О проблеме инвариантности межгеномной дистанции у прокариот

[1]Василенко О.В., [2]Георгиева З.Д.

[1]ФИЦ «Пущинский научный центр биологических исследований РАН»,
(Институт биохимии и физиологии микроорганизмов им. Г.К. Скрябина РАН, ВКМ)
[2]ФГБОУ ВО Кубанский Государственный университет, биологический факультет;
ovvasilenko@gmail.com

Рассмотрена проблема устойчивости значений относительной межгеномной дистанции (ОМД) при изменении качества геномной сборки, или, иначе говоря, - зависимость относительной межгеномной дистанции, вычисленной через среднюю нуклеотидную идентичность (ANI) от значения среднего покрытия геномной сборки и общепринятых критериев качества сборки. Соотношение между ними простое: ОМД = 100% - ANI.

С развитием геномики роль так называемых геномных индексов - особых показателей, вычисляемых непосредственно из последовательности нуклеотидов генома, - возрастает. Они стали главнейшим критерием в определении принадлежности штамма к одному из известных или к новому виду (1). Среди них ANI (ОМД) выделяется тем, что имеет объективный и интуитивно понятный общий алгоритм вычисления, высокую точность вычисления и вполне определенный фундаментальный (эволюционный) смысл. Несколько способов вычисления ANI соответствуют нескольким разновидностям ANI. Пригодны для практического применения три из них (ANIb, orthoANI и FastANI), которые при соблюдении определенных условий дают почти идентичные результаты (1-5). Особое значение приобрел ANI после того, как было показано, что секвенирование геномов решает те проблемы в идентификации клинических образцов бактерий, которые не решает масспектрометрический подход (MALDI-TOFF) (6). Важность проблемы, эффективность подхода, вектор научно-технического прогресса вкупе с авторитетом и энергией авторов данной инициативы не оставляют сомнений в скором утверждении этого подхода в качестве клинического стандарта в дополнении к фундаментальному таксономическому. Тем более удивляет то обстоятельство, что до настоящего момента не изучен вопрос о том, как зависит ОМД от качества сборки. Есть общие слова о пригодности даже неполных драфт-геномов для вычисления ANI, но нет других, кроме эстетических, критериев пригодности геномной сборки для видовой идентификации штаммов. И если отвергнутый журналом геном по таким поснованиям, как "слишком большое количество контигов" - это неприятность, которую можно пережить, то несвоевременное принятие клинически важного решения по аналогичному поводу - это то, что может иметь, без преувеличения, тяжкие последствия.

Наиболее существенная причина, по которой бывает невозможно собрать геном нужного качества, - это недостаток подходящих первичных ридов. Их дополнительная наработка - это новый дорогостоящий запуск секвенатора, который надо не только оплатить, но и собрать нужный объем библиотек, а затем просто провести сам запуск. Всегда ли это необходимо и оправдано? Это очень важный вопрос, и он до настоящего времени был без ответа. Мы поставили задачу ответить на него моделированием ситуаций с различным качеством сборок, которое является следствием изменяющегося покрытия. Для этого мы выбрали пять различных сборок бактерий хорошего и удовлетворительного качества, для которых были нам также доступны и первичные риды в формате fastq. Список геномов и коды доступа приведены в Таблице 1. Мы отбирали псевдослучайным образом только часть ридов так, что прогнозируемое покрытие составляло меньшую величину, чем исходная "образцовая" сборка, собирали риды при помощи программы SPAdes 3.13.1 (7), получая несколько - до десяти - сборок, отличающихся покрытием и зависящим от него качеством. Сборки сравнивали между собой и исходной, взятой в качестве "референса", при помощи

программы QUAST 5.0.2 (8). Настройки программы были таковы, что она выдавала ряд параметров сборки не только формально-статистического описательного типа (N50, число контигов/скаффолдов и т.п.), но и из области функциональной геномики.

Табл. 1. Список геномных сборок, использованных в работе

Код доступа	Видовое название	Штамм	Платформы	Покрытие
MASI00000000.1	*Methyloligella halotolerans*	VKM B-2706^T (C2^T)	IonTorrent	115
MCRI00000000.1	*Methylophaga muralis*	VKM B-3046^T (Bur 1^T)	IonTorrent	75
MUKN01000000	*Rathayibacter* sp.	VKM Ac-2630	IonTorrent	34
FXBM00000000.1	*Rathayibacter oskolensis*	VKM Ac-2121^T	Illumina IonTorrent	222
NZ_FXAY00000000.1	*Agreia pratensis*	VKM Ac-2510	Illumina	214

Межгномную дистанцию между сборками и внешним, но филогенетически близким геномом вычисляли при помощи программы orthoANI (4). В программу Microsoft Excel загружали данные QUAST, orthoANI и заданные величины покрытия. Строили графики зависимостей от величины покрытия таких эмпирических функций, как ОМД, а также вычисленных в QUAST показателей качества сборки (число контигов с длиной больше нуля (# contigs (>= 0 bp)); общая длина конигов (Total length (>= 0 bp)); наибольший контиг (Largest contig); длина медианного контига/скаффолда (N50); число неопределенностей на 100 килобаз (# N's per 100 kbp); доля полных обязательных ортологичных генов в сборке (Complete BUSCO (%)); и, соответственно, неполных генов (Partial BUSCO (%)), а также GC-состав (GC (%)). Для того, чтобы все функции можно было отразить на одном графике, их значения нормировали по максимальному из их выборки, которое принимали за 100%, конечно, если значения функций не были выражены в процентах в пределах 0-100% по умолчанию.

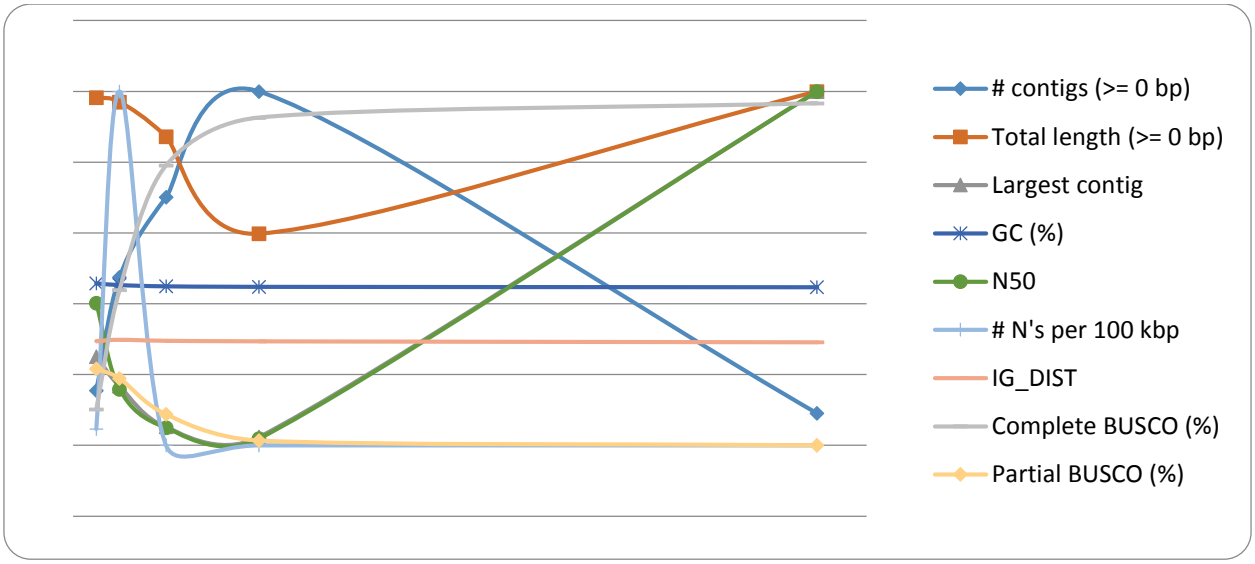

Рис. 1. График зависимостей от глубины секвенирования (покрытия, ось абсцисс) показателей качества сборок и других параметров (объяснения в тексте выше). Ось ординат - все значения функций в процентах. Объект: сборки *Methylophaga murali.*

Типичный результат показан на Рис. 1. Отчетливо видно, что ОМД не зависит ни от одного из показателей качества сборки и ни с одним из них не коррелирует, не зависит и от величины покрытия. Точно так же ведет себя GC-состав, который, очевидно, является коровой характеристикой генома и стабилен по закону больших чисел, если в сборке есть достаточное количество "букв". Аналогия с ОМД не случайная, но для описания данного явления не существует общепризнанных терминов. Образно говоря, коровое свойство генома - это его собственные характеристические координаты в пространстве эволюционных путей. Поэтому как только появляется модель генома (сборка), у него появляются определенные характеристические расстояния до соседних геномов как следствие фиксированности координат. Действует закон "все или ничего": есть геном или сборка - есть координаты и расстояния, нет сборки (генома) - нет координат.

Вывод. ОМД и ANI являются наиболее нетребовательными к качеству сборки параметрами, производными из генома. В области малых покрытий, много меньших, чем десять, погрешность определения межгеномной дистанции до внешнего штамма не превышает обычно 1%, в то время как внутривидовая изменчивость межгеномной дистанции может быть до 5%. Нет необходимости чрезмерно пренебрегать качеством секвенирования, если геномные сборки будут использованы только для систематики или типирования штаммов в прикладных целях (в медицине), но так же и нет необходимости уклоняться от принятия таксономических или, тем более, диагностических решений на том основании, что сборка выглядит неэстетично, - для этого не найдено ни оснований ни оправданий.

Работа выполнена при поддержке гранта РФФИ 18-04-01347.

Литература

1. Chun J, Oren A, Ventosa A, Christensen H, Arahal D, da Costa M, Rooney A, Yi H, Xu X, De Meyer S, Trujillo M. 02/01/2018. Int J Syst Evol Microbiol 68(1):461-466 doi:10.1099/ijsem.0.002516.
2. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl Acad. Sci. 2009;106:19126–19131. doi: 10.1073/pnas.0906412106.

3. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr. 2016;4:e1900v1.
4. Lee I, Kim YO, Park SC, Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int. J. Syst. Evol. Microbiol. 2016;66:1100–1103. doi: 10.1099/ijsem.0.000760.
5. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114. Published 2018 Nov 30. doi:10.1038/s41467-018-07641-9.
6. Ha SM, Kim CK, Roh J, Byun JH, Yang SJ, Choi SB, Chun J, Yong D. Application of the Whole Genome-Based Bacterial Identification System, TrueBac ID, Using Clinical Isolates That Were Not Identified With Three Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) Systems. Ann Lab Med. 2019 Nov;39(6):530-536. https://doi.org/10.3343/alm.2019.39.6.530.
7. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics, 2013, 29 (8), 1072-1075.
8. Bankevich A., Nurk S., Antipov D., Gurevich A., Dvorkin M., Kulikov A. S., Lesin V., Nikolenko S., Pham S., Prjibelski A., Pyshkin A., Sirotkin A., Vyahhi N., Tesler G., Alekseyev M. A., Pevzner P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 2012.