make_otu_heatmap_html.py -i otus/otu_table.biom -o otus/OTU_Heatmap/
```

An html file is created in the directory `otus/OTU_Heatmap/`. You can open this file with any web browser, and will be prompted to enter a value for "Filter by Counts per OTU". Only OTUs with total counts at or above this threshold will be displayed. The OTU heatmap displays raw OTU counts per sample, where the counts are colored based on the contribution of each OTU to the total OTU count present in that sample (blue: contributes low percentage of OTUs to sample; red: contributes high percentage of OTUs). Leave the filter value unchanged, and click the "Sample ID" button, and a graphic will be generated like the figure below. For each sample, you will see in a heatmap the number of times each OTU was found in that sample. You can mouse over any individual count to get more information on the OTU (including taxonomic assignment). Within the mouseover, there is a link for the terminal lineage assignment, so you can easily search Google for more information about that assignment.



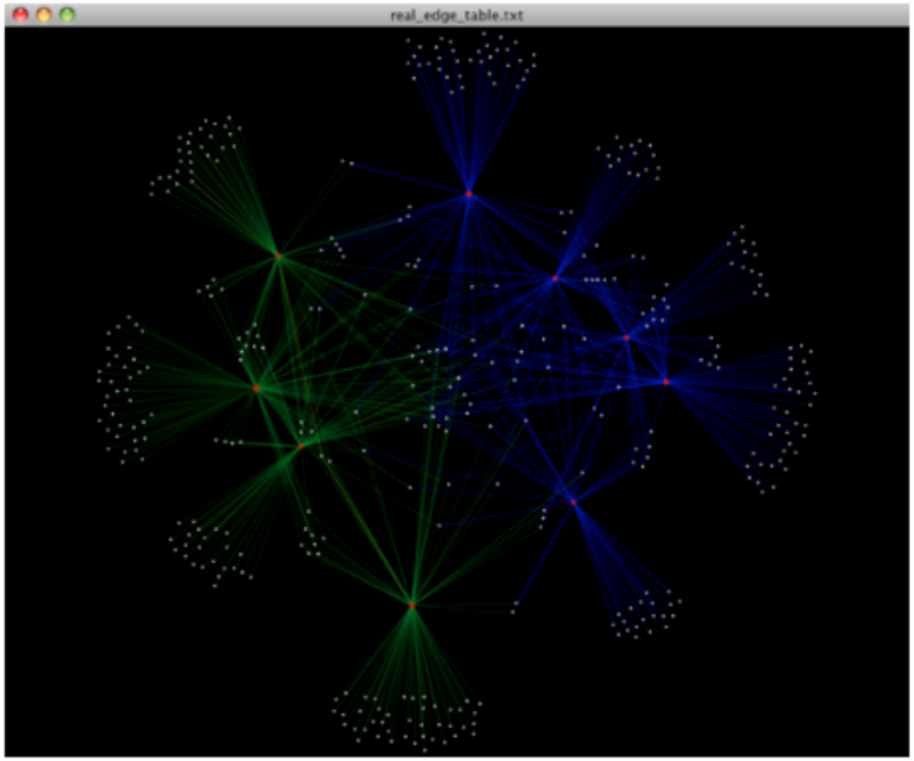
Network

Make OTU Network

An alternative to viewing the OTU table as a heatmap is to create an OTU network, using the following command.:

```
make_otu_network.py -m Fasting_Map.txt -i otus/otu_table.biom -o otus/OTU_Network
```

To visualize the network, we use the Cytoscape program (which you can run by calling cytoscape from the command line – you may need to call this beginning either with a capital or lowercase 'C' depending on your version of Cytoscape), where each red circle represents a sample and each white square represents an OTU. The lines represent the OTUs present in a particular sample (blue for controls and green for fasting). For more information about opening the files in Cytoscape please refer to Making Cytoscape Networks.



Compute Alpha Diversity within the Samples and Generate Rarefaction Curves

Community ecologists typically describe the microbial diversity within their study. This diversity can be assessed within a sample (alpha diversity) or between a collection of samples (beta diversity). Here, we will determine the level of alpha diversity in our samples using a series of scripts from the QIIME pipeline. To perform this analysis, we will use the `alpha_rarefaction.py` workflow script. This script performs the following steps:

1. Generate rarefied OTU tables (for more information, refer to multiple_rarefactions.py)
2. Compute measures of alpha diversity for each rarefied OTU table (for more information, refer to alpha_diversity.py)
3. Collate alpha diversity results (for more information, refer to collate_alpha.py)
4. Generate alpha rarefaction plots (for more information, refer to make_rarefaction_plots.py)

Although we could run this workflow with the (sensible) default parameters, this provides an excellent opportunity to illustrate the use of custom parameters. To see what measures of alpha diversity will be computed by default, type:

```
alpha_diversity.py -h
```

You should see, among other information:

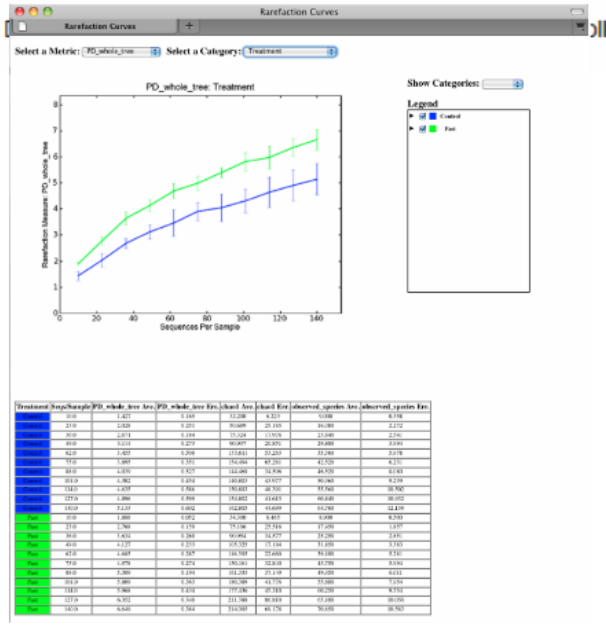
```
-m METRICS, -metrics=METRICS
   Alpha-diversity metric(s) to use. A comma-separated
   list should be provided when multiple metrics are
   specified. [default:
   PD_whole_tree,chao1,observed_species]
```

to also use the shannon index, create a custom parameters file by typing:

```
echo "alpha_diversity:metrics shannon,PD_whole_tree,chao1,observed_species" > alpha_params.txt
```

Then run the workflow, which requires the OTU table (-i) and phylogenetic tree (-t) from above, and the custom parameters file we just created:

```
alpha_rarefaction.py -i otus/otu_table.biom -m Fasting_Map.txt -o wf_arare/ -p alpha_params.txt -t otus/rep_set.tre
```



Compute Beta Diversity and Generate Beta Diversity Plots

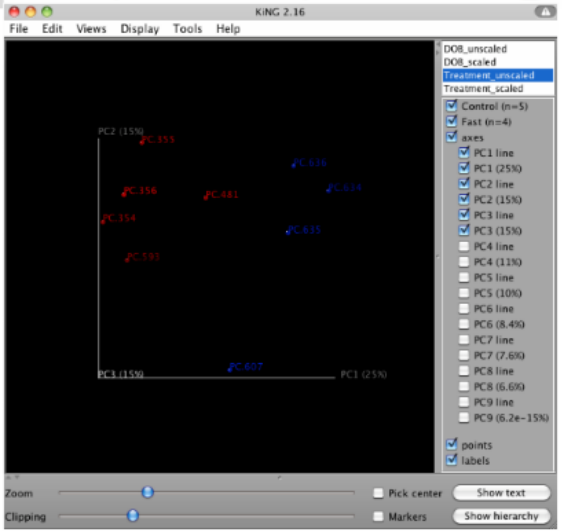
Beta diversity represents the explicit comparison of microbial (or other) communities based on their composition. Beta-diversity metrics thus assess the differences between microbial communities. The fundamental output of these comparisons is a square matrix where a "distance" or dissimilarity is calculated between every pair of community samples, reflecting the dissimilarity between those samples. The data in this distance matrix can be visualized with analyses such as Principal Coordinate Analysis (PCoA) and hierarchical clustering. Like alpha diversity, there are many possible metrics which can be calculated with the QIIME pipeline - the full list of options can be found here beta diversity metrics. Here, we will calculate beta diversity between our 9 microbial communities using the default beta diversity metrics of weighted and unweighted unifrac, which are phylogenetic measures used extensively in recent microbial community sequencing projects. To perform this analysis, we will use the beta_diversity_through_plots.py workflow script. This script performs the following steps:

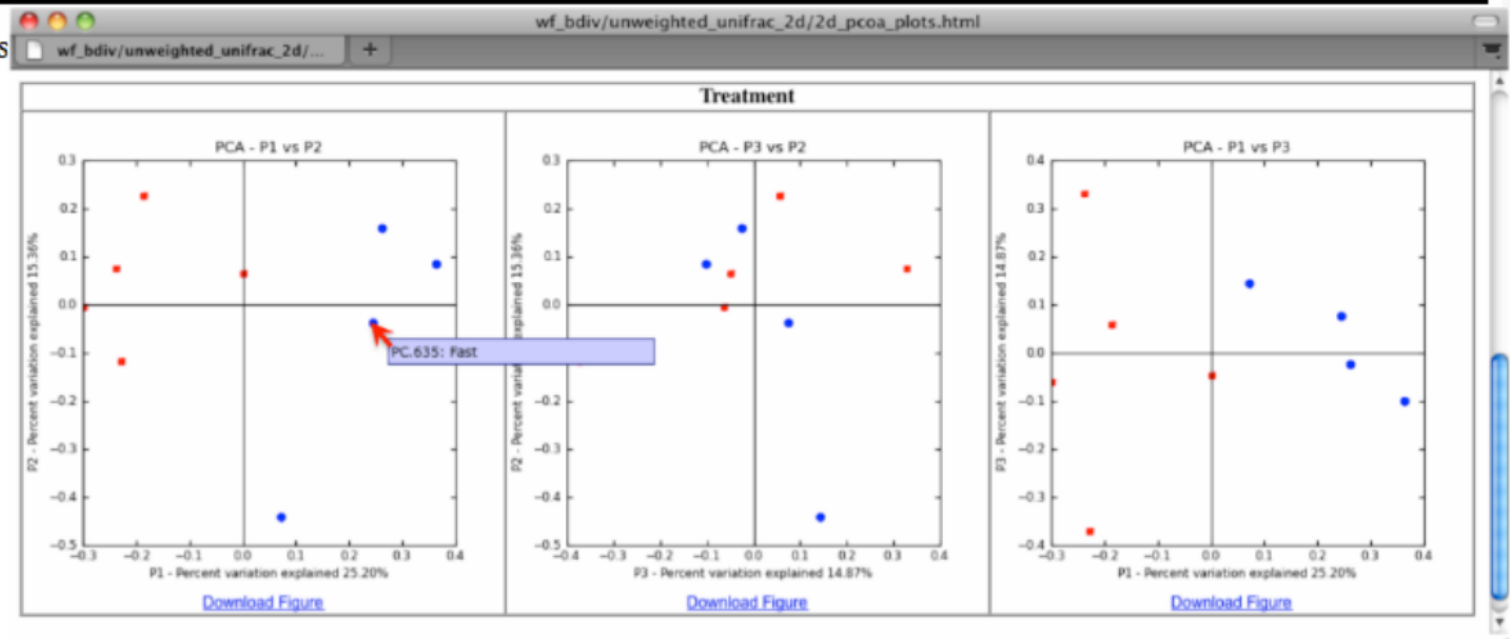
1. Rarify OTU table (for more information, refer to single_rarefaction.py)
2. Make preferences file (for more information, refer to make_prefs_file.py)
3. Compute Beta Diversity (for more information, refer to beta_diversity.py)
4. Generate Principal Coordinates (for more information, refer to principal_coordinates.py)
5. Generate 3D PCoA plots (for more information, refer to make_3d_plots.py)
6. Generate 2D PCoA plots (for more information, refer to make_2d_plots.py)
7. Make Distance Histograms (for more information, refer to make_distance_histograms.py)

To run the workflow, type the following command, which defines the input OTU table "-i" and tree file "-t" (from pick_otus_through_otu_table.py), the user-defined mapping file "-m", the output directory "-o", and the number of sequences per sample (sequencing depth) as 146:

```
beta_diversity_through_plots.py -i otus/otu_table.biom -m Fasting_Map.txt -o wf_bdiv_even146/ -t otus/rep_set.tre -e 146
```

Descriptions of the steps involved in *beta_divers*





Jackknifed Beta Diversity and Hierarchical Clustering

This workflow uses jackknife replicates to estimate the uncertainty in PCoA plots and hierarchical clustering of microbial communities. Many of the same concepts relevant to beta diversity and PCoA are used here. For this analysis we use the script *jackknifed_beta_diversity.py*, which performs the following steps:

1. Compute the beta diversity distance matrix from the full OTU table (and tree, if applicable) (for more information, refer to beta_diversity.py)
2. Build UPGMA tree from full distance matrix; (for more information, refer to upgma_cluster.py)
3. Build rarefied OTU tables (for more information, refer to multiple_rarefactions.py)
4. Compute distance matrices for rarefied OTU tables (for more information, refer to beta_diversity.py) <../scripts/beta_diversity.html>`_)
5. Build UPGMA trees from rarefied distance matrices (for more information, refer to upgma_cluster.py)
6. Compare rarefied UPGMA trees and determine jackknife support for tree nodes. (for more information, refer to tree_compare.py and consensus_tree.py)
7. Compute principal coordinates on each rarefied distance matrix (for more information, refer to principal_coordinates.py)
8. Compare rarefied principal coordinates plots from each rarefied distance matrix (for more information, refer to make_3d_plots.py and make_2d_plots.py)

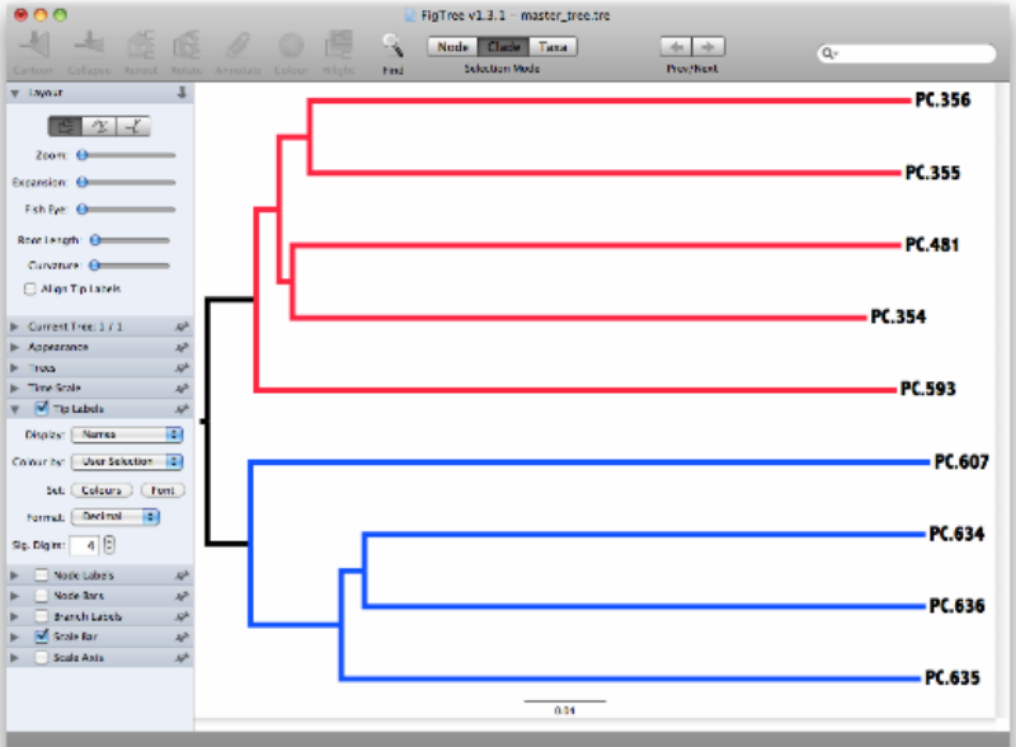
To run the analysis, type the following:

```
jackknifed_beta_diversity.py -i otus/otu_table.biom -t otus/rep_set.tre -m Fasting_Map.txt -o wf_jack -e 110
```

Steps 1 and 2. UPGMA Clustering

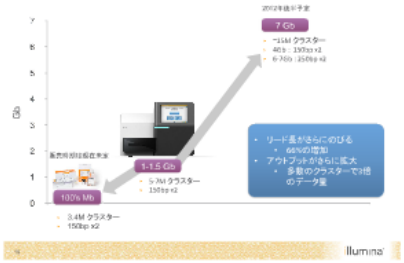
Unweighted Pair Group Method with Arithmetic mean (UPGMA) is type of hierarchical clustering method using average linkage and can be used to interpret the distance matrix produced by beta_diversity.py.

The output is a file that can be opened with tree viewing software, such as FigTree.



Quantitative Insights Into Microbial Ecology

MiSeqのパフォーマンスの向上



Classification accuracy by query size (Naïve Bayesian Classifier)

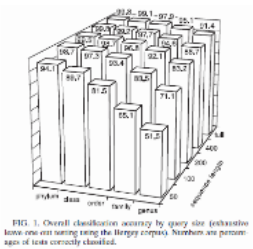
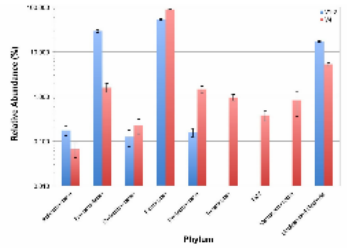


FIG. 1. Overall classification accuracy by query size (exhaustive leave-one-out testing using the Bergey corpus). Numbers are percentages of tests correctly classified.

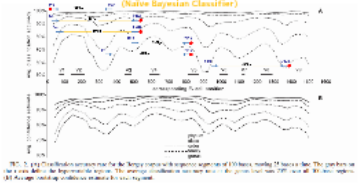
Wang, Q., et al., Appl. Environ. Microbiol. 2007, 73(16):5261

Difference of amplicon between V1-2 region and V4 region



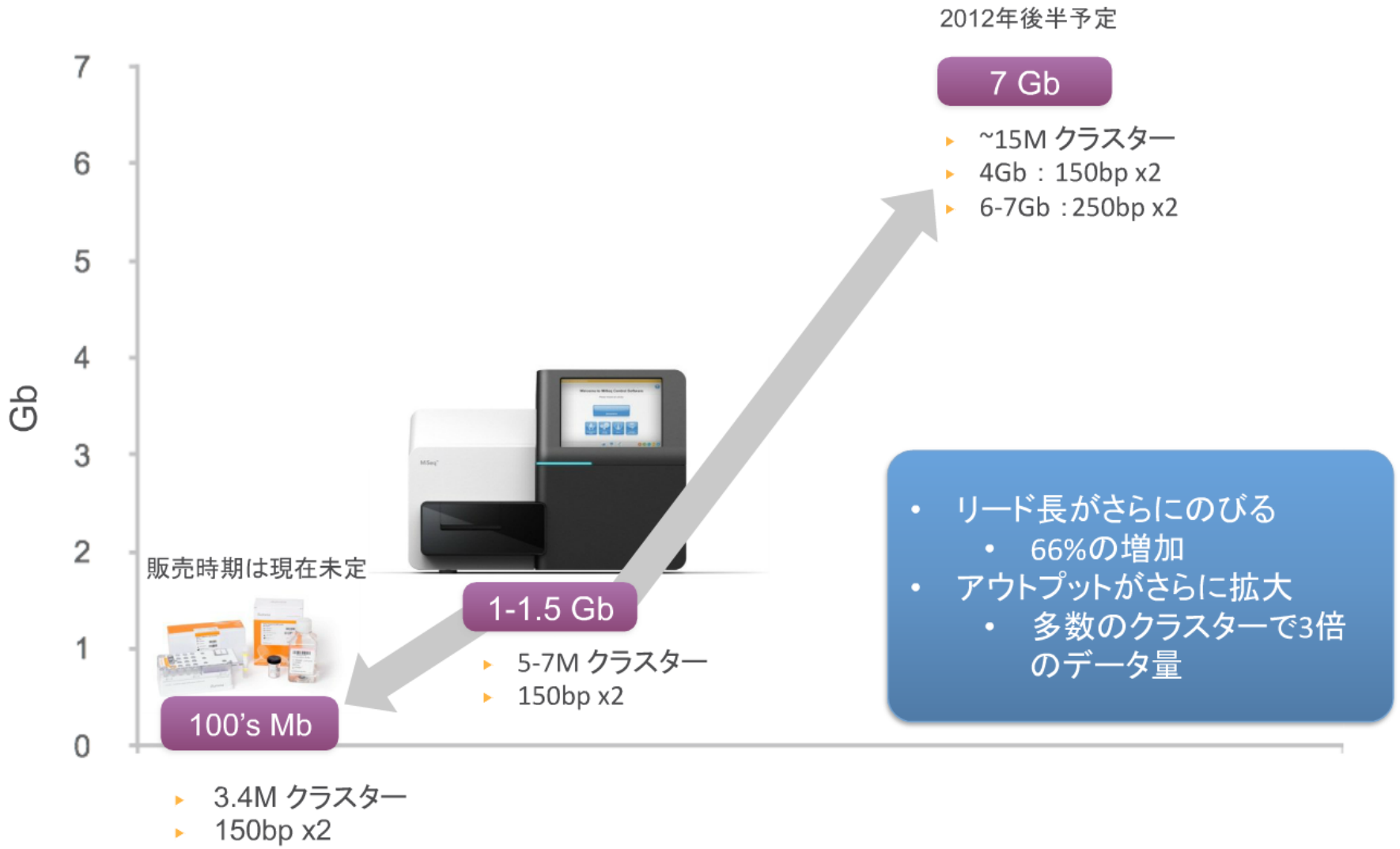
ISME J. 2012 Nov;6(11):2033-44.

Classification accuracy rate for Bergey corpus with sequence segments (Naïve Bayesian Classifier)



Wang, Q., et al., Appl. Environ. Microbiol. 2007, 73(16):5261

MiSeqのパフォーマンスの向上



illumina

Classification accuracy by query size (Naïve Bayesian Classifier)

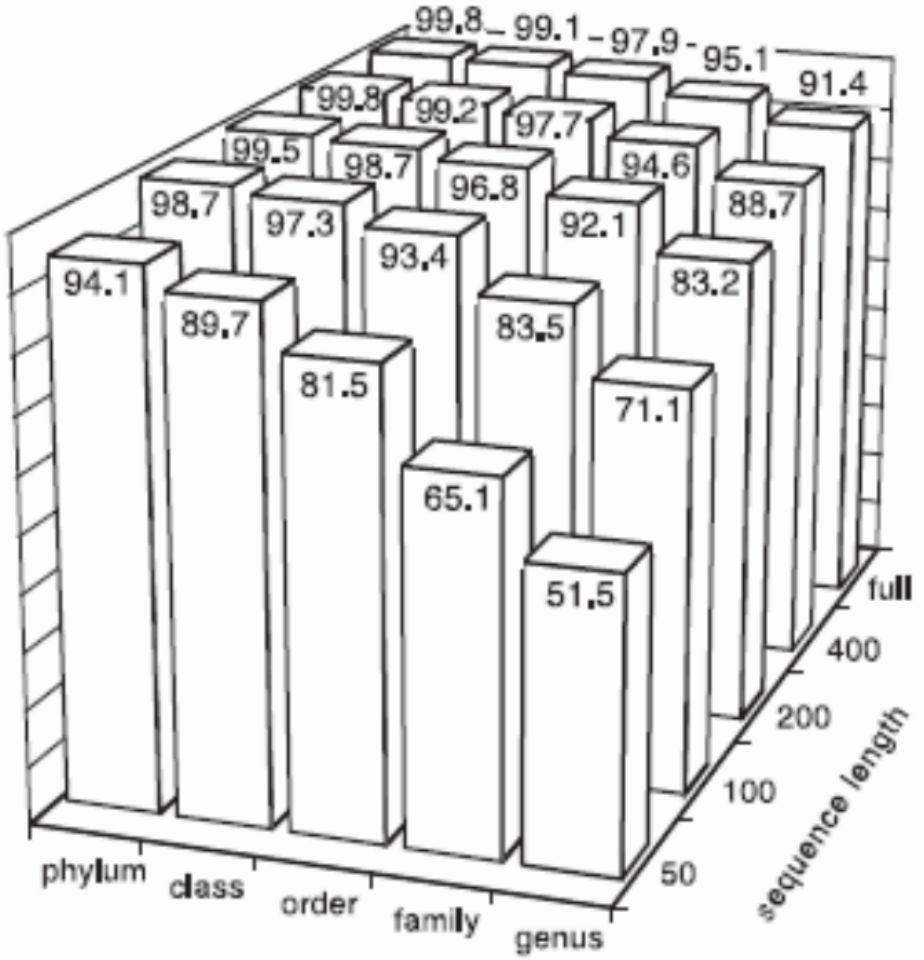
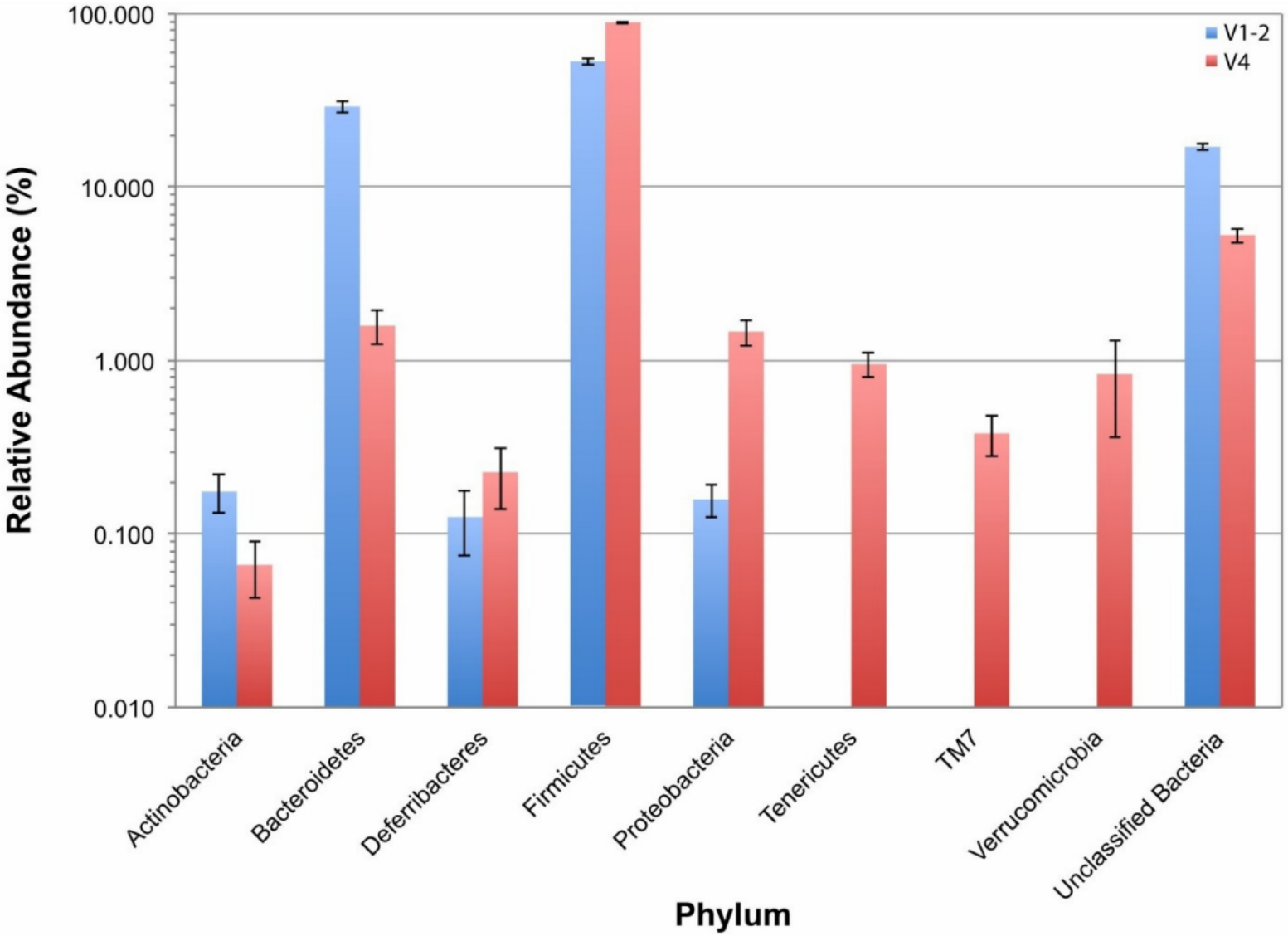


FIG. 1. Overall classification accuracy by query size (exhaustive leave-one-out testing using the Bergey corpus). Numbers are percentages of tests correctly classified.

Wang, Q., et al., Appl. Environ. Microbiol. 2007, 73(16):5261

Difference of amplicon between V1-2 region and V4 region



ISME J. 2012 Nov;6(11):2033-44.

Classification accuracy rate for Bergey corpus with sequence segments (Naïve Bayesian Classifier)

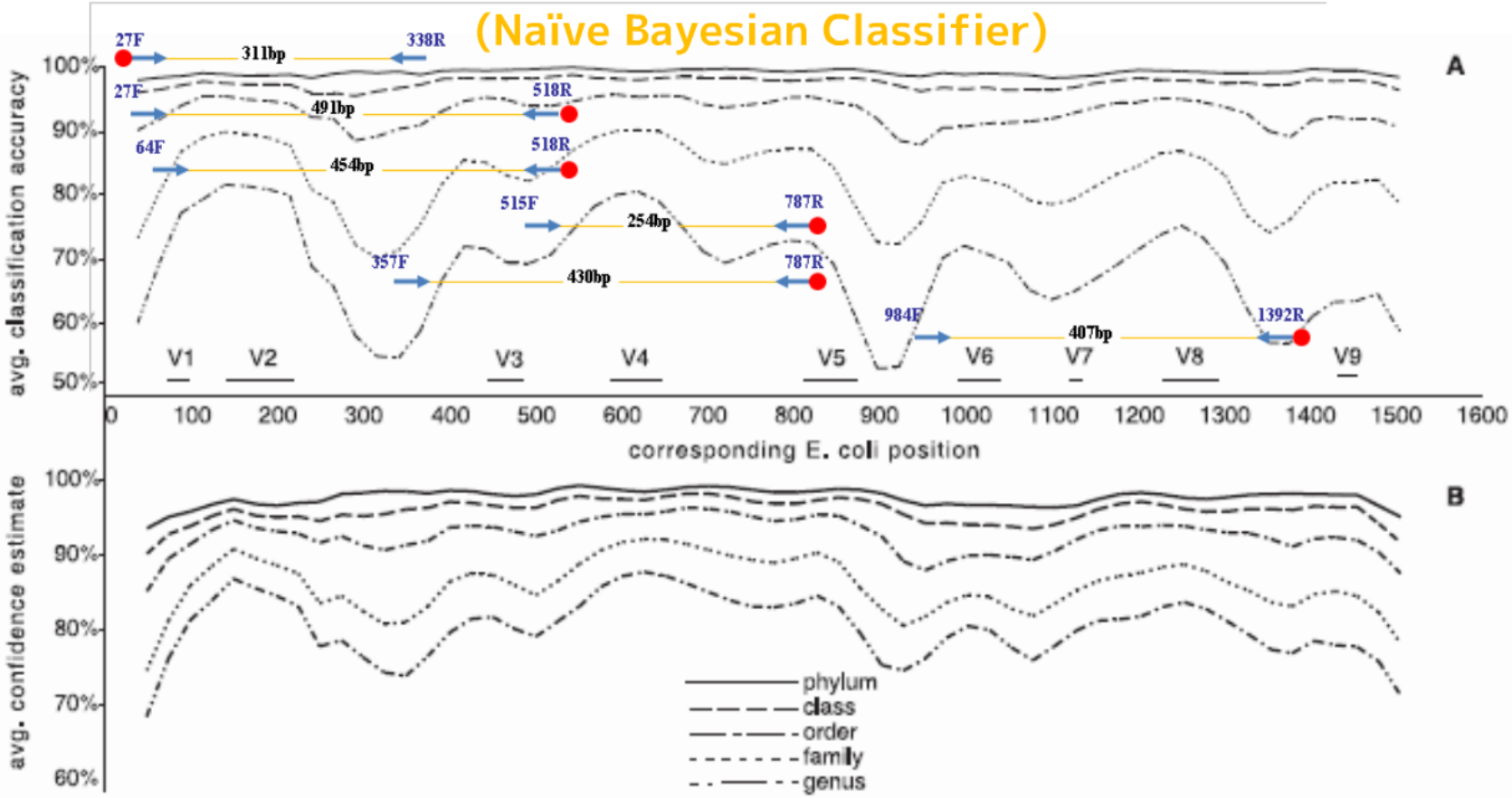


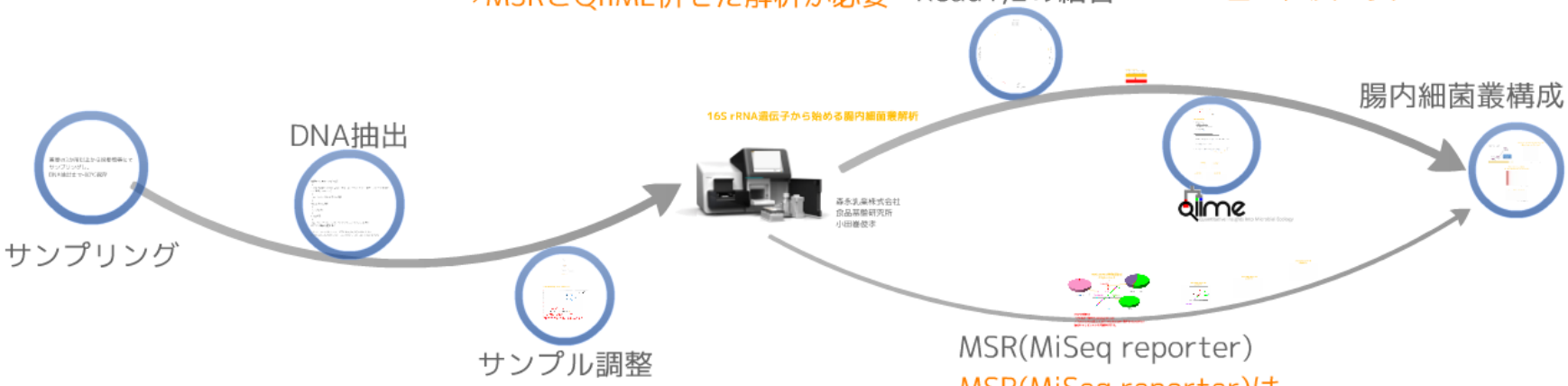
FIG. 2. (A) Classification accuracy rate for the Bergey corpus with sequence segments of 100 bases, moving 25 bases a time. The gray bars on the *x* axis define the hypervariable regions. The average classification accuracy rate at the genus level was 70% over all 100-base regions. (B) Average bootstrap confidence estimate for each segment.

Wang, Q., et al., Appl. Environ. Microbiol. 2007, 73(16):5261

Workflow

Read1,2の結合は
・PhiXやクオリティスコアの低い配列を除去
・リードの同定率向上(Unclassifiedが減少)
する上で有効だが、
・設定が厳しいとActinobacteriaを低く見積もる可能性がある。
⇒MSRとQIIME併せた解析が必要

カスタムプライマー
・V4領域はBacteroidetesを低く見積もってしまう
⇒250bp×2に合った増幅部位の検証
・12塩基のIndex配列は>Q30リードが激減
⇒６塩基程度が良い？



サンプルの濃度測定は最重要

サンプルライブラリーは推奨値よりも高め
・PCR product 16pM, PhiX 6pM
（MiSeq v2はこの1.5倍程度？）

MSR(MiSeq reporter)は
PhiXのリードが数%混入してしまう。