

№18 2001 год ЧТО СОДЕРЖИТ ПОЛНЫЙ ГЕНОМ *ESCHERICHIA COLI*?

В настоящее время E. coli, несомненно, представляет собой самую изученную клетку из всех существующих.

Г.Стент

Кишечная палочка *Escherichia coli* — классический объект молекулярной генетики, на котором исследованы наиболее принципиальные проблемы организации генетического материала. Штамм *E. coli* K12 был успешно использован Дж.Ледербергом и Э.Тейтумом в 1946 г. для доказательства существования рекомбинаций у бактерий. Позже Дж.Ледерберг построил для нее первую генетическую карту, а Ф.Жакоб и Э.Вольман — первую кольцевую карту. В 1963 г. Дж.Кернс сфотографировал кольцевой геном *E. coli* в процессе его репликации.

Лет 30 назад огромным событием в генетике было секвенирование первого гена. К концу 1970-х годов технология секвенирования упростилась и стала рутинной процедурой. После этого были секвенированы геномные ДНК и РНК сначала

малых фагов (*R17, Q β , ϕ X174* и др.), затем больших фагов и вирусов (фага λ , вирусов коровьей оспы, оспы и др.), клеточных органелл (митохондрий и хлоропластов) и плазмид (F-фактора и др.). Наконец, в конце 1980-х годов были начаты первые международные программы секвенирования полных клеточных геномов бактерий, грибов, растений, насекомых, млекопитающих, человека.

Зачем это нужно? Клеточный геном представляет собой сбалансированную систему генов — архив генетической информации, достаточной для контроля всего клеточного метаболизма, развития, морфогенеза, самовоспроизведения [1, 2]. В частности, геном клетки содержит гены всех основных генетических процессов — репликации, транскрипции, трансляции, репарации, рекомбинации, сегрегации и т.д. Полное секвенирование генома позволяет сопоставить и оценить генетическую сложность тех или иных молекулярных систем и геномов, выявить ранее неизвестные гены, выполнить сравнительный анализ функционального и структурного сходства различных генов и геномов, выявить общие принципы организации сложных клеточных молекулярно-генетических систем управления.

Работа по проекту полного секвенирования генома *E. coli* K12 была начата в 1991 г. под руководством д-ра Фреда Блаттнера (лаборатория генетики, Висконсинский университет, г. Медисон, США). В январе 1997 г. основные результаты были переданы в компьютерную базу данных *GenBank* [3], а в сентябре 1997 г. в американском журнале «*Science*» появилась итоговая статья коллектива участников секвенирования [4]. Полная последовательность ДНК генома *E. coli* K12 стала достоянием науки. Ниже мы приведем в сводной форме основные результаты этих работ с необходимыми комментариями, имея в виду, что такой уникальный материал позволяет ответить на многие принципиальные вопросы молекулярно-генетической организации и эволюции.

Таблица 1

Общие характеристики некоторых секвенированных прокариотических и эукариотических клеточных геномов

Биологический объект, вид	L (Мб)	Число цистронов	Год
1. <i>Mycoplasma genitalium</i>	0.580	470	1995
2. <i>Mycoplasma pneumoniae</i>	0.816	677	1996
3. <i>Borrelia burgdorferi</i>	0.910	853	1997
4. <i>Aquifex aeolicus</i>	1.551	1512	1998
5. <i>Methanococcus jannaschii</i>	1.66	1738	1996

6. <i>Helicobacter pylori</i>	1.667	1590	1997
7. <i>Methanobacterium thermoautotrophicum</i>	1.751	1855	1997
8. <i>Haemophilus influenzae</i>	1.830	1743	1995
9. <i>Archaeoglobus fulgidus</i>	2.178	2436	1997
10. <i>Bacillus subtilis</i>	4.214	4100	1997
11. <i>Escherichia coli</i>	4.639	4288	1997
12. <i>Saccharomyces cerevisiae</i>	12.068	5885	1996
13. <i>Caenorabditis elegans</i>	100.0	12178	1998
14. <i>Drosophila melanogaster</i>	120.0	13600	2000
15. <i>Homo sapiens</i>	2910.0	38588	2001

Примечание. Таблица составлена по данным оригинальных работ, опубликованных в журналах «Nature», «Science», «NAR» и др. в 1995-1998 гг., а также базы данных GenBank [3] и последних публикаций. Методы оценки и сравнения возможных цистронов и их белков допускают некоторые неоднозначности в интерпретации. Поэтому оцененные числа цистронов не надо воспринимать как окончательные. Оценки некоторых ORF как цистронов могут быть уточнены.

Заметим также, что параллельно с *E. coli* были секвенированы многие другие клеточные геномы бактерий и эукариот. К концу 1997 г. было опубликовано 8 полных клеточных геномов, а к лету 1998 г. — уже 15 (табл. 1). Среди них геномы микоплазм, энтеробактерий, археобактерий, дрожжей, нематоды. На очереди геномы других бактерий и грибов, а также дрозофилы, арабидопсиса, пшеницы, риса, кукурузы, мыши и, наконец, человека. В целом это направление теперь называют **геномикой**. Это, вероятно, одна из главных точек роста современной молекулярной генетики.

Непосредственно для секвенирования была выбрана линия *E. coli K12 MG 1655*, из которой были исключены **профаг λ и F-плазида**, а другие генетические манипуляции были сведены к минимуму. Длина генома этой линии $L=4639221$ н.п. Эта последовательность отвечает кольцевой генетической карте *E. coli K12*, калиброванной на 100 минут по времени конъюгационного переноса. Начало отсчета карты выбрано между генами *lasT* и *thrl*.

Общие характеристики последовательности ДНК генома *E. coli K12* таковы: 87,8% генома занимают реальные и вероятные белок-кодирующие гены, или цистроны. Примерно 1/3 из них была известна ранее, а остальные выбраны среди огромного числа новых открытых рамок трансляции (возможных цистронов, или ORF) путем сложного сопоставления многих свойств, имеющих характерные различия между кодирующими и не кодирующими районами. Функции 38% этих цистронов неизвестны.

— 0,8% — гены стабильных фракций РНК (т-РНК, р-РНК и др.).

— 0,7% — не кодирующие повторы.

— 11,0% генома — функциональные сайты и другие участки, выполняющие регуляторные и другие функции.

Таким образом, геном *E. coli K12* очень плотно нагружен генами (~ 88,5%), а межгенные участки занимают относительно малую долю (~11%). Среди 4288 выявленных или предсказанных цистронов 1853 описаны ранее, а 2435 — новые. Самый большой цистрон содержит 7149 нп. (2383 кодона), функция его неизвестна. Средний размер цистрона 951 нп. (317 кодонов). Средний интервал между цистронами — 118 нп. Однако межгенные интервалы в большинстве своем содержат различные функциональные сайты, то есть выполняют регуляторные функции. Кроме того, цистроны не содержат интронов — внутренних не кодирующих участков.

Известно, что цистроны выделяются в ДНК и м-РНК начальными и конечными знаками пунктуации. В общей форме они были известны ранее и внесены в генетический код. Однако в геноме *E. coli* они встречаются с различными частотами:

Начальные знаки пунктуации

Конечные знаки пунктуации

GTG — 612

TGA — 1257

TTG — 130

TAG — 326

ATT — 1

CTG — 1

Интересно, что у 405 пар смежных цистронов вообще нет межгенных интервалов: знак начала трансляции одного частично перекрывается с конечным знаком другого:

(нач)	(нач)	(нач)	(нач)	
ATGA,	TAATG,	TGATG,	GTGA,	и др.
[кон]	[кон]	[кон]	[кон]	

По данным на январь 1998 г. [5] сложность молекулярно-генетической системы управления и метаболической сети *E. coli* можно охарактеризовать следующим образом:

1. Длина ДНК генома (Мб)	4.6
2. Полное число генов	4909
3. Число цистронов	4288
4. Число кодируемых ими ферментов	804
5. Число метаболических реакций	988
6. Число метаболических путей	123
7. Число химических веществ, участвующих в метаболизме	1303
8. Число фракций т-РНК (генов т-РНК)	79 (86)
9. Число регуляторных белков	60

В таких случаях специалисты говорят: «жизнь при 4909 генах». Метаболизм сложен, но не запредельно. В дальнейшем приведенные цифры могут возрасти в ходе исследований за счет новых знаний.

Более подробная классификация цистронов по 22 функциональным классам представлена в таблице 2. Здесь примерно 1/4 клеточных ресурсов связана с метаболизмом малых молекул, 1/8 — с метаболизмом макромолекул и 1/5 — с клеточными структурами и процессами. В метаболизме малых молекул ключевую роль играет синтез, распад и преобразование нуклеотидов (58 цистронов), аминокислот (131); энергетические процессы (243), транспорт (146), центральный промежуточный метаболизм (188) и другие процессы. В частности, системы, выполняющие основные генетические процессы, содержат:
— репликацию, рекомбинацию, модификацию и репарацию ДНК — 115 (2,68%);
— транскрипцию, синтез, метаболизм и модификацию РНК — 55 (1,28%);
— трансляцию и посттрансляционную модификацию белков — 182 (4,24%) + 21 ген р-РНК + 86 генов т-РНК.

Кроме того, найдено 9 цистронов, контролирующих синтез **шаперонов** — вспомогательных белков, способствующих формированию правильной пространственной упаковки всех остальных белков. Этот процесс называется самоорганизацией, или фолдингом белков.

Помимо различных вспомогательных функций, эти системы в совокупности образуют **сайзер** — универсальную систему самовоспроизведения клетки [1, 2]. Сайзер составляет сердцевину молекулярно-генетической системы управления клетки. Несмотря на внушительное число участвующих генов (~460, свыше 10% всех генов), принципиальная блок-схема сайзера достаточно проста [1, 6].

Далее рассмотрим разнообразие функциональных единиц транскрипции. Для генома *E. coli* и других энтеробактерий характерно присутствие управляемых единиц транскрипции — **оперонов**. Первые опероны были открыты именно у *E. coli*: *lac*-оперон, контролирующий сбраживание сахара лактозы, *trp*-оперон, контролирующий синтез аминокислоты триптофана, и др. [см. 6]. Важной особенностью оперонов является наличие обратной связи между концентрацией контролируемого метаболита и наработкой ферментов его синтеза или распада. Всего в геноме *E. coli* выявлено и предсказано 2584 оперона. Среди них:

- 73% содержат 1 цистрон;
- 16% — 2 цистрона;
- 4,6% — 3 цистрона (в том числе *lac*-оперон);
- 6% — 4 и более цистронов (в том числе *trp*-, *his*-опероны).

Все они имеют не менее 1 **промотора**— начального знака транскрипции.

Опероны управляются регуляторными белками через специфические функциональные сайты управления. Например, **белок-репрессор** *lac*-оперона узнает его **оператор** — функциональный сайт и через него подавляет функцию инициации транскрипции. Иногда опероны подчинены нескольким регуляторным белкам и имеют несколько регуляторных сайтов [1, 6].

Всего по данным секвенирования выявлены 45 цистронов белков с регуляторными функциями и еще цистроны 133 предполагаемых регуляторных белков. Большинство из них, вероятно, участвует в управлении оперонами.

Так, внутри областей с предсказанными сайтами управления (в основном оперонов)

- 89,2% регулируются 1 белком (в том числе *trp*-оперон);
- 8,4% — 2 белками (в том числе *lac*-оперон);
- 2,4% — 3 и более белками.

В свою очередь эти области содержат

- 81,2% — 1 сайт управления;
- 12,2% — 2 сайта управления (в том числе *lac*-оперон);
- 6,6% — 3 и более сайтов управления.

Это значит, что большинство оперонов регулируется достаточно просто. Этим они существенно отличаются от генов эукариот, которые подвержены действию многих общих и специфических белковых факторов управления.

Геном *E. Coli* содержит 2 функциональные единицы репликации. Ф.Блаттнер и др. назвали их **реплихорами** [4].

Общее двустороннее начало репликации (*ori*, *origin*) локализовано на участке примерно 84,5 мин конъюгационного переноса и занимает ~ 250 нп. В этой зоне иницируется двусторонняя репликация. Реплихор 1 ориентирован по часовой стрелке, реплихор 2 — против нее. Оба процесса заканчиваются на противоположном участке генетической карты, ~ 34-35 мин., где каждый из них имеет свой отдельный ориентированный терминальный знак (*ter*) — **T1** и **T2**. Следует отметить, что традиционно участки репликации, ограниченные знаками *ori* и *ter*, называют **репликонами** [6].

Таблица 2

Распределение цистронов и белков *E. coli* по 22 функциональным классам

Функциональный класс	Число белков	%
1. Регуляторная функция	45	1.05
2. Предполагаемая регуляторная функция	133	3.10
3. Структура клетки	182	4.24
4. Предполагаемые мембранные белки	13	0.30
5. Предполагаемые структурные белки	42	0.98
6. Фаги, транспозоны, плазмиды	87	2.03
7. Транспортные и связывающие белки	281	6.55
8. Предполагаемые транспортные белки	146	3.40
9. Энергетический метаболизм	243	5.67

10. Репликация, рекомбинация, модификация и репарация ДНК	115	2.68
11. Транскрипция, синтез, метаболизм и модификация РНК	55	1.28
12. Трансляция, посттрансляционная модификация белков	182	4.24
13. Клеточные процессы, включая адаптацию и защиту	188	4.38
14. Биосинтез кофакторов, простетических групп и носителей	103	2.40
15. Предполагаемые шапероны	9	0.21
16. Биосинтез и метаболизм нуклеотидов	58	1.35
17. Биосинтез и метаболизм аминокислот	131	3.06
18. Метаболизм жирной кислоты и фосфолипидов	48	1.12
19. Катаболизм соединений углерода	130	3.03
20. Центральный промежуточный метаболизм	188	4.38
21. Предполагаемые ферменты	251	5.85
22. Другие известные гены (генные продукты и фенотипы неизвестны)	26	0.61
23. Гипотетические, неклассифицированные, неизвестные	1632	38.06
24. Всего	4288	100.00

Геном *E. coli* K12 содержит также значительное число необязательных (факультативных) включений — профагов, плазмид и транспозонов. Выявлено 87 цистронов и белков этих включений (табл. 2). Число их может быть различным, поскольку они подвижны, способны к внедрению в геном и выщеплению из него. Наилучшим образом это продемонстрировано для умеренного фага I и полового фактора (плазмиды) F, которые в данной линии отсутствуют. Многие фаги исключаются из генома не полностью, оставляя там в качестве следа некоторые свои гены. Эти остатки, не способные к самостоятельному перемещению и развитию, называют «криптическими» фагами. Среди факультативных включений в этой линии найдены 41 копия различных транспозонов (*IS*), которые участвуют в процессах внедрения и исключения плазмид.

Наконец, следует отметить, что геном *E. coli* содержит ряд функциональных и нефункциональных повторов. Октамер GCTGGTGG отвечает «горячим точкам рекомбинаций» (так называемым **χ-сайтам**). Он встречается в сотнях позиций в обеих ориентациях и играет ключевую роль в конъюгационной рекомбинации и других генетических процессах. Найдено большое число копий (581) небольшого палиндромного повтора *REP* длиной ~ 40 нп. Функция их неизвестна. В сумме они занимают 0,54% ДНК генома. Известны и другие повторы. В основном они попадают в межцистронные интервалы.

Таким образом, молекулярно-генетическая система управления *E. coli* оказалась хотя и сложной, но вполне обозримой. Геном кодирует белки всех основных генетических процессов и систем: контролирует синтез и метаболизм мономеров, энергетику, транспорт, клеточные процессы, защитные реакции. Хотя функции 38% выявленных белков пока не известны, скорее всего, они пополняют уже обозначенные функциональные группы. Опероны, как управляемые единицы транскрипции, являются доминирующим вариантом организации генов *E. coli* K12 и других прокариотических клеток.

Сравнительный анализ молекулярно-генетических систем *E. Coli* и других объектов (см. табл. 1) позволил выявить многие гомологичные гены, оценить степень сходства геномов, а также высказать предположения о минимальной сложности

гипотетической (а может быть, первичной?) клетки. Ясно, что для организации клетки необходим некоторый минимум молекулярных структур и процессов. В таблице 1 приведены суммарные данные по числу генов и размерам первых секвенированных клеточных геномов. Минимальный геном имеет микоплазма *M. Genitalium* — 0.58 Мб, 470 генов. Путем сравнения геномов выявлено, что минимальная клетка, способная к автономной жизнедеятельности и самовоспроизведению, должна была бы содержать не менее 250-300 наиболее существенных генов. Группа японских исследователей [7] показала, что можно обеспечить все основные метаболические потребности клетки, трансляцию и репликацию РНК-генома в системе со 127 генами. Правда, при этом клетка должна быть лишена архива ДНК, репарационных и других важных систем защиты и помехоустойчивости, что делает ее эволюционно беззащитной.

В заключение выражаю благодарность К.С.Макаровой, Ю.И.Вульф и А.Э.Келю за содействие в адаптации последних данных по клеточным геномам.

Литература

1. Ратнер В.А. Концепция молекулярно-генетических систем управления. Новосибирск: Наука, 1993. 120 с.
2. Ratner V.A., Zharkikh A.A., Kolchanov N.A. et al. Molecular Evolution. Berlin e.a.: Springer-Verlag, 1996. 433 p.
3. Сайт в INTERNET: <http://ncbi.nlm.nih.gov/genbank/genomes>
4. Blattner F.R., Plunket III.G., Bloch C.A. et al. The Complete Genome Sequence of *Escherichia coli* K12 // Science. 1997. V. 277. P. 1453-1462.
5. Karp P.D., Riley M. EcoCyc: encyclopedia of *E. coli* genes and metabolism. <http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html>
6. Ратнер В.А. Молекулярная генетика: принципы и механизмы. Новосибирск: Наука, 1983. 256 с.
7. Tomita M. et al. A virtual cell with 127 genes // Proc. 1st Intern. Conf. «Bioinformatics of Genome Regulation and Structure (BGRS'98)», Novosibirsk: Inst. Cytol. Genet., 1998. V. 1. P. 97-99.

В.А.Ратнер, д.б.н., профессор, Институт цитологии и генетики СО РАН, Новосибирск