Анализ сложных хроматограмм популяционного секвенирования

Юрий Фантин
Центральный НИИ Эпидемиологии
fantinys@mail.ru

Алексей Неверов
Центральный НИИ Эпидемиологии;
ФББ МГУ
neva_2000@mail.ru

Александр Фаворов
ИОГЕН РАН; ГосНИИГенетика;
Johns Hopkins University School of Medicine
favorov@sensi.org

Андрей А. Миронов
ФББ МГУ
mironov@bioinf.fbb.msu.ru

Владимир Чуланов
Центральный НИИ Эпидемиологии
vladimir.chulanov@pcr.ru

Аннотация

Метод прямого или популяционного секвенирования продукта ПЦР широко применяется в медицинских диагностических и научных целях. Хроматограммы полученные данным методом содержат информацию о смеси типов ДНК, одновременно амплифицированных при ПЦР. Важной задачей является извлечение информации, характеризующей генетическое разнообразие этих вариантов без применения клонирования ПЦР продукта. В настоящей работе мы предлагаем новый метод расшифровки хроматограммы популяционного секвенирования и новый жадный алгоритм расшифровки структуры популяции - BCV. На входе алгоритма — последовательность пиков хроматограммы и словарь (множественное выравнивание последовательностей). Мы предполагаем, что хроматограмма является результатом секвенирования локуса генома, гомологичного последовательностям словаря. Существует три основных варианта использования программы: - Определение нуклеотидной последовательности хроматограммы, - Определение типов ДНК, составляющих секвенируемую популяцию. Предсказанные ДНК-типы далее могут быть охарактеризованы методами поиска по базам данных биологических последовательностей (например, blast) или с помощью филогенетического анализа совместно с последовательностями словаря. - Определение наличия и размера делеций/вставок и их положений относительно референс последовательности. Построение консенсусной последовательности, соответствующей доминирующей в смеси субпопуляции типов ДНК. Мы показали применимость метода для разделения смеси генотипов вируса гепатита B, типирования бактериальных сообществ в клинических образцах человека по 16S РНК и определения делеций/вставок в гене pncA M. tuberculosis.

1. Введение

Для определения нуклеотидной последовательности по хроматограмме секвенирования по методу Сенгера первоначально разрабатывались методы анализа, допускающие единственный тип ДНК: например phred [1, 2]. Такие методы, в общем случае, не могут использоваться для анализа результатов прямого секвенирования. Благодаря активному развитию проектов по ресеквенированию геномов были разработаны приложения для обработки хроматограмм диплоидных геномов, например, TraceTuner [3]. Ряд приложений [4, 5, 6] осуществляют повторный анализ хроматограмм (re-basecalling) после предварительной обработки «сырых» результатов секвенирования программами, определяющими первичную последовательность (например phred).

Целью проектов ресеквенирования геномов является определение полиморфизмов, в первую очередь SNP. Для того, чтобы уменьшить вероятность принять артефакт секвенирования за SNP, рядом алгоритмов используется предварительная сборка по-

следовательностей, полученных в результате многократного прочтения одного локуса у различных индивидов. Этот подход используется для ресеквенирования генома человека и обнаружения новых SNP следующими приложениями: polyphred [7, 8] и polybayes [9], AutoEditor [4]. Эти алгоритмы требуют относительно высокой степени покрытия последовательностями каждого сайта, разработаны для ресеквенирования геномов эукариот с невысокой степенью вариабельности и не позволяют выявлять вставки и делеции.

Задача расшифровки хроматограммы не может быть сведена к определению соответствующей последовательности IUPAC кодов, соответствующих нуклеотидному составу каждой позиции (к задаче base-calling), если в составе образца присутствуют аллели, характеризующиеся делециями или вставками. Большинство приложений при анализе хроматограмм, полученных методом прямого секвенирования, при наличии вставок или делеций в одном из аллелей, неудовлетворительно определяют соответствующую нуклеотидную последовательность, либо рассматривают второй аллель как загрязнение неспецифическим продуктом амплификации. В настоящее время существуют, алгоритмы позволяющие решать поставленную задачу при наличии гетерозиготы — двух типов ДНК в смеси [6, 10, 11].

Хроматограмма, получаемая методом прямого секвенирования, является суперпозицией сигналов от гомологичных фрагментов ДНК, присутствующих в продукте ПЦР. Амплитуды пиков хроматограммы несут информацию о пропорции аллелей в образце. В ряде приложений предпринимаются попытки восстановить частоты типов ДНК, присутствующих в смеси, по хроматограмме прямого секвенирования, при условии, что известен ее состав [12, 13], либо известны последовательности всех возможных типов ДНК которые могут быть в составе смеси [13]. В ряде случаев представляет интерес именно задача определения количественного состава образца. В работе [13] авторы показали применимость метода прямого секвенирования для количественной оценки изменения экологического состава микробных популяций во времени.

Задача определения типов ДНК и количественной характеристики секвенируемой популяции в достаточно общем виде (без ограничения на количество типов) была сформулирована в работе [14]. Авторы рассмотрели следующую постановку задачи: задана последовательность дикого типа, словарь разрешенных мутаций (типы мутаций: вставка, удаление и замена подстроки) и экспериментальный профиль - строка, где для каждой позиции последовательности определено множество нуклеотидов. Необходимо найти минимальный набор мутаций, объясняющий профиль. Задача в такой постановке является частным случаем задачи покрытия множества.

Постановка задачи в виде, предложенном в [14], не является универсальной. Для быстро эволюционирующих организмов, таких как многие РНК-вирусы, невозможно заранее определить последовательность, которая могла бы быть выбрана как последовательность дикого типа. По той же причине, во многих случаях невозможно составить словарь разрешенных мутаций. Используя словарь известных последовательностей по секвенируемому локусу генома, мы можем рассчитывать только на то, что в образце присутствует гомолог какой либо последовательности словаря. Быстро эволюционирующим вирусным геномам, таким как HIV, HBV, HCV, свойственно формировать в организме хозяина популяцию квазивидов — высоко гомологичных штаммов. Алгоритм Indelligent [10] позволяет анализировать хроматограммы смесей типов ДНК с делециями, для которого не требуется знать последовательность дикого типа. Indelligent определяет две максимально близкие последовательности которые могут содержать делеции/вставки по отношению друг к другу смесь которых соответствует хроматограмме.

- Определение нуклеотидной последовательности хроматограммы.
- Определение типов ДНК, составляющих секвенируемую популяцию. Предсказанные ДНК-типы далее могут быть охарактеризованы методами поиска по базам данных биологических последовательностей (например, blast) или с помощью филогенетического анализа совместно с последовательностями словаря.
- Определение наличия и размера делеций/вставок и их положений относительно референс последовательности. Построение консенсус-последовательности, соответствующей доминирующей в смеси субпопуляции типов ДНК.

Мы показали применимость метода BCV для разделения смеси генотипов вируса гепатита B, типирования бактериальных сообществ в клинических образцах человека по 16S РНК и определения делеций/вставок в гене *pncA M. tuberculosis*.

2. Материалы и методы

2.1. Образцы

Сравнение качества предсказания нуклеотидных последовательностей хроматограмм проводилось на аннотированных вручную множествах хроматограмм, полученных для для фрагментов геномов вирусов гепатита A (HAV, 2C, 39 образцов) и

D (HDV, 3'-HTP+DeltaAg, 79 образцов). Хроматограммы тестового множества HAV не содержали вырожденных позиций, длина последовательности хроматограммы в среднем составила 680 нк. Хроматограммы тестового множества HDV содержали 3-14% вырожденных позиций, средняя длина хроматограммы — 410 нк. Для определения главной консенсус последовательности, наличия типов ДНК содержащих делеции и вставки были использованы хроматограммы полученные для образцов: вируса иммунодефицита человека тип 1 (HIV, *gag* (*p1,p6*) и *PR* ген), образца вируса гепатита B (HBV, RT домен гена *P*) и 123 образцов *M. tuberculosis* (ген *pncA*). Образцы были секвенированы в двух встречных направлениях: HIV *M. tuberculosis* с ПЦР праймеров, образец HBV с трех праймеров. Из двух образцов биопсии слизистой желудка ID= 95 и 97 была выделена ДНК. Фрагмент гена, кодирующего 16S рРНК длиной ок. 840 нк., был амплифицирован с помощью ПЦР. Продукт ПЦР был секвенирован с трех праймеров, кроме того, образцы были охарактеризованы клонированием и последующим секвенированием отобранных клонов (17 клонов для ID=95 и 10 для ID=97). Для обоих образцов было проведено выявление *Helicobacter pylori* двумя методами: ПЦР ("Амплисенс-Helicobacter pylori-FL", ЦНИИЭ, Россия) и быстрый уреазный тест ("ХЕЛПИЛ", Ассоциация Медицины и Аналитики, Россия).

2.2. Алгоритм

Мы рассматриваем обработку исходной хроматограммы, как процесс, состоящий из следующих стадий:

- выделение пиков (event detection),
- определение нуклеотидной последовательности хроматограммы (base-calling),
- определение типов ДНК, смесь которых соответствует последовательности хроматограммы
- сборка консенсус-последовательности, соответствующей исследуемому фрагменту генома, определение наличия в смеси типов ДНК, содержащих делеции/вставки по отношению к этой последовательности.

Выделение пиков хроматограммы соответствующих четырем типам нуклеотидов в последовательности хроматограммы производится на основании записи флуоресцентного сигнала секвенатора. В настоящей работе мы используем для первичной обработки хроматограммы TraceTuner [3] для определения первичной последовательности нуклеотидов и Polyscan для повторного анализа хроматограммы [6]. Мы не используем группировку пиков по позициям (разбиение), полученное с помощью Polyscan. Позиции могут содержать до четырех пиков, соответствующих различным типам нуклеотидов, таким образом разбиению соответствует последовательности IUPAC кодов. Позиции последовательности, в которых находится более одного пика, называются вырожденными.

Хроматограмма и ее разбиения представляются в виде скрытой Марковской модели (HMM) которая является обобщением модели предложенной в работе [15]. Наилучшее разбиение соответствует нуклеотидной последовательности хроматограммы.

Для наилучшего разбиения, составляющие ее типы ДНК определяются жадным способом на основании спектра частот нуклеотидов в позициях (профиля) и гомологии с последовательностями словаря. На каждой итерации алгоритма понижаются амплитуды пиков через которые прошло наилучшее выравнивание последовательностей из словаря, таким образом происходит вычитание предсказанной последовательности. Предсказанные жадным способом типы ДНК объединяются в кластеры на основании парных расстояний между профилями амплитуд пиков, которые были получены на соответствующих итерациях алгоритма. Алгоритм максимизации ожидания (EM - expectation maximization), аналогичный предложенному в работе [17], используется для определения числа кластеров и соответствующих им консенсус последовательностей, а также их частот в смеси.

Алгоритм определения делеционных мутаций анализирует множественное выравнивание типов, порожденных жадным алгоритмом до этапа EM. На этом этапе осуществляется сборка главной консенсус последовательности из данных полученных для всех хроматограмм, относящихся к одному образцу. Главная консенсус-последовательность соответствует ДНК типам, составляющим наибольшую долю в смеси, которые не содержат делеций по отношению друг к другу. Точность определения положения делеций положительно зависит от покрытия (количества хроматограмм).

3. Результаты

Мы исследовали применимость алгоритма на примере следующих задач: определение нуклеотидной последовательности хроматограмм для HAV и HDV, сборки главных консенсус последовательностей и определения наличия делеций/вставок в последовательностях вирусов иммунодефицита человека-1 (HIV), гепатита B (HBV) и *M. tuberculosis*, внутри фрагментов, геномов, кодирующих гены, вовлеченные в формирование лекарственной устойчивости и анализа сложных смешанных хроматограмм для 16S рРНК бактериальных сообществ биоптатов желудка.

3.1. Определение нуклеотидной последовательности

Алгоритм BCV показал высокую специфичность (более 90%), преимущество по сравнению с другими приложениями (PolyScan [6], TraceTuner [3] и ABI basecaller [16]) составило от 1 до 5

3.2. Определение делеционных мутаций

Из 123 образцов *M. tuberculosis* в четырех образцах была обнаружена смесь где наряду с последовательностью дикого типа была последовательность, содержащая делеции/вставки. В таблице 1 представлены результаты определения положений делеций/вставок для шести образцов секвенированных фрагментов геномов, мутации в которых определяют устойчивость к терапии HIV, HBV и *M. tuberculosis*. Протяженность делеций в исследованных образцах составляла 1 - 12 нуклеотидов. Для некоторых образцов предсказания, сделанные BCV были подтверждены секвенированием после клонирования (см. таб. 1). В том случае, когда предсказания положений делеций/вставок совпадают более чем по одной хроматограмме встречного секвенирования, результаты также были подтверждены клонированием (таб. 1) Таким образом, чувствительность и специфичность метода определения положений делеций/вставок зависит от покрытия секвенируемой области генома.

Организм	Образец	Аннотированные делеции/вставки (тип, позиция)	Направление и число хроматограмм (nF/mR)	Предсказанные делеции/вставки (тип, позиция)
HIV	vqa10_01	(0,+12), 2138	R	(0,+12), 2138
			F	(0,+11), 2150
				(0,+1), 2179
M. tuberculosis	11042	(-6,0), pncA 297	F/R	(-6,0), pncA 297
		(0,+1), pncA 591	F/R	(0,+1), pncA 591
	ms41	n/a	F/R	(-1,0), pncA 252
	2243	n/a	F	(0,+2), pncA 198
	2687	n/a	F/R	(+1,0), pncA 488
HBV	BV1	(0,+1), 589	2F	(0,+1), 589

Таблица 1: Обнаруженные делеции/вставки в смесях типов ДНК. Тип делеции/вставки - это доминирующая группа типов ДНК в сравнении с “диким типом”, минорная группа типов ДНК в сравнении с “диким типом”. Например (0,+1): минорная группа типов ДНК имеет вставку 1 нк. по сравнению с последовательностью “дикого типа”.

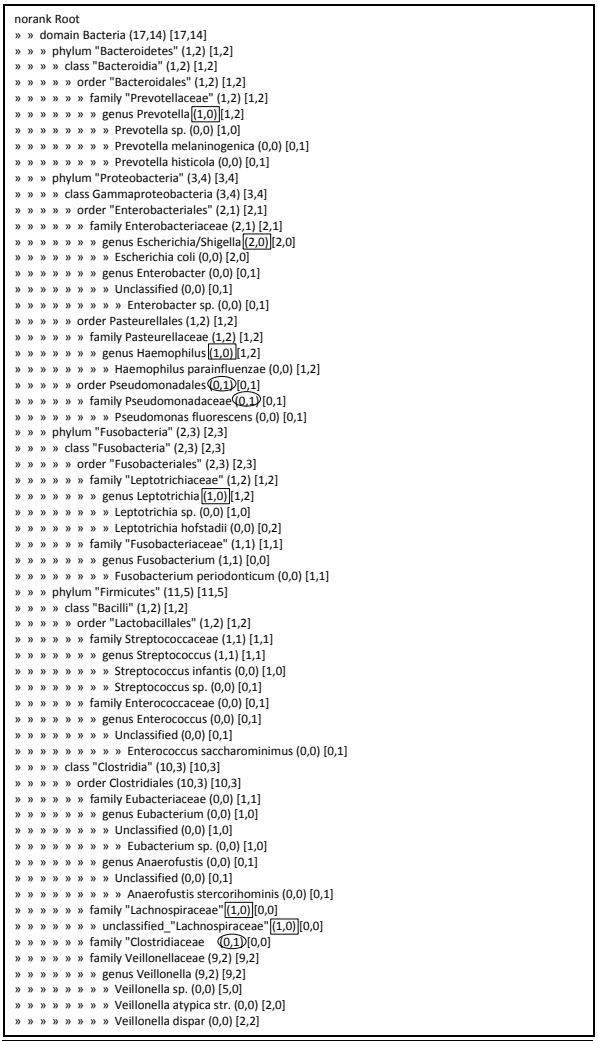
3.3. Анализ микробных сообществ в клинических образцах по 16S рРНК

Мы использовали алгоритм BCV для анализа микробных сообществ в клинических образцах человека по 16S рРНК. ПЦР и уреазный тесты на наличие H. pylori были отрицательны в образце биопсии слизистой желудка 95 и положительны в образце 97. Классификация последовательностей была получена с помощью RDP Classifier [18] для клонов и модифицированного метода основанного на STAP [19] для предсказанных типов ДНК.

Результаты сравнения двух методов исследования разнообразия микробной популяции клинических образцов человека по 16S рРНК представлены на рисунке 1. Данный метод в некоторых случаях обеспечивает более детальную классификацию по сравнению со STAP [19]. Для всех последовательностей предсказанных с помощью BCV классификация STAP [19] однозначно отображалась на классификацию RDP [18]. Таксономические категории, к которым относятся последовательности, полученные только одним из методов, выделены цветом. Разнообразие бактерий в образце оцененное секвенированием небольшого числа клонов (10-20), хорошо совпадает с результатами прямого секвенирования с последующим анализом с помощью BCV. Последовательности полученные после клонирования, в основном были классифицированы до рода, соответствующие этим родам семейства содержат последовательности предсказанные обоими методами (рис. 1). Это объясняется двумя факторами: для некоторых таксономических групп последовательности базы STAP имеют ограниченную глубину классификации (обычно до семейств), и уровень шума в некоторых предсказанных последовательностях не позволяет обеспечить большую глубину классификации. Несмотря на это, несомненным достоинством метода классификации, основанном на STAP, является достоверность таксономической категории, к которой была отнесена последовательность.

Мы провели поиск предсказанных последовательностей по базе 16S рРНК Greengenes [20]. Максимальное сходство предсказанных последовательностей с известными 16S рРНК составляет 80-99%. Для 19 из 21 предсказанных BCV последовательностей для обоих образцов наиболее близкие последовательности из базы Greengenes относятся к тому же семейству или роду к которому были отнесены соответствующие предсказанные последовательности классификатором STAP (рис. 1). В тех случаях, когда нет противоречия между таксономиче-

А.



Б.

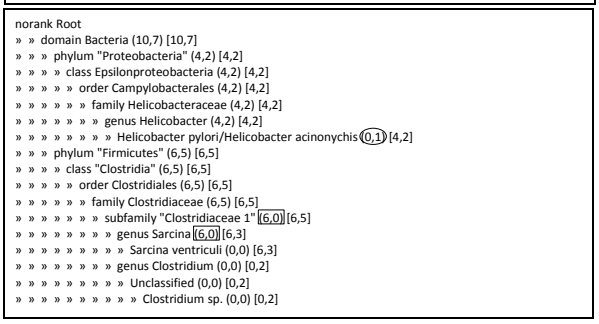


Рис. 1: Сравнение результатов классификации последовательностей ДНК типов предсказанных программой BCV с секвенированными клонами ПЦР продукта. Числа в скобках число секвенированных клонов и число предсказанных последовательностей в таксономической категории. Сравнение методов производится относительно классификации RDP. Различия выделены: подчёркиванием — категория представлена только предсказанными типами ДНК; серым фоном — категория представлена только клонами.

ской принадлежностью классификации предсказанных последовательностей с помощью blastn и таксономической категорией, определяемой методом, основанным на STAP, можно рассматривать blastn как подходящий метод для уточнения классификации.

4. Обсуждение

Новый алгоритм расшифровки хроматограмм прямого секвенирования — BCV обладает двумя уникальными возможностями: определять типы ДНК присутствующие в исследуемом образце и определять наличие в образце типов ДНК содержащие делеции/вставки по отношению к главной консенсус-последовательности. Существующие к настоящему времени альтернативные подходы к решению данной задачи предусматривают, что изначально известна последовательность дикого типа и список разрешенных мутаций [14] или число возможных типов ДНК жестко фиксировано (диплоидный геном) [6, 10]. Многие микроорганизмы и вирусы, особенно РНК-содержащие вирусы, обладают очень высокой скоростью эволюции, для них сложно определить, какую последовательность считать диким типом. Предложенный нами алгоритм свободен от обоих ограничений.

В нашей работе показаны варианты использования программы, представляющие интерес для применения в клинической практике: выявления мутаций в гене *pncA M. tuberculosis*, ассоциированных с устойчивостью к антибактериальному препарату пиразинамиду, определение главной консенсус-последовательности сложных хроматограмм, полученных для гена протеазы ВИЧ и оценки состава сложных смесей, таких как бактериальные популяции по 16S рРНК. Метод прямого секвенирования может использоваться для анализа бактериальных популяций только для исследования клинического материала, который в норме является стерильным, что связано с высоким пределом детекции для метода секвенирования по Сенгеру.

Основным ограничением метода является высокая вариация амплитуды пиков в хроматограммах секвенирования по Сенгеру, именно это ограничение затрудняет предсказание ДНК типов, для которых в словаре нет надлежащего гомолога. Модификация BCV для пиросеквенаторов (имеется ввиду не NGS секвенаторы), обладающих значительно лучшим отношением сигнал/шум, может существенно улучшить предсказательную способность метода.

Список литературы

[1] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, *Basecalling of automated sequencer traces using phred.*

I. Accuracy assessment., 1998, Genome Research, Vol. 8, No. 3, 175-185

[2] B. Ewing, and P. Green, *Basecalling of automated sequencer traces using phred. II. Error probabilities.*, 1998, Genome Res, Vol. 8, No. 3, 186-194

[3] G. Denisov, D. Ho, M. Mettler, J. Candlin, and T. Hunkapiller, *TraceTuner–Next Generation Base Calling*, `www.paracel.com`, A Celera Business, pp. 1-7, Sep. 2000

[4] P. Gajer, M. Schatz, and S. L. Salzberg, *Automated correction of genome sequence errors*, 2004, Nucleic Acids Research, vol. 32, no. 2, pp. 562-569

[5] J. A. Galves, A. Quitzau, and Z. Dias, *New strategy to detect single nucleotide polymorphisms*, 2006, Genetics and Molecular Research: GMR, vol. 5, no. 1, pp. 143-153

[6] K. Chen, M. D. Mclellan, L. Ding, M. C. Wendl, Y. Kasai, R. K. Wilson, and E. R. Mardis, *PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data*, Genome Research, Vol. 17, No. 5., May 2007, pp. 659-666.

[7] D. A. Nickerson, V. O. Tobe, and S. L. Taylor, *PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing*, 1997, Nucleic Acids Research, vol. 25, no. 14, pp. 2745-2751

[8] M. Stephens, J. S. Sloan, P. D. Robertson, P. Scheet, and D. A. Nickerson, *Automating sequence-based detection and genotyping of SNPs from diploid samples*, 2006, Nature Genetics, vol. 38, no. 3, pp. 375-381

[9] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P. Y. Kwok, and W. R. Gish, *A general approach to single-nucleotide polymorphism discovery*, 1999, Nat. Genet 23, 452-456

[10] D. A. Dmitriev, and A. R. Rakitov, *Decoding of Superimposed Traces Produced by Direct Sequencing of Heterozygous Indels*, 2008 PLoS Comput Biol, 4(7)

[11] E. Seroussi, M. Ron, and D. Kedra, *ShiftDetector: detection of shift mutations*, Bioinformatics, 2002, vol. 18, no. 8, pp. 1137-1138

[12] A. Pozhitkov, K. Stemshorn, and D. Tautz, *An algorithm for the determination and quantification of components of nucleic acid mixtures based on single sequencing reactions*, 2005, BMC Bioinformatics, vol. 6, p. 281

[13] P. Trosvik, B. Skanseng, K. S. Jakobsen, N. C. Stenseth, T. N?s, and K. Rudi, *Multivariate Analysis of Complex DNA Sequence Electropherograms for High-Throughput Quantitative Analysis of Mixed Microbial Populations*, 2007, Applied and Environmental Microbiology, vol. 73, no. 15, pp. 4975-4983

[14] A. Wildenberg, S. Skiena, and P. Sumazin, *Deconvolving sequence variation in mixed DNA populations*, 2003, J Comput Biol., 10(3-4), pp. 635-652

[15] L. Andrade-Cetto, and E.S. Manolakos, *A Graphical Model Formulation of the DNA Base-Calling Problem*, 2005 IEEE Workshop on Machine Learning for Signal Processing, 369-374

[16] T. L. Hagemann, and S.-P. Kwan, *ABI Sequencing Analysis: Manipulation of Sequence Data from the ABI DNA Sequencer*, 1999, Molecular Biotechnology 13 (2): 137-152, doi:10.1385/MB:13:2:137

[17] C. Quince, A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan, *Accurate determination of microbial diversity from 454 pyrosequencing data*, 2009, Nat Meth 6 (9): 639-641

[18] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy*, Applied and Environmental Microbiology, vol. 73, no. 16, pp. 5261-5267, Aug. 2007

[19] D. Wu, A. Hartman, N. Ward, and J. A. Eisen, *An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP)*, PLoS ONE, vol. 3, no. 7, p. e2566, Jul. 2008

[20] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB*, Applied and Environmental Microbiology, vol. 72, no. 7, pp. 5069-5072, Jul. 2006