



**Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Российской академии наук**

Н.Н. Козлов, С.С. Грязнов

Некоторые численные
характеристики
больших геномов.

Препринт №

Москва

Институт прикладной математики
имени М.В. Келдыша
Российской Академии Наук

Н.Н. Козлов, С.С. Грязнов

Некоторые численные характеристики
больших геномов.

Москва 2006

Н.Н. Козлов, С.С. Грязнов НЕКОТОРЫЕ ЧИСЛЕННЫЕ
ХАРАКТЕРИСТИКИ БОЛЬШИХ ГЕНОМОВ. Препринт ИПМ им. М.В.
Келдыша РАН, М., 2006, 28с

Работа посвящена применению математических и компьютерных методов для исследования больших геномов. Представлены результаты первых расчетов для некоторых характеристик 10 больших геномов *C. familiaris*, *G. gallus*, *D. rerio*, *M. musculus*, *R. norvegicus*, *A. mellifera*, *D. melanogaster*, *A. thaliana*, *C. elegans*, в том числе генома человека *H. sapiens*. Полное число исследуемых генов около 200 000. Работа основывается на Теореме (Козлов Н.Н. ДАН (математика) 2007. Т. 412. № 5. (в печати)), которая устанавливает потенциал стандартного генетического кода для решения т.н. задачи блокировки генов применительно к наиболее обширным областям ДНК, где каждый ген кодирует только один белок. Показано, что блокировка не является случайной, не является экстремальной (как минимальной так и максимальной). Для генома человека, собаки, курицы, мыши, крысы, рыба-зебры среднее ее значение плотности блокировки более чем втрое меньше теоретического максимума и в 2.5 раз больше теоретического минимума. Теоретические значения устанавливаются Теоремой. Показано, что плюс - и минус-цепи ДНК блокированы как правило одинаково. Установлено не равновероятное участие блокирующих кодонов в блокировках: кодон TGA использован более 50% или почти столько же как и оба других кодона TAA и TAG. Наибольшее отклонение по этой и другим численным характеристикам (всего анализируются 5 характеристик) обнаружены у медоносной пчелы (*Apis Mellifera*), которую биологи считают вершиной эволюции среди насекомых.

Ключевые слова: структурные гены, блокировка гена, блокировка генома.

N.N.Kozlov, S.S.Gryaznov. Some numeric characteristics of greates genomes. Preprint
KIAM RAN Moscow, 2006, 28p.

In work results of the calculations for some characteristics of 10 greates genomes *C. familiaris*, *G. gallus*, *D. rerio*, *M. musculus*, *R. norvegicus*, *A. mellifera*, *D. melanogaster*, *A. thaliana*, *C. elegans*, including genome *H. sapiens*. Full number of investigated genes nearby 200 000. Work is based on the Theorem (Kozlov N.N. DAN (Mathematics). 2007. T. 412. N 5. (in a press)) which establishes potential of a standard genetic code for the decision of a so-called problem of blocking of genes with reference to the most extensive areas of DNA where each gene codes only one protein. It is shown, that blocking is not casual, not extreme (as minimal and maximal). For human, dog, chicken, mouse, rat and zebra-fish genomes its average value of density of blocking more than three times less than a theoretical maximum and in 2.5 times more a theoretical minimum. Theoretical values are established by the Theorem. It is shown, that plus – chain and a minus-chain of DNA are blocked as a rule equally. Not equiprobable participation terminated codons in blocking is established: TGA it is used more than 50 % or nearly so as much as well as both of others TAA and TAG. The greatest deviation on this and other numerical characteristics (all 5 characteristics are analyzed) are found out in a melliferous bee (*Apis Mellifera*) as which biologists consider as top of evolution among insects.

Key words: Structural genes , gene blocking, genome blocking.

Работа выполнена при финансовой поддержке Программы фундаментальных исследований Президиума РАН «Фундаментальные проблемы информатики и информационных технологий», Российского фонда фундаментальных исследований (коды проектов 04-01-00320, 05-01-00113 и 05-07-90026), а также гранта ведущих научных школ (НШ-2448.2006.1)

Предисловие.

После завершения в 2003 году гигантского Международного проекта по геному человека, приуроченного к 50-летию классической работы Уотсона и Крика по пространственной структуре ДНК [1] наступила постгеномная эпоха. Многомиллиардные вливания в исследования по всестороннему анализу человеческого генома и других больших геномов ежегодно увеличиваются и это направление исследований признано одним из приоритетных в мировой науке. Показательно в связи с этим недавнее сообщение о награде в 10 млн. \$ за создание быстрого и дешевого метода расшифровки генома человека. Премия объявлена Фондом X-премии (X-Prize Foundation). По условиям конкурса быстрый метод расшифровки генома означает возможность расшифровать 100 любых человеческих геномов за 10 дней, а за 6 месяцев – еще 100 геномов, но по выбору фонда [2]. Специалисты полагают, что использование такого метода позволит получать сведения об индивидуальных геномах за 1000 \$, или более чем в миллион раз дешевле стоимости первого человеческого генома. Таким образом, речь может идти не о десятках или тысячах геномов, а даже и о миллионах. При этом возникает задача математического анализа подобной информации. При этом может быть поставлен ряд совершенно новых математических задач. Ниже рассматривается одна из них , связанная с математическим анализом блокировки структурных генов, которая, как полагают биологи, наблюдается для подавляющего большинства генов [3].

Данная работа возникла в ходе исследования по математическому анализу особых случаев записи генетической информации т.н. перекрывающихся генов [4]. В ходе проведения этих исследований возникла постановка задачи, связанной с изучением кодирующих областей ДНК, не содержащих генетических перекрытий, т.е. наиболее протяженных областей ДНК.

Постановка такой задачи не была случайной, некоторые из полученных

выводов позволили разрешить одну из задач, решение которой оказалось невозможным при анализе только перекрывающихся генов. Этот этап исследования в настоящее время завершается и будет представлен в последующих публикациях.

1. Введение.

Структурными генами называют гены, кодирующие белковые последовательности. Каждая из таких последовательностей состоит из аминокислот. Число аминокислот равно 20, и каждая из них кодируется различающимися кодонами. Любой кодон или триплет состоит из трех букв четырехбуквенного алфавита: А, Т, С, G – это нуклеотиды аденин, тимин, цитозин и гуанин соответственно. Таким алфавитом записаны ДНК всех живых организмов, на одной из двух цепей которой располагается структурный ген, записанный последовательностью указанных букв без всяких пропусков, запятых и т.п. Экспериментально установленная таблица стандартного генетического кода показала, что каждая из аминокислот кодируется одним, двумя, тремя, четырьмя, либо шестью кодонами (таблица 1). Помимо кодировок аминокислот были установлены три кодона

ter: TAA, TAG, TGA (1)

являющиеся кодонами терминации – ter, на которых синтез белка останавливается. Если в результате мутации в структурном гене образуется один из трех кодонов ter, это приводит к преждевременному останову синтеза белка в том месте, где расположен мутантный кодон. При этом нарушится функция белка, т.к. синтезируется только часть белковой молекулы.

2. Блокировка генов.

В работе [5] сформулирована Теорема, устанавливающая потенциал стандартного кода, который использует природа для блокировки последовательностей, альтернативных данному гену. Эти 5 последовательностей или 5 рамок считывания – PC1-PC5 – можно вычислить с помощью сдвига на -1 либо +1 нуклеотид в той же цепи ДНК, что и данный ген (это PC1, PC2) и сдвига на -1, 0, +1 в цепи ДНК комплементарной данной (это PC3-PC5) (рис.1). Сама блокировка связана с возникновением в одной или нескольких PC из PC1-PC5, какого-либо из трех кодонов терминации *ter* из (1). Появление такого кодона останавливает синтез белка по альтернативной последовательности. Биологи говорят о мощной биологической защите, которая состоит в том, что если за счет мутаций от PC0 перейдем на одну из пяти PC (PC1-PC5), то белок (как правило эфемерный) синтезируется лишь частично, что не должно повлиять на нормальное функционирование клетки.

Основная задача состоит в том, чтобы установить потенциал генетического кода, который использует природа для указанной блокировки. В формулировке основного результата - Теоремы из [5] - ограничились рассмотрением только блокировок, образующихся парами аминокислот. Были установлены все пары аминокислот, создающие блокировки для каждой из PC. Оказалось, что суммарное число пар аминокислот фигурирующих в Теореме свелось к 175 различающимся парам (рис.2 из [5]). Эти пары были разделены на две группы: пары аминокислот по которым возможна блокировка лишь одной PC, их всего 125. Оставшиеся 50 пар аминокислот могут быть использованы в блокировках от 2 до 4 PC, но, как показал анализ, одновременно могут блокировать не более двух PC. Оказалось также, что среднее значение число блокирующих пар

аминокислот около 9. Минимум – 0 аминокислот соответствует Trp (триптофан), как оказалось он не участвует в блокирующих парах ни как первая аминокислота, ни как вторая аминокислота в паре. Кроме того также три аминокислоты: Gln, Lys, Glu не участвуют в качестве первой аминокислоты в блокирующих парах, а 4 аминокислоты: Phe, Cys, Pro, Leu – в качестве второй аминокислоты в парах.

Множество из указанных 175 пар аминокислот было разбито на два подмножества. Первое из них соответствует т.н. неизбежным блокировкам это 31 пара аминокислот. Эти пары выделены на рис.2. из [5]. Можно видеть, что для этих пар могут блокироваться все PC от PC1 до PC5. Дадим пример такой блокировки (рис2). Так, для пары аминокислот ValLys имеют место блокировки для PC1, PC2, PC5. Для любого N: A, C, T, G имеет место блокировка TAA (при N=A), TGA в PC5 (при N=C), TGA в PC2 (при N=G) либо дважды TAA одновременно в PC1 и PC5 (при N=T). Эта ситуация существенно отличается, скажем, от пары ValAla, где блокировки возникают лишь для PC2 при N=A, а при трех других значениях N блокировки не возникают. Подобные пары отнесем к множеству возможных блокировок. Для поиска полного числа этого множества обратимся к блокировкам PC4, которые могут возникать лишь для некоторых кодировок ser либо leu. Эти блокировки также целесообразно свести к парам аминокислот. Это сведение привело к тому, что вместо 175 пар получим 210- различающихся пар аминокислот.

Тем самым в [5] получена двухкомпонентная интегральная характеристика стандартного генетического кода, в которой из полного числа 400 пар установлено минимальное число блокирующих пар – $q_{\min}$ и их максимальное число – $q_{\max}$. Имеем

$$q: q_{\min}=31, \quad q_{\max}=210 \quad (2)$$

Т.о. оказалось, что в процессе блокировки из 400 возможных пар могут участвовать не более чем 210, а оставшиеся 190 пар не имеют никакого отношения к блокировкам и кодировки в этих парах используются для решения некоторых других задач, которые нами были поставлены и исследованы [6]. Еще раз подчеркнем, что стандартный код устроен так, что независимо от кодировок в 31 паре аминокислот, которые соответствуют $q_{\min}$, неизбежно возникают блокировки. Или для гена, который образован из равновероятного набор пар аминокислот из всех 400 неизбежно блокируются $31/400 \sim 0.08$, или в среднем возникает (в каких-либо PC1 – PC5) по 8 кодонов ter на каждые 100 смысловых кодонов гена.

2. Постановка задачи

Наиболее важным было понять, как же на самом деле природа блокирует гены, используя интегральную характеристику q . В связи с этим, на основе Теоремы и ее следствий были проведены расчеты природных блокировок для 10 больших геномов (около 200000 генов), в т.ч. для всех известных на сегодня генов из генома человека.

Для математического анализа введем ряд характеристик. Назовем плотностью блокировки β отношение суммарного числа блокирующих PC1-PC5 кодонов терминации к числу смысловых кодонов в гене. Для стандартного кода можно вычислить два значения β - $\beta^{\circ\min}$, $\beta^{\circ\max}$ - экстремальные значения β при условии равновероятного представления в белковой последовательности каждой из 400 пар аминокислот. На основе [5] нетрудно видеть, что

$$\beta^{\circ\min}=0.0775, \beta^{\circ\max}=0.63. \quad (3)$$

Различие в этих значениях более чем в 8 раз, а разность составляет более 0.55. Отметим, что можно также показать, что при любом представлении пар аминокислот в белке имеет место неравенство

$$0 \leq \beta \leq 1 \quad (4)$$

гипотетические белковые последовательности с $\beta=0$ либо $\beta=1$ легко построить с учетом рис. 2 из [5].

Введем также другую характеристику - γ , как отношение числа кодонов терминации в PC0 (эта PC может иметь блокировки при условии, когда ген записывается на одной из трех PC- PC3, PC4, PC5), PC1, PC2 к подобному числу в PC3-PC5. Т.е. характеристика γ соответствует различию блокировки различных цепей ДНК, т.е. плюс- либо минус- цепей ДНК. При равновероятной блокировке этих цепей имеем $\gamma = 1$.

3. Исходные данные

Все данные по геномам взяты с сайта “National Center Biotechnology Information” - национального центра биотехнологической информации США [7]

Учитываем только по одной копии каждого гена. Следует отметить, что в геноме может содержаться несколько копий одного и того же гена. Например, гены хромосомы 2 и 4 содержатся в других хромосомах [8]. Последняя из публикаций по числу генных копий приводится в [9].

Для того, чтобы оценить число блокировок на целых геномах опубликованных в Интернете, а так же построить и сравнить распределения блокировок для различных геномов была написана программа для математической обработки генетической информации и вывода данных анализа. Подробное описание программы будет опубликовано отдельно.

4. Предварительный анализ расчетов.

В таблице 2 приводятся результаты расчетов для около 200 000 генов, для 10 больших геномов: собаки(*C.familiaris*), курицы(*G. gallus*), рыбы-зебры(*D. rerio*), мыши(*M. musculus*), крысы(*R. norvegicus*), медоносной пчелы(*A.mellifera*), дрозофилы(*D. melanogaster*), арабидопсиса(*A. thaliana*), червя(*C. elegans*), человека (*H. sapiens*).

В таблице 2 приводятся: $\bar{\beta}$ и $\max \beta$ - среднее и максимальное значение β соответственно, $\bar{\gamma}$ - среднее значение γ . При этом среднее и максимальное берется по всем генам в геноме. Приводится также значение $\Delta\beta$, которое вычислялось из условия, что интервалу разброса от среднего значения ($\bar{\beta} - \Delta\beta, \bar{\beta} + \Delta\beta$) принадлежит не менее 90% полного числа генов из соответствующего генома. Кроме того, дается процент содержания каждого кодона *ter* из (1) в альтернативных РС.

В таблице 2 организмы располагаются в порядке убывания характеристики $\bar{\beta}$, которая изменяется от 0.17 (*D. melanogaster*) до 0.30 (*A. mellifera*). Этот диапазон принадлежит насекомым. Отметим, что насекомые – самая многочисленная и разнообразная группа животных на Земле. При таком диапазоне изменения имеем: размер геномов изменяется в 33 раза (от 0.1 Gbp для *C. elegans* и *A. thaliana* до 3.3 Gbp для *M. musculus*). Кодированная область изменяется в 100 раз (от 0.3% для *D. rerio* до 33% *A. thaliana*), а число генов в 8 раз (от 4282 для *D. rerio* до 33461 для *C. familiaris*).

Проанализируем полученные результаты последовательно по пяти характеристикам на основании таблицы 2 и рис 3-5.

1. Вновь обратимся к характеристике $\bar{\beta}$. Видим, что $\bar{\beta}$ для шести геномов из таблицы 2 почти одинаково: $\bar{\beta} = 0.2 \pm 0.01$. Или среднее число блокирующих кодонов около 2-х на 10 смысловых кодонов в гене или 1 блокировка на 5 кодонов. Отклонение указанной величины наблюдается у четырех оставшихся геномов: для *D. melanogaster* имеем $\bar{\beta} = 0.17$, для *C. elegans* – 0.24, для *A. thaliana* – 0.27 и наконец для *A. mellifera* – 0.3. Последнее значение соответствует максимуму $\bar{\beta}$ среди всех 10 геномов. Отметим, что генетики считают *A. mellifera* вершиной эволюции среди насекомых [10]. Для другого насекомого – классического объекта генетики

D. melanogaster $\bar{\beta} = 0.17$ или почти в 1.8 раза меньше.

На Рис 4. показаны распределения блокировок для 10 геномов. Слева на право (сплошным) $\beta_{\min}$, β , $\beta_{\max}$ в геномах и их (пунктиром) случайных аналогах (каждая аминокислота кодируется случайным кодоном синонимом). Здесь $\beta_{\min}$ ($\beta_{\max}$) – это распределение минимальных(максимальных) значений β в генах, которое достигается путем варьирования кодонов благодаря вырожденности генетического кода при сохранении всех аминокислот всех белков. Горизонтальная ось показывает относительное число блокировок $\beta * 100$ в генах, вертикальная – число генов. Зеленая и красная пунктирные вертикальные линии - экстремальные значения $\beta^{\circ\min} = 0.0775$ и $\beta^{\circ\max} = 0.63$ соответственно [5]. Наименьшее отклонение в распределении β от распределения случайного аналога наблюдается у *A. thaliana*. Сдвиг случайного распределения на $\Delta\beta = -0.02$ приводит к совпадению с реальным распределением. Случай с *C. Elegans* - исключительный. Пик случайного распределения совпадает с реальным. Во всех остальных случаях он смещен, причем в сторону увеличения β , за исключением *A.mellifera*, где пик смещен в сторону уменьшения. Пик распределения $\beta_{\max}$ всюду смещен влево и лишь случае *A. Thaliana* совпадает со значением $\beta^{\circ\max} = 0.63$. Пик распределения $\beta_{\min}$ как правило смещен в разные стороны от $\beta^{\circ\min} = 0.0775$ и совпадает с ним лишь для *M. Musculus* и *C.familiaris*. Пик реального распределения максимален для *A.mellifera* – 0.37 и минимален для *H. sapiens* – 0.15.

2. *Характеристика $\max \beta$* . Ни в одном из геномов $\max \beta$ не для какого либо гена не достигает теоретического максимума $\beta^{\circ\max}=0.63$. Наиболее приближенное к $\beta^{\circ\max}$ значение $\max \beta$ у *M. musculus* – 0.60, наименьшее у *D. rerio* – 0.42.

3. *Характеристика $\Delta\beta$* . Интервал $(\bar{\beta} - \Delta\beta, \bar{\beta} + \Delta\beta)$ содержит 90% генов и $\Delta\beta$ имеет наименьшее 0.07 значение для трех геномов *D. melanogaster*, *A. thaliana*, *D. rerio*, 0.08 для *C. elegans* и наибольшее 0.15 для

A. mellifera. Для пяти оставшихся геномов $\Delta\beta = 0.1 \pm 0.01$.

Характеристику $\Delta\beta$ нетрудно представить на Рис. 3. Здесь для каждого из 10 геномов показано соответствие “ген – величина β ”, дающее наглядное представление о блокировках в геноме в целом. Гены на графиках представлены по возрастанию длины. Налицо отличие в “разбросе” числа блокировок у организмов. Если же рассматривать конкретный график, то видно, что с увеличением длины у генов “разброс” уменьшается всюду, кроме *H. sapiens*.

Наибольшее уменьшение $\Delta\beta$ наблюдается у *A. thaliana*. Затемнение на рисунках для генов большой длины связано, видимо, с тем что число генов для различных длин примерно одинаково. Также, нельзя не заметить повышенное содержание блокировок в генах *M. Musculus* и *R. norvegicus*, имеющих длину в интервале $(942-25, 942+25)$ нуклеотидов(скачки на графиках). Число таких генов 1700 и у *M. Musculus* и у *R. norvegicus*.

4. *Характеристика $\bar{\gamma}$* . Наиболее приближенная к значению $\bar{\gamma} = 1$, что соответствует равновероятной блокировке плюс-цепи и минус-цепи ДНК наблюдается у *A. mellifera*: $\bar{\gamma} = 0.98$, наименьшая – у *D. melanogaster*: 0.82.

На рис.5 показано отношение числа блокировок в плюс-цепи к числу блокировок в минус-цепи в разрезе генов, а так же $\bar{\gamma}$ для соответствующего генома. Для всех геномов наблюдаются скачки по γ . Максимальное значение равно $\gamma = 25$ наблюдается у *R. Norvegicus* и *C. familiaris*. Это пики лишь для отдельных генов и в основном наблюдаются для генов малой длины.

5. *Встречаемость каждого из трех кодонов *ter* из (1)*. Для 9-ти геномов из 10-ти наблюдается примерно одинаковый процент встречаемости TGA, что составляет $54 \pm 6\%$. Получается, что блокировка TGA встречается столь же часто в процентном соотношении как и сумма

ТАА и TAG. Исключение и на этот раз - *A.mellifera*: практически равновероятно участвуют ТАА(40%) и TGA(38%), и почти вдвое меньше TAG(22%).

Итак, по четырем характеристикам: $\bar{\beta}$, $\Delta\beta$, $\bar{\gamma}$, а также встречаемость γ оказалось, что для *A.mellifera* наблюдается отличие по этим характеристикам от всех девяти остальных геномов. Биологический анализ полученных результатов будет опубликован.

Литература

1. Watston J.D., Crick, F.H.C. A structure for Deoxyribose Nucleic Acid. Nature, 171, 737-738, 1953.
2. Heidi Ledford. Kudos, not cash, is the X-factor. Nature 19 oct 2006 ,vol 443, p.733.
3. Льюин Б. Гены. М.: Мир, 1987. 544с.
4. Козлов Н.Н. Математический анализ генетических кодов. Математическая биология и биоинформатика. 2005. Т. 1. № 1. С. 70-96, [http://www.matbio.org/downloads_en/Kozlov2006\(1_70\).pdf](http://www.matbio.org/downloads_en/Kozlov2006(1_70).pdf)
5. Козлов. Математический анализ структурных генов. ДАН (Математика). 2007. Т. 412. № 5. (в печати).
6. Козлов Н.Н. , Бахарев М.Г. Математический анализ одной биологической структуры. Препринт ИПМ им М.В.Келдыша РАН ,2006, 28 с.
7. LaDeana W.Hiller “Generation and annotation of the DNA sequences of human chromosomes 2 and 4” Nature 7 April 2005, vol 434, p 724-731.
8. Национальный центр биотехнологической информации США: <ftp://ftp.ncbi.nih.gov>
9. Richard Redon et al. Global variation in copy number in the human genome. Nature, 23 nov 2006 ,vol 444, p.444-454.
10. Ичас М. О природе живого: механизмы и смысл. М.: Мир, 1994. 496с.

PC0 (0) MetSerIleLysLeuSerTyrArgGluSerPheSerIleLeuGluGluValTyrGluGlyGlyArgValHisAlaPro *

PC1 (-1) TyrGluHis * Thr * Leu * ArgValIle * TyrIleArgGlyGlyLeu * ArgArgAlaArgSerArgAlaIle

PC2 (+1) * AlaLeuAsnLeuValIleGluSerHisLeuValTyr * ArgArgPheMetLysGluGluAlaPheThrArgHis

→ AGCTAGCTATGAGCATTAAACTTAGTTATAGAGAGTCATTAGTATATTAGAGGAGGTTTATGAGGAGGGCGCGTTCACGCGCCATAA

← TCGATCGATACTCGTAATTTGAATCAATATCTCTCAGTAAATCATATAATCTCCTCCAATACTTCCTCCCGCGCAAGTGCGCGGTATT

PC3 (-1) IleLeuMetLeuSerLeu * LeuSerAspAsnLeuIleAsnSerSerThr * SerProProArgThr * AlaGlyTyr

PC4 (0) HisAlaAsnPheLysThrIleSerLeu * LysThrTyr * LeuLeuAsnIlePheSerProAlaAsnValArgTrpLeu

PC5 (+1) SerCys * Val * AsnTyrLauThrMet * TyrIleLeuProProLysHisLeuLeuAlaArgGluArgAlaMet

13

Рис. 1. Блокировки и рамки считывания.

Пять альтернативных рамок считывания (PC1 – PC5) для гена PC0 из плюс - цепи ДНК. В скобках указаны сдвиги (в нуклеотидах) относительно PC0. Кодоны блокировки * : TAA, TAG, TGA выделены. Стрелками указывается направление чтения генетических текстов в плюс - цепи (→) и минус - цепи (←) ДНК.

Если ген записан на минус – цепи ДНК, например, на PC3, то PC0 становится блокированной наряду с другими PC, кроме PC3.

Таблица 1.

Стандартный генетический код.

	1		2	3
1	Met	M	1	ATG
2	Trp	W	1	TGG
3	Phe	F	2	TTY
4	Tyr	Y	2	TAY
5	His	H	2	CAY
6	Asn	N	2	AA $\bar{Y}$
7	Asp	D	2	GA $\bar{Y}$
8	Cys	C	2	TG $\bar{Y}$
9	Gln	Q	2	CA $\bar{X}$
10	Lys	K	2	AA $\bar{X}$
11	Glu	E	2	GA $\bar{X}$
12	Ile	I	3	AT $\bar{Y}$, AT $\bar{A}$
13	Val	V	4	GT $\bar{N}$
14	Pro	P	4	CC $\bar{N}$
15	Thr	T	4	AC $\bar{N}$
16	Ala	A	4	GC $\bar{N}$
17	Gly	G	4	GG $\bar{N}$
18	Ser	S	6	TC $\bar{N}$, AG $\bar{Y}$
19	Leu	L	6	CT $\bar{N}$, TT $\bar{X}$
20	Arg	R	6	CG $\bar{N}$, AG $\bar{X}$
	ter	*	3	TAX, TGA

Примечание. Для каждой из 20-и аминокислот приводятся:

1 – стандартные трехбуквенные и однобуквенные сокращения названий аминокислот, 2 – число кодонов, 3 – трехбуквенные представления кодонов. Обозначение: X: A, G; Y: T, C; N: A, G, T, C. В последней строке приводятся три терминаторных кодона – ter, каждый из которых обозначает останов синтеза белка.

Таблица 2. Исходные данные и результаты расчета для 10 больших геномов

Исходные данные:

Размер геномов указан в миллиардах нуклеотидов (1 Gbp = 10^9 пар нуклеотидов).

Размер кодирующей области: из данного генома берется только по одной копии каждого известного гена (копий гена в геноме может быть несколько [9]). Суммарный размер такой области и составляет указанный процент по отношению к размеру всего генома.

Результаты расчета:

β -плотность блокировки или отношение числа кодонов ter в 5-и альтернативных рамках считывания к числу смысловых кодонов в гене. $\bar{\beta}$ – среднее по β по всем генам в геноме.

Max β – максимальное β по всем генам в геноме.

γ - отношение остановов в плюс-цепи к остановам в минус-цепи в гене. $\bar{\gamma}$ - среднее по γ по всем генам в геноме.

TAA(%), TAG(%), TGA(%) – количество каждого из кодонов ter: TAA, TAG, TGA в геноме в процентах.

Исходные данные				Результаты расчета						
Организм	Размер генома	Кодирующая область	Число генов	$\bar{\beta}$	Max β	$\Delta\beta$	$\bar{\gamma}$	TAA(%)	TAG(%)	TGA(%)
A.mellifera	0,27Gbp	4%	5588	0.30	0.53	0.15	0.98	40	22	38
A. thaliana	0,1Gbp	33%	26525	0.27	0.57	0.07	0.97	29	23	48
C. elegans	0.1Gbp	25%	20050	0.24	0.58	0.08	0.84	27	18	55
M. musculus	3,3 Gbp	1%	25319	0.21	0.60	0.09	0.93	21	23	56
R. norvegicus	3,1 Gbp	1%	21612	0.21	0.57	0.09	0.92	21	23	56
D. rerio	1,7Gbp	0.3%	4282	0.21	0.42	0.07	0.90	22	18	60
C. familiaris	2.4 Gbp	2%	33461	0.20	0.45	0.11	0.95	23	22	55
G. gallus	1.2 Gbp	2%	14776	0.20	0.55	0.1	0.96	24	21	55
H. sapiens	3,2 Gbp	1%	25613	0.19	0.5	0.11	0.92	23	21	56
D. melanogaster	0,17Gbp	13%	13651	0.17	0.51	0.07	0.82	25	24	51

3.а. Первый тип блокировок:

	V	a	l	L	y	s	
	G	T	N	A	A	X	
N: A	G	T	A	A	A	X	
		t	e	r			→
N: C	G	T	C	A	A	X	
		A	G	T			←
		t	e	r			
N: G	G	T	G	A	A	X	
		t	e	r			→
N: T	G	T	T	A	A	X	→ T A A (ter)
	C	A	A	T	T	Y	← T A A (ter)

3.б. Второй тип блокировок:

	V	a	l	A	l	a	
	G	T	N	G	C	N	
N: A	G	T	A	G	C	N	
		t	e	r			→
N: C	G	T	C	G	C	N	
	C	A	G	C	G	N	
N: G	G	T	G	G	C	N	
	C	A	C	C	G	N	
N: T	G	T	T	G	C	N	
	C	A	A	C	G	N	

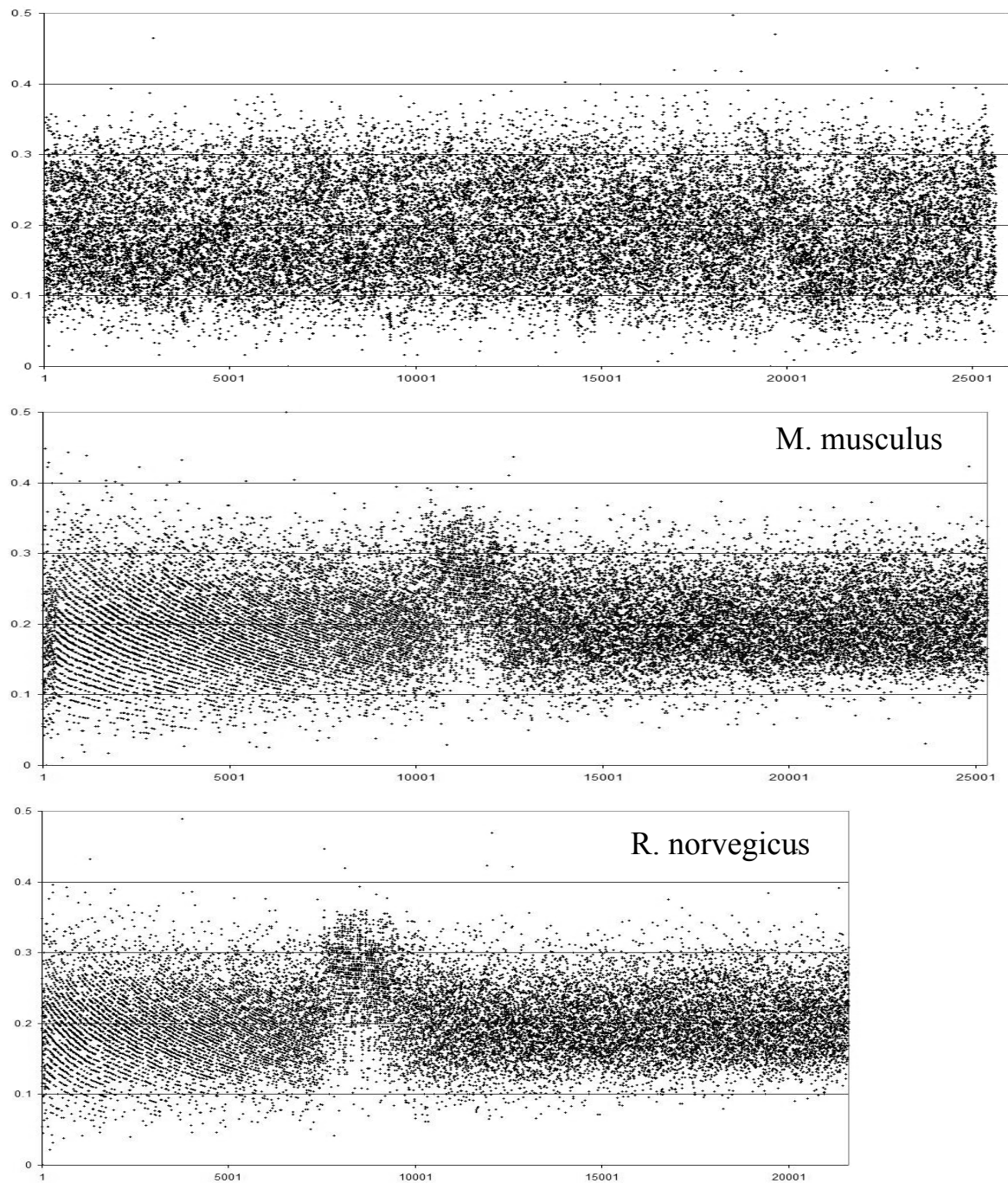
« → » означает чтение по плюс-цепи ДНК.

« ← » означает чтение по минус-цепи ДНК.

Рис. 2. Типы блокировок.

2.а. Блокировка возникает независимо от выбора нуклеотидов. Такие типы блокировок встречаются у генов с маленьким числом относительных остановов.

2.б. Блокировка возникает не всегда, а в зависимости от выбранного природой нуклеотида.

β 

номер гена

Рис 3.

Относительные блокировки в геномах.
 Гены указаны в порядке возрастания их длины.
 Случай геномов *H. Sapiens*, *M. musculus*, *R. norvegicus*.

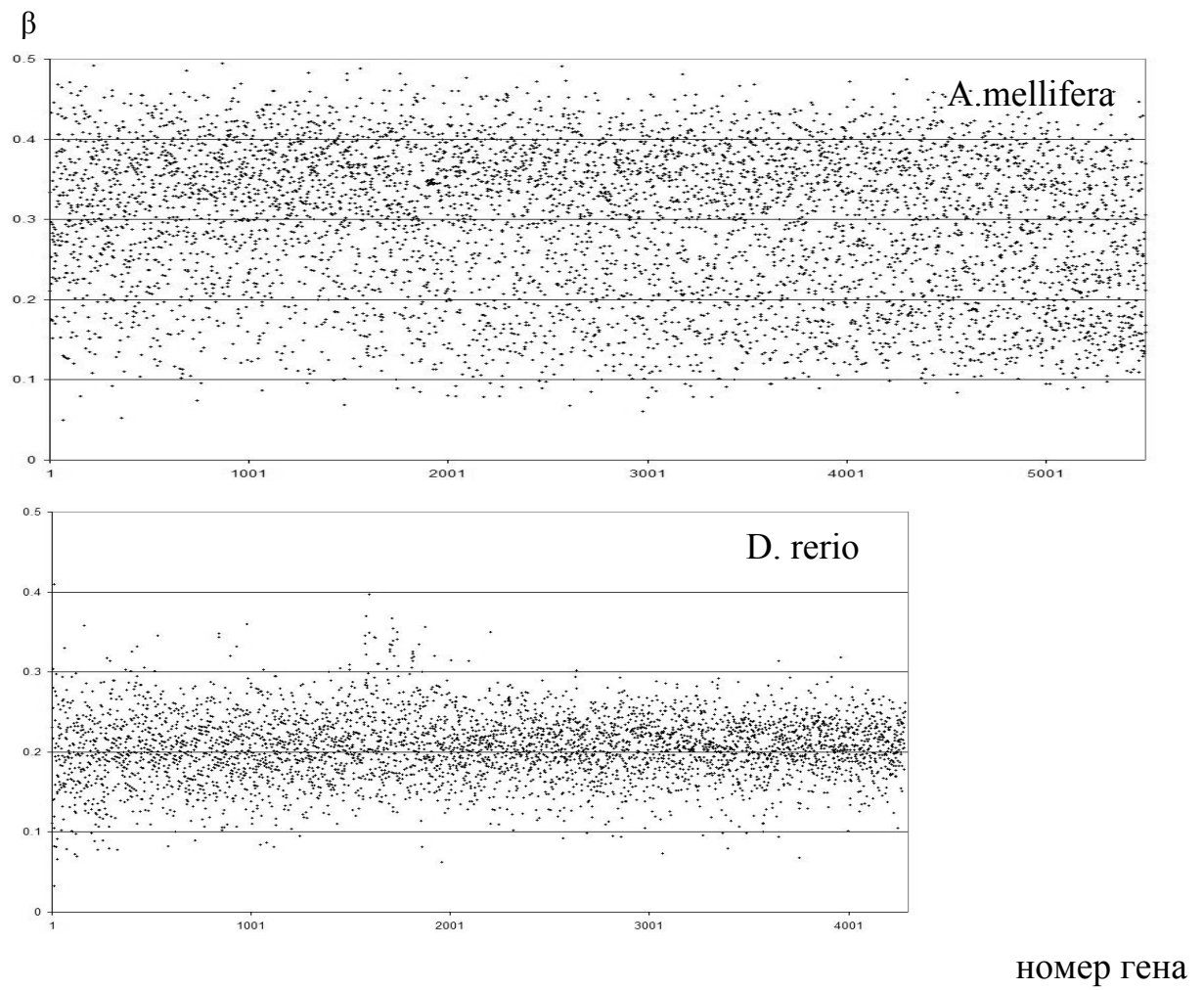


Рис 3. (продолжение)

Случай геномов *A. mellifera*, *D. rerio*.

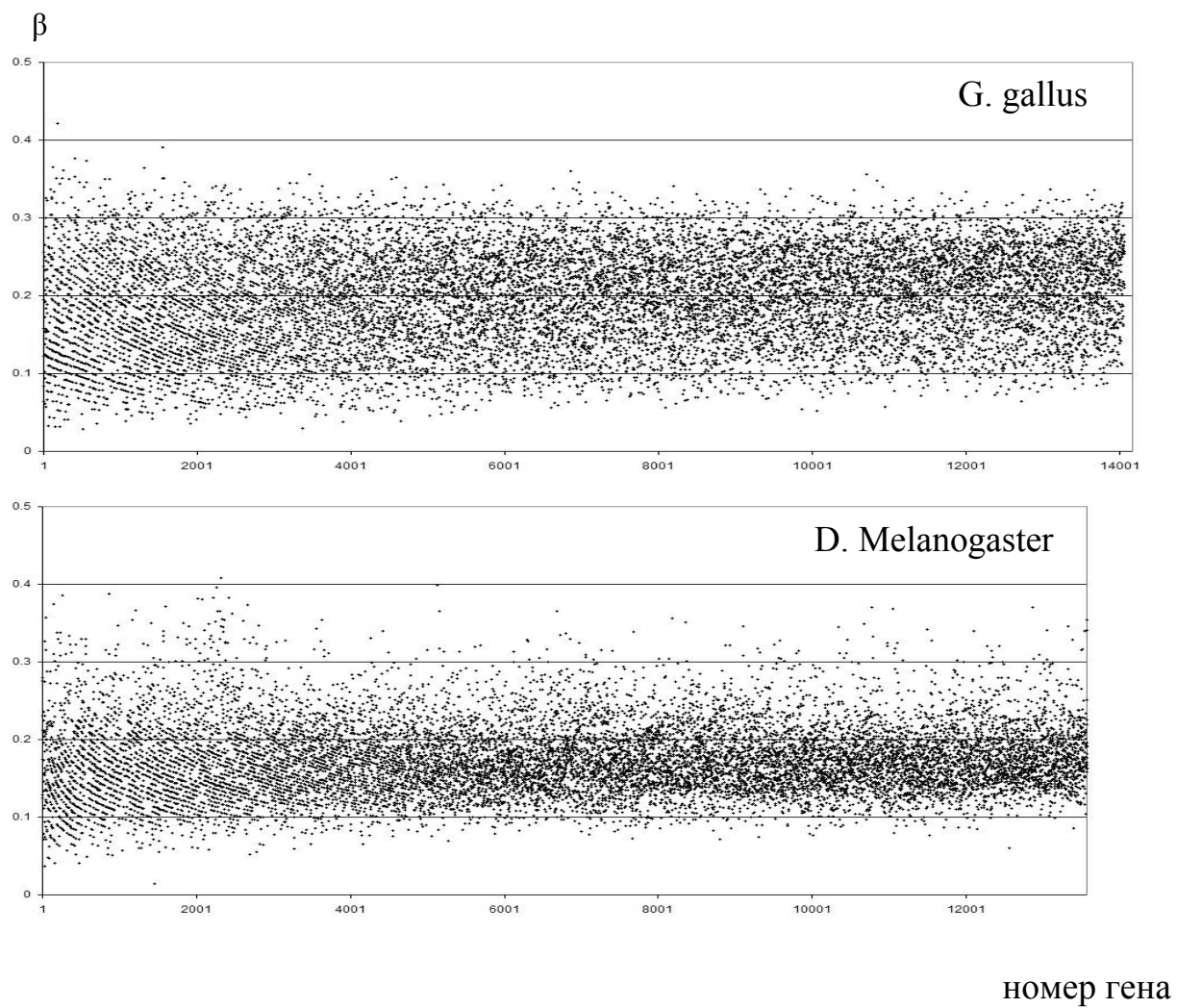


Рис 3.(продолжение)

Случай геномов *G. gallus*, *D. Melanogaster*.

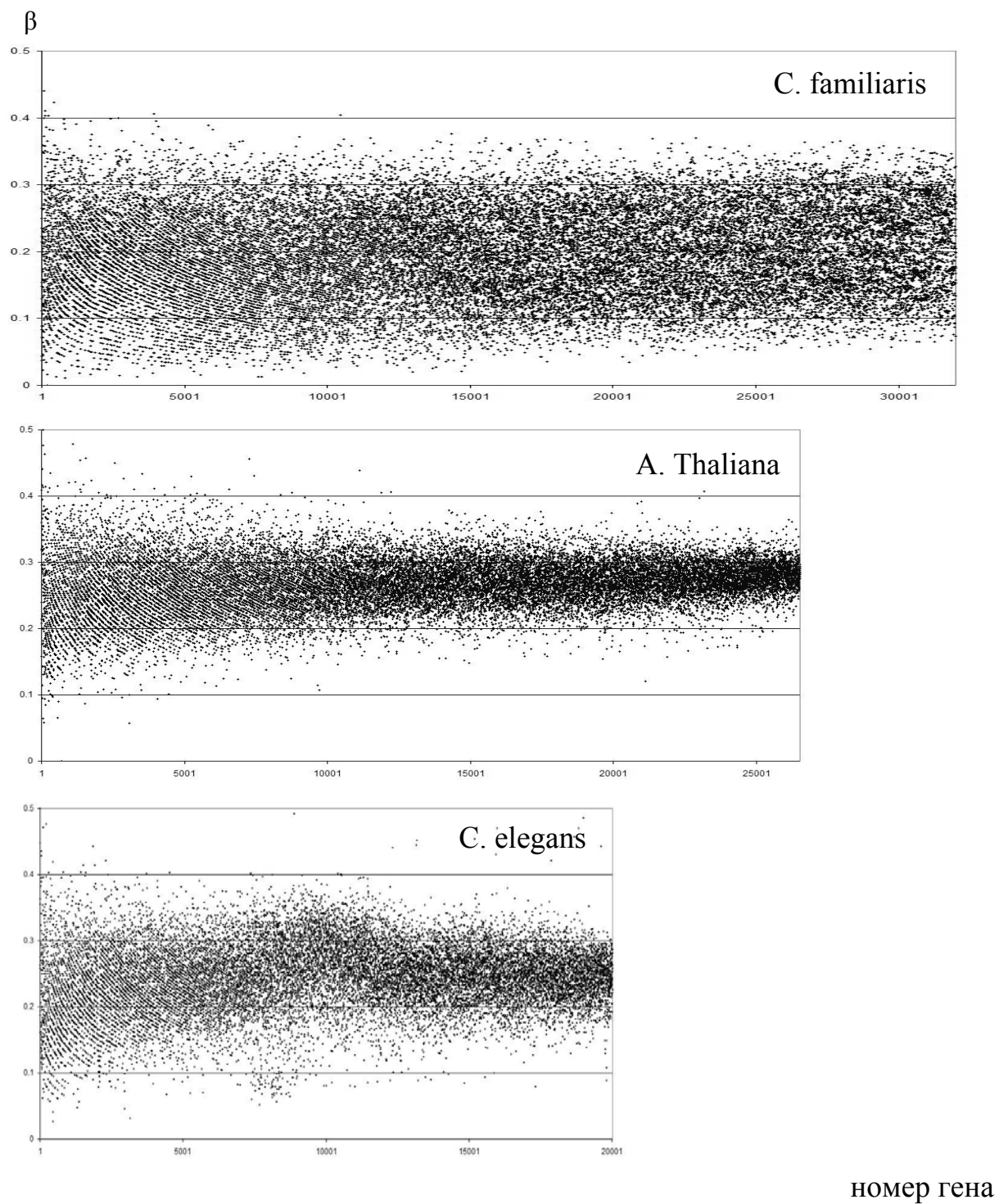
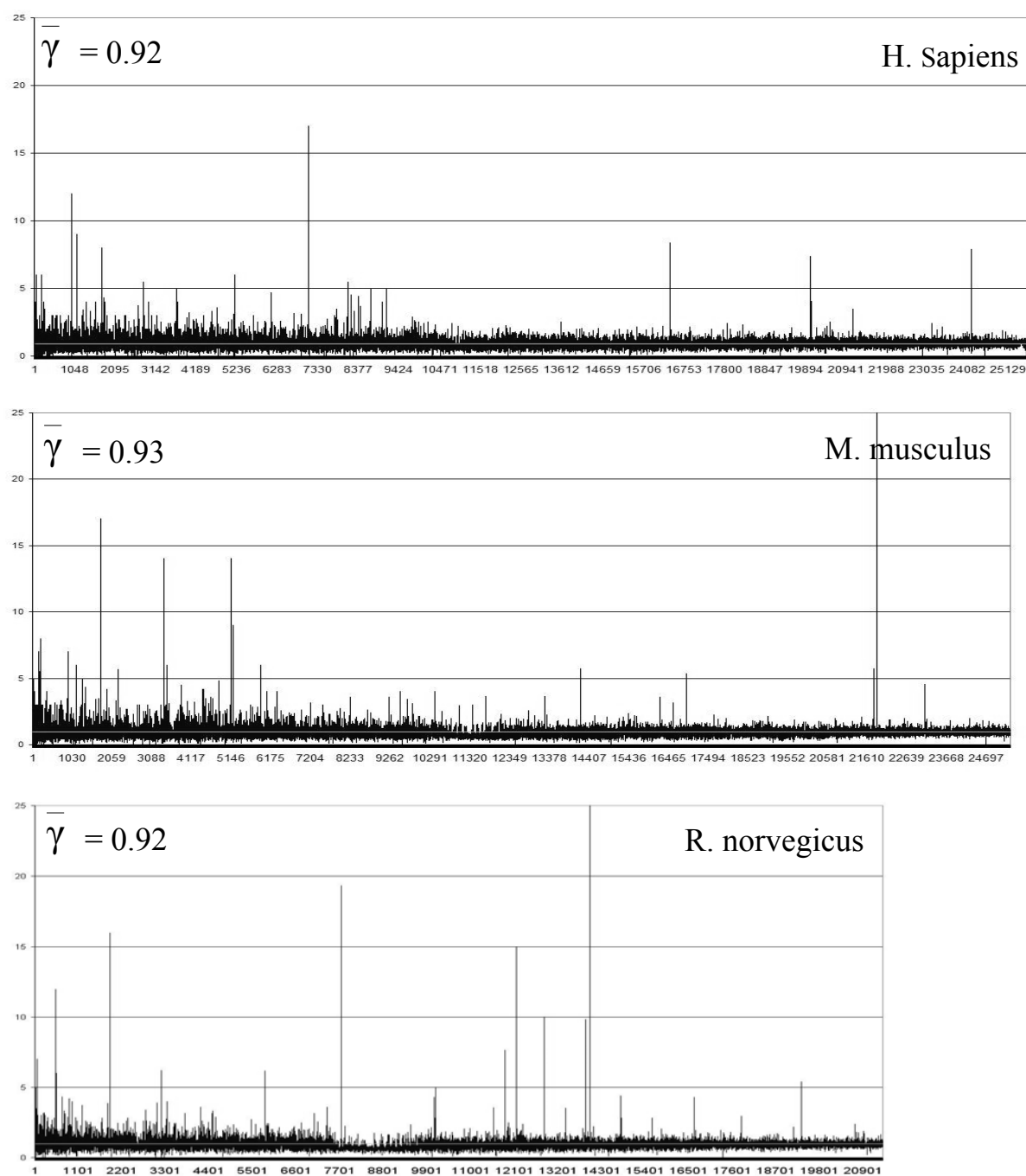


Рис 3. (продолжение)

Случай геномов *C. familiaris*, *A. Thaliana*, *C. elegans*.

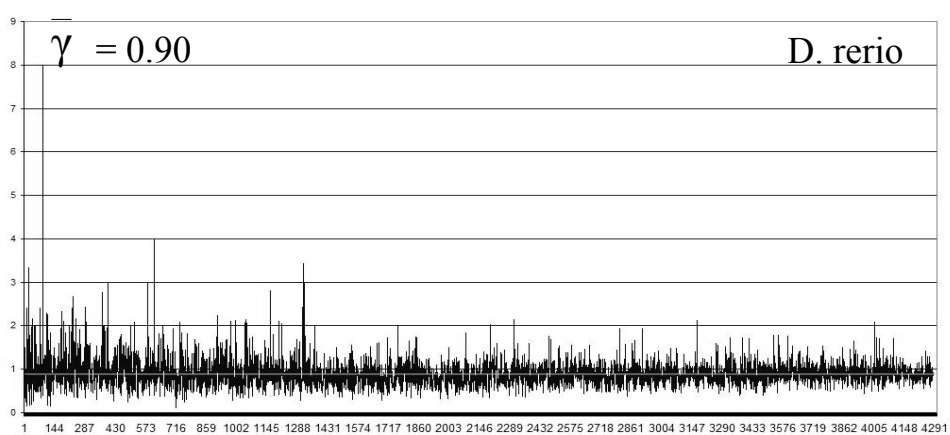
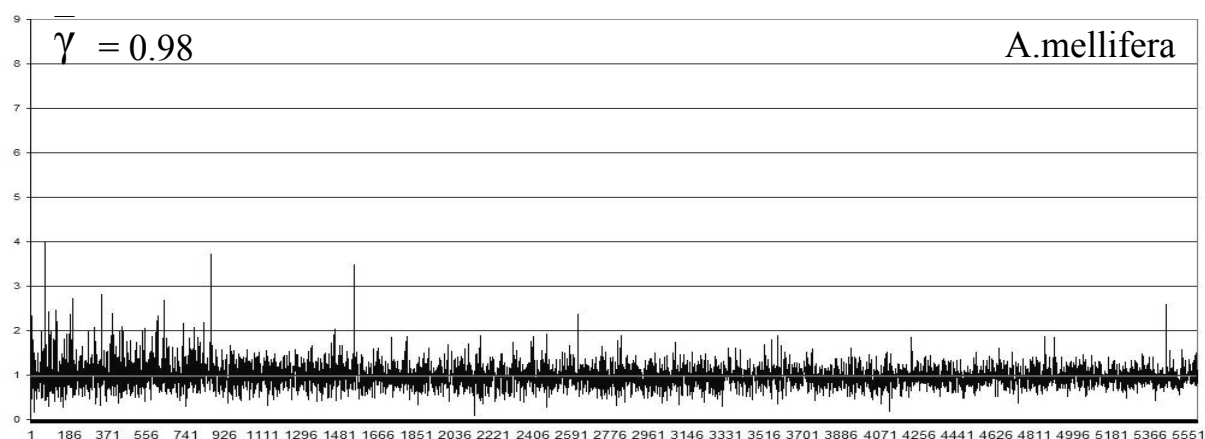
γ 

Число генов

Рис 5.

Отношение числа блокировок в плюс - цепи к числу блокировок в минус - цепи в разрезе белков. Гены указаны в порядке возрастания их длины.

Случай геномов *H. Sapiens*, *M. musculus*, *R. norvegicus*.

γ 

Число генов

Рис 5.(продолжение)

Случай геномов A. mellifera, D. rerio.

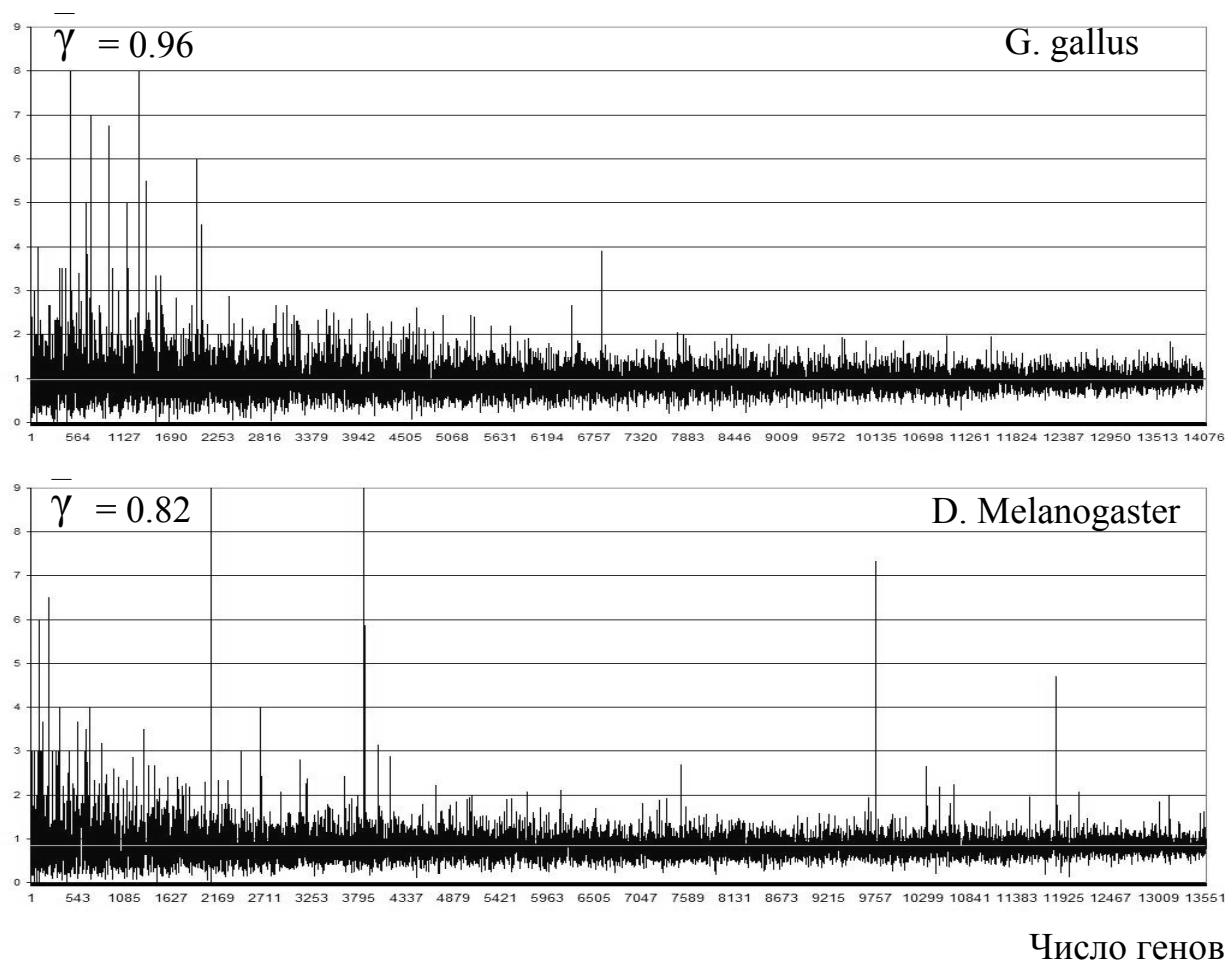
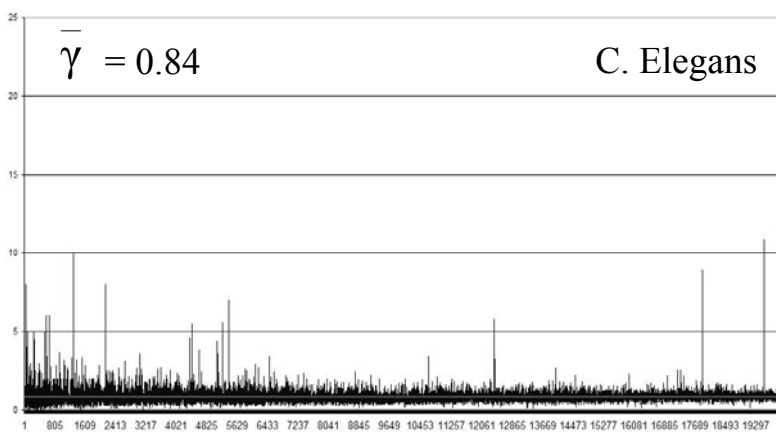
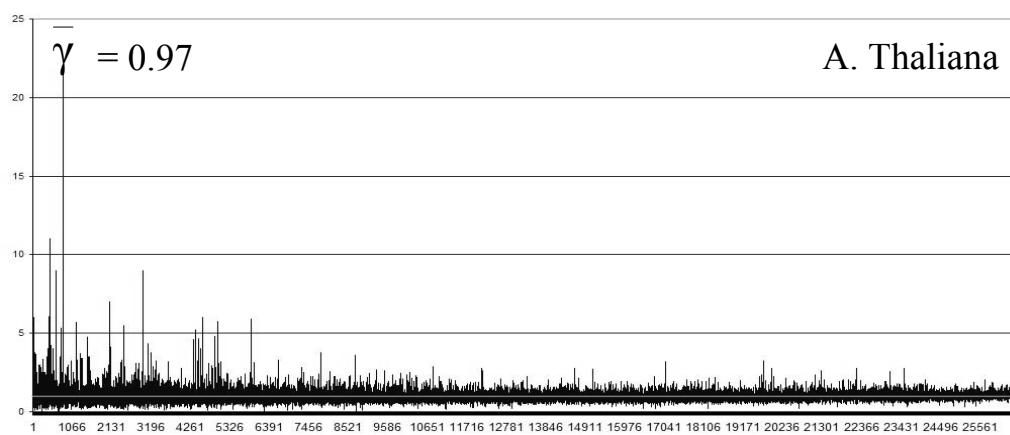
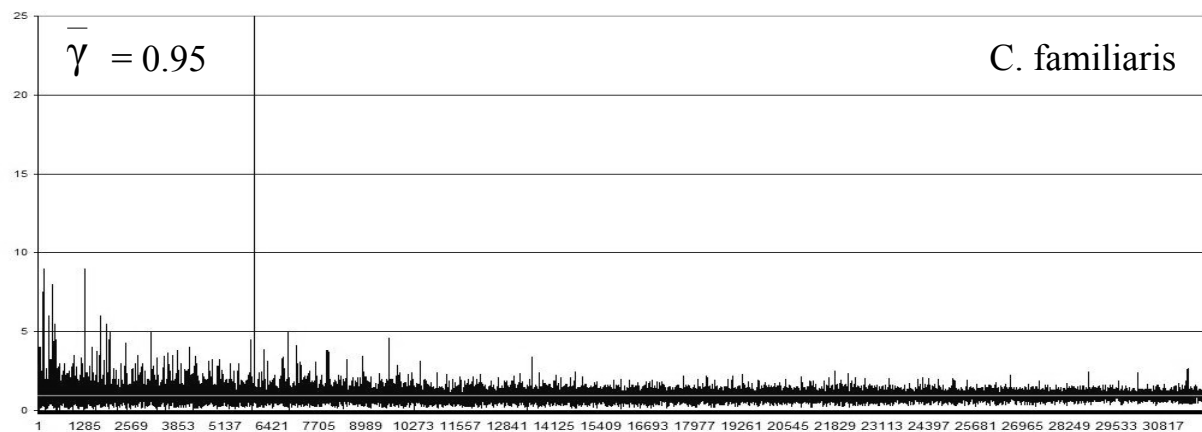
γ 

Рис 5. (продолжение)

Случай геномов *G. gallus*, *D. Melanogaster*.

γ 

Число генов

Рис 5.(продолжение)
Случай геномов *C. familiaris*, *A. Thaliana*, *C. Elegans*.

Число генов

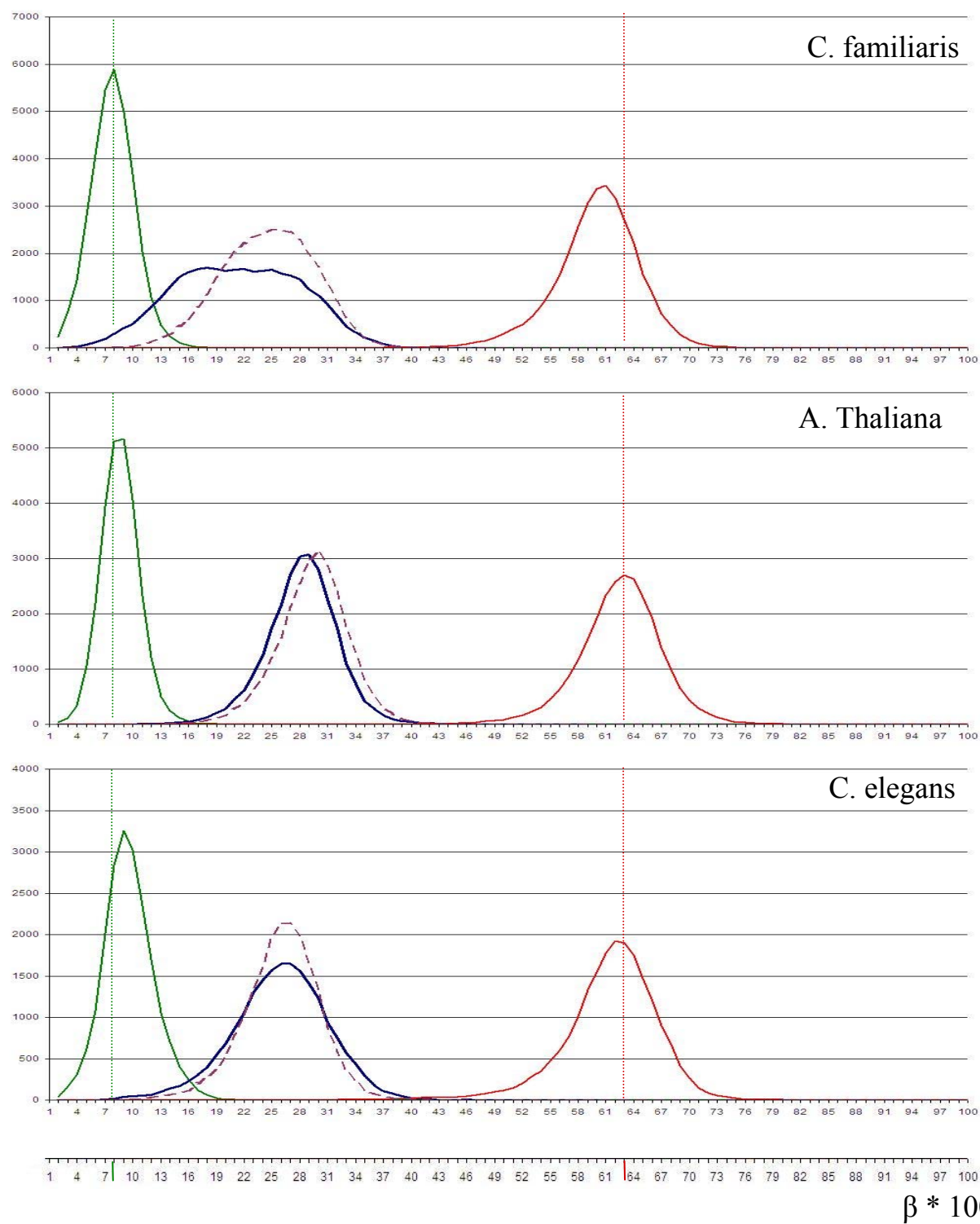


Рис 4. (продолжение)

Случай геномов *C. familiaris*, *A. Thaliana*, *C. elegans*.

Число генов

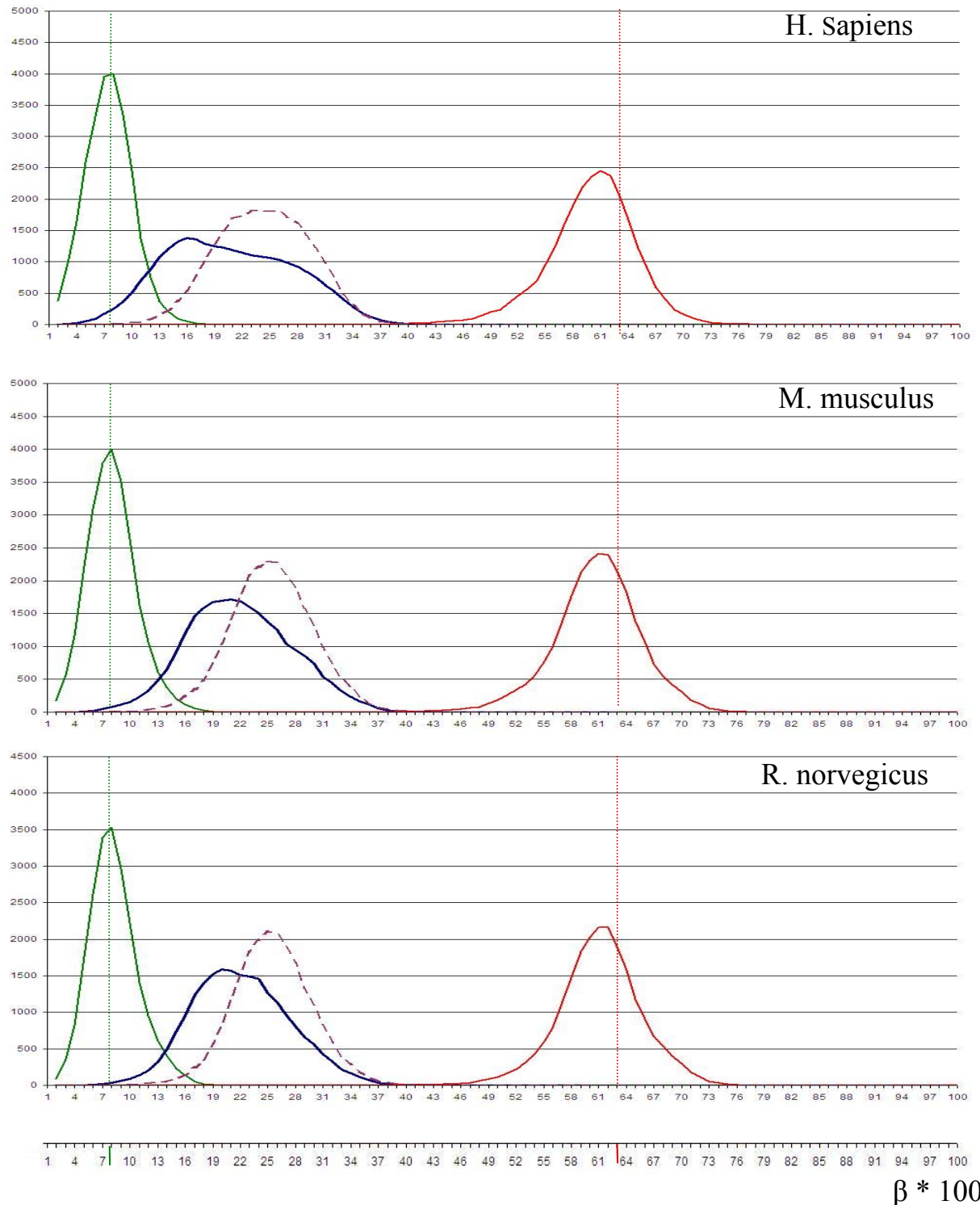


Рис 4. Распределение блокировок.

Слева на право (сплошным) β_{min} , β , β_{max} в геномах и их (пунктиром) случайных аналогов (каждая АК кодируется случайным кодоном синонимом). Горизонтальная ось показывает относительное число блокировок $\beta * 100$ в генах, вертикальная – число генов. Зеленая и красная пунктирные вертикальные линии - экстремальные значения $\beta^{\circ min}=0.0775$ и $\beta^{\circ max}=0.63$ соответственно [5].

Случай геномов *H. Sapiens*, *M. musculus*, *R. norvegicus*.

Число генов

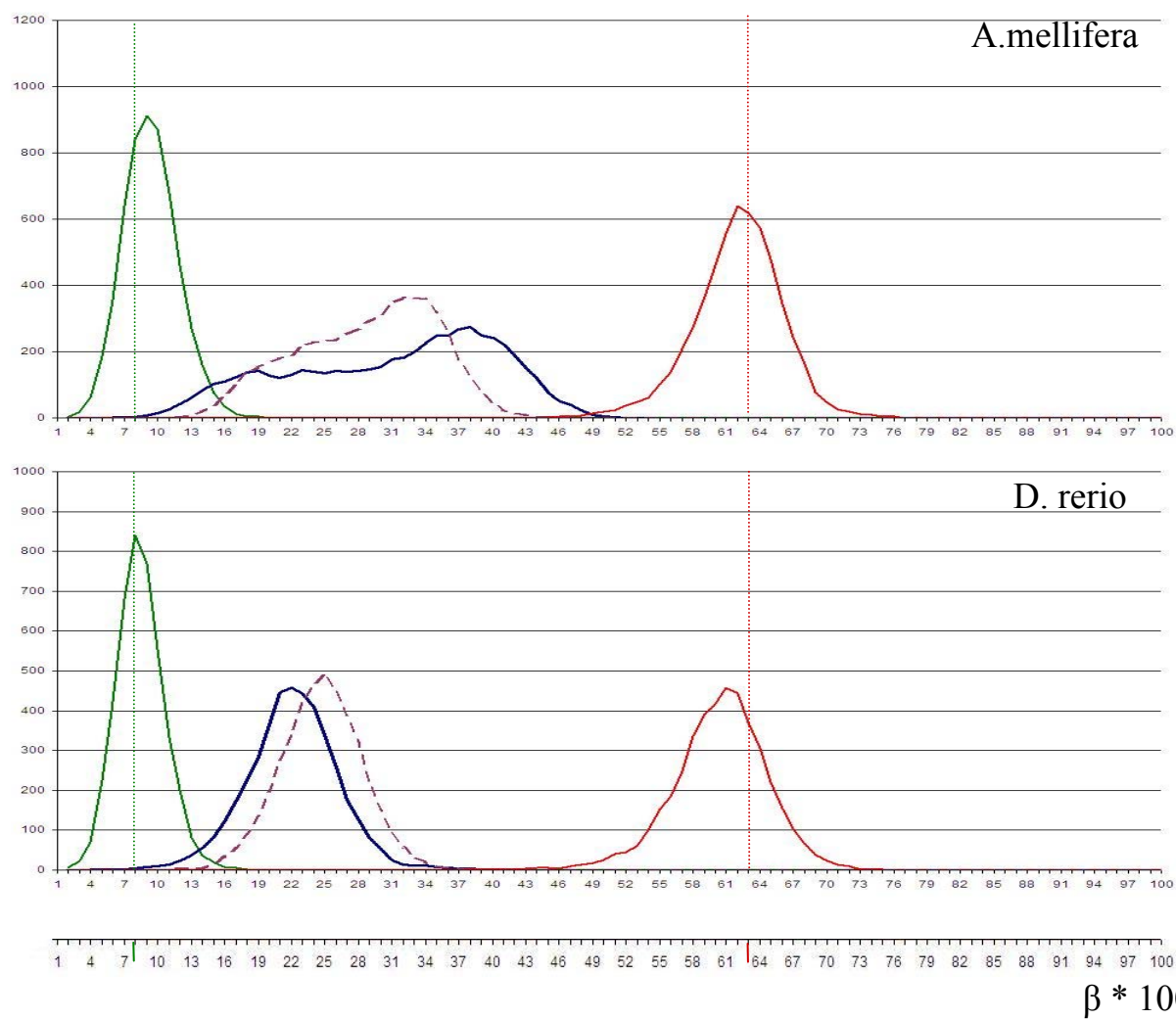


Рис 4.(продолжение)

Случай геномов *A. mellifera*, *D. rerio*.

Число генов

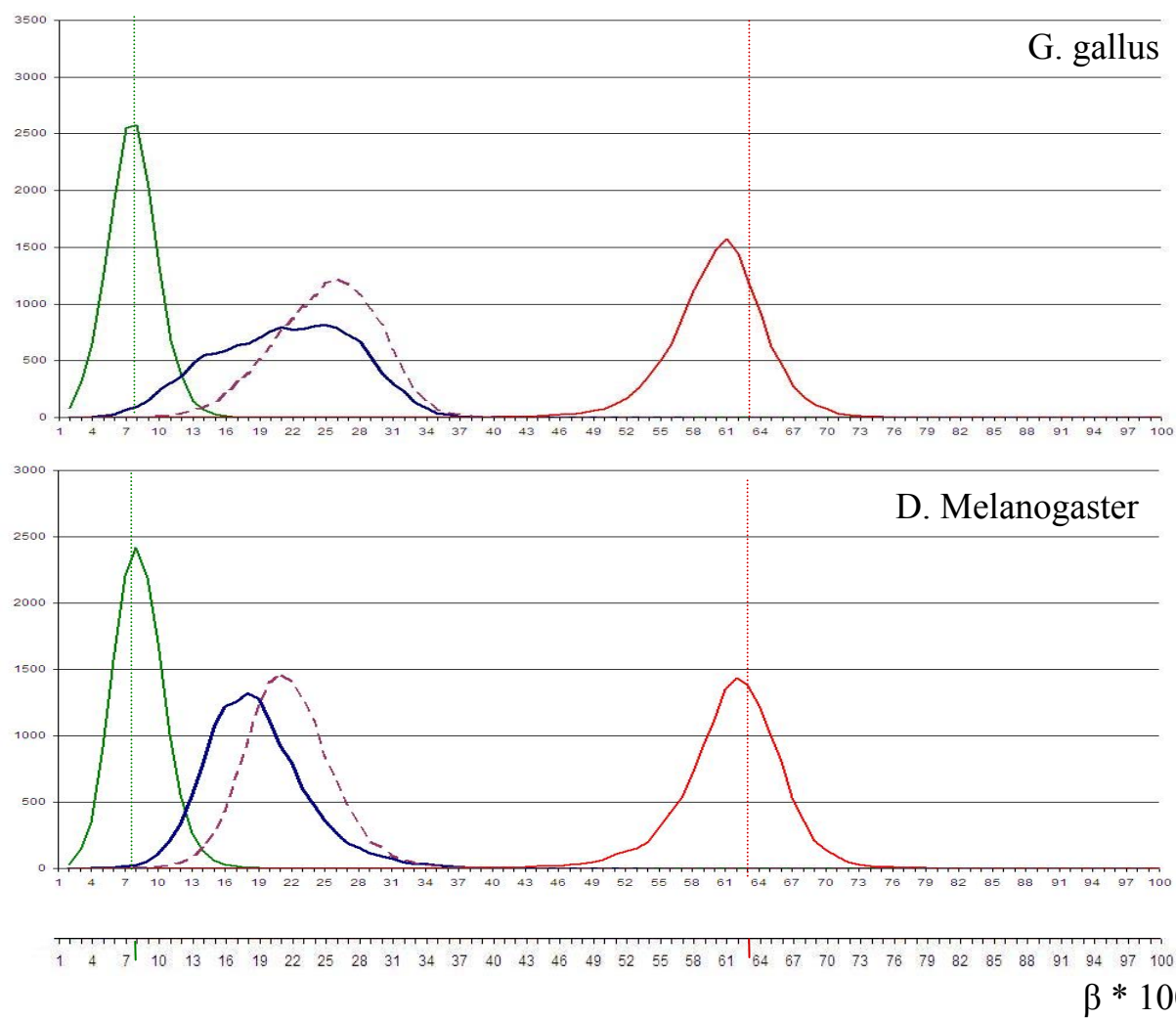


Рис 4.(продолжение)

Случай геномов *G. gallus*, *D. Melanogaster*.