БИООРГАНИЧЕСКАЯ ХИМИЯ, 2004, том 30, № 6, с. 621–625

УДК 577.214.(337+622)

НОВЫЕ ГЕНЫ НА ХРОМОСОМЕ 7 ЧЕЛОВЕКА: БИОИНФОРМАЦИОННЫЙ АНАЛИЗ КЛАСТЕРА ГЕНОВ ИЗ СЕМЕЙСТВА *POLR2J*

© 2004 г. Д. Г. Шпаковский, Е. К. Шематорова, Г. В. Шпаковский[#]

Институт биоорганической химии им. акад. М.М. Шемякина и Ю.А. Овчинникова РАН, 117997 ГСП, Москва, В-437, ул. Миклухо-Маклая, 16/10

Поступила в редакцию 24.02.2004 г. Принята к печати 13.04.2004 г.

В составе хромосомы 7 человека выявлены четыре независимых гена, кодирующих различные варианты субъединицы hRPB11 РНК-полимеразы II *Homo sapiens*. Три гена (*POLR2J1*, *POLR2J2*, *POLR2J3*) расположены в виде единого кластера суммарной протяженностью 214530 п. о. в генетической области 7q22.1 на длинном плече хромосомы (в составе контига NT_007933); еще один ген, *POLR2J4* (31040 п. о.), обнаружен в цитогенетическом локусе 7p13 короткого плеча хромосомы 7 (контиг NT_007819). Проведенный анализ позволил также уточнить имевшиеся разнородные экспериментальные данные о картировании гена субъединицы hRPB11 на хромосоме 7, в частности объяснить наличие трех мест его локализации, согласно данным гибридизации с флуоресцентномечеными специфическими зондами (метод FISH). Установлено, что при экспрессии четырех генов *POLR2J* человека могут синтезироваться по крайней мере 14 видов зрелых, кодирующих слегка различные изоформы hRPB11, мРНК, 11 из которых обнаружены нами (в виде полноразмерных копий или четко соотносимых фрагментов) в доступных базах EST (expressed sequence tags) и кДНК. Предложена и обоснована наиболее вероятная схема происхождения множественных генов семейства *POLR2J* в результате трех увеличивающихся в размерах дупликаций, позволяющая сделать ряд интересных наблюдений о путях эволюции отдельных человеческих генов и о механизмах генерирования белкового разнообразия у высших эукариот.

Ключевые слова: Homo sapiens, хромосома 7; ядерная РНК-полимераза II; субъединица hRPB11; экзоны; дупликации; эволюция генов и белков.

ВВЕДЕНИЕ

Одним из эпохальных достижений науки начала XXI в. стало установление первичной структуры генома человека [1]. Параллельно с появлением в 2000–2001 годах первоначального, еще очень несовершенного "чертежа" (draft) генома [2, 3], стали появляться статьи, описывающие высокоточные, почти полные первичные структуры пяти хромосом *Homo sapiens* (№ 22, 21, 20, 14 и 7) [4–9]. Хромосома 7 выделяется в этом списке по целому ряду параметров: во-первых, это первая секвенированная метацентрическая хромосома; во-вторых, она содержит множество протяженных повторяющихся последовательностей (8.2%), большинство из которых представляют собой внутрихромосомные дупликации [8, 9].

В 1999–2000 годах в совместной работе с французскими коллегами нами было впервые установлено, что, в отличие от всех других изученных к тому времени эукариотических организмов (дрожжи, нематода, растения, мышь), субъединица РНК-полимеразы II Rpb11 кодируется у человека целым семейством генов, картированных на хромосоме 7 [10, 11]. Были изучены два типа генов из этого семейства (*POLR2J* = *hRPB11a* и *POLR2J2* = *hRPB11b*) и показано, что первый ген кодирует основную изоформу субъединицы hRPB11 (hRPB11a, идентичную единственной субъединице mRPB11 мыши), а экспрессия второго гена приводит в результате альтернативного сплайсинга к синтезу по крайней мере двух различных мРНК и соответственно к генерированию двух новых полипептидов, названных нами hRPB11bα и hRPB11bβ. Кроме того, было установлено, что в обеих мРНК *hRPB11b* терминальный экзон формируется из нуклеотидных последовательностей, характерных для генов семейства *hPMS2L*, кодирующих белки системы репарации дефектов ДНК [11]. Еще один вариант субъединицы hRPB11, кодируемый геном *hRPB11b*, изоформа hRPB11bγ, был обнаружен американскими авторами с помощью метода дрожжевой двухгибридной системы – как белок, взаимодействующий *in vivo* с белком SATB1, который синтезируется исключительно в лимфоидной ткани и специфически связывается с участками прикреп-

[#] Автор для переписки (тел.: (095) 330-65-83; эл. почта: gvs@ibch.ru; факс: (095) 335-71-03).

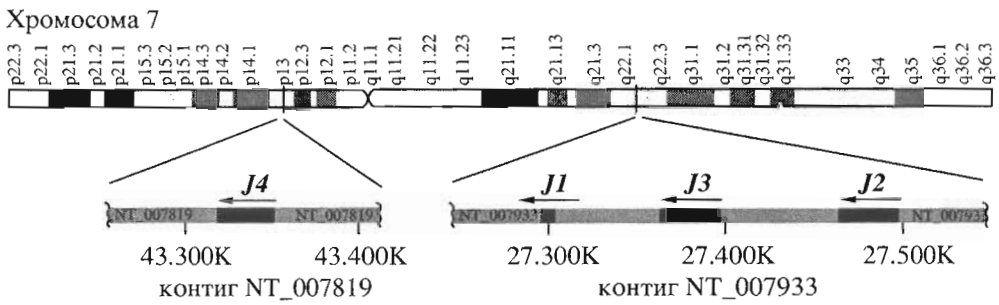

Рис. 1. Схема расположения на хромосоме 7 человека известных ранее (*POLR2J1*, *POLR2J2*) и выявленных в этой работе (*POLR2J3*, *POLR2J4*) генов, кодирующих изоформы субъединицы hRPB11 РНК-полимеразы II *H. sapiens*. Структурные части генов показаны в масштабе черными прямоугольниками, стрелки указывают направление транскрипции генов. На карте хромосомы 7 в виде светлых и темных полос отмечены ее цитогенетические локусы, обнаруживаемые после окрашивания хромосомы специфическим красителем Гимза-Романовского. K – kilobases, т.п.о.

ления ДНК к ядерному матриксу (MAR/SAR) в тимоцитах [12]. Наши экспериментальные данные, такие, как частота встречаемости последовательностей *hRPB11* в геномной клонотеке в сравнении с генами других субъединиц РНК-полимеразы II, а также результаты количественной ПЦР, указывали на то, что в геноме человека может присутствовать от 4 до 12 независимых копий гена *hRPB11* [11] (*POLR2J* по номенклатуре, утвержденной Международной организацией по изучению генома человека HUGO /Human Genome Organization/ [13]), по крайней мере, двух типов – ***a*** и ***b*** [11].

Цель данной работы – проведение биоинформационного поиска в нуклеотидной последовательности хромосомы 7 человека других (всех возможных!) участков, способных кодировать субъединицу hRPB11, а также попытка логически объяснить наличие на хромосоме 7 двух (p15 и q11.22 [14], эти данные фигурируют в подавляющем большинстве ссылок базы данных GenBank) или трех (p12, q11.23 и q22 [11]) различных участков, способных к гибридизации с флуоресцентно-мечеными зондами, специфическими для *hRPB11*.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Компьютерный анализ базы данных генома человека в GenBank/EMBL/DDBJ с помощью программы TBLASTN [15] на присутствие последовательностей экзона, кодирующего *C*-концевой участок изоформы hRPB11bγ (53 а. о., номер депонирования в GenBank AF468111 [12]), выявил четыре гомологичных участка (восемь "попаданий" /hits/ с учетом того, что хромосома 7 *H. sapiens* представлена в GenBank в виде двух альтернативных, несколько различных по сборке суммарных нуклеотидных последовательностей, полученных соответственно фирмой "Celera" (США) [9] и Международным консорциумом по секвенированию генома человека [8]). Три из четырех обнаруженных последовательностей оказались расположенными в контиге NT_007933 (в составе ранее описанных нами генов *POLR2J* и *POLR2J2* [11], а также в области между этими генами), четвертая же входит в состав контига NT_007819 [8] (рис. 1). Одновременно было выявлено (рис. 2), что терминальный экзон изоформы hRPB11bγ присутствует в указанных выше четырех последовательностях в двух формах (в соотношении 2 : 2) – как в гене ***b***-типа (в *POLR2J2* и в новом гене в составе контига NT_007933, обозначенном нами *POLR2J3*), так и в форме, характерной для генов класса ***a*** (в *POLR2J* и в новом гене в составе контига NT_007819, названном нами *POLR2J4*), с единичной аминокислотной заменой Gln138/Gly139,140 в соответствующей изоформе. Такое же расщепление 2 (*J1* и *J4*) : 2 (*J2* и *J3*) наблюдается при анализе генов по основному, характеристическому признаку генов *hRPB11a*- и *hRPB11b*-типов – наличию соответственно двух (Lys17 и Lys18) или одного (Lys17) остатков лизина на границе первого и второго экзонов (рис. 2).

Сходный описанному выше анализ на присутствие последовательностей терминального экзона изоформы hRPB11bβ (40 а. о., номер депонирования в GenBank AJ277740 [11]) выявил только три гомологичных участка (шесть "попаданий" /hits/, все на хромосоме 7), расположенных в генах *POLR2J2*, *POLR2J3* и *POLR2J4*, причем только последний ген имеет в кодирующей части одну аминокислотную замену (Leu92/Pro93) в составе этого экзона по сравнению с характеристической последовательностью. Другой характерной чертой гена *POLR2J4* является то, что, в отличие от трех других генов, в качестве стартового, инициирующего кодона в нем вместо классического ATG (Met) используется достаточно малоэффективный при трансляции (см. [16]) GTG (Val) (рис. 2).

В силу отсутствия в составе гена *POLR2J1* последних экзонов, характерных для изоформ

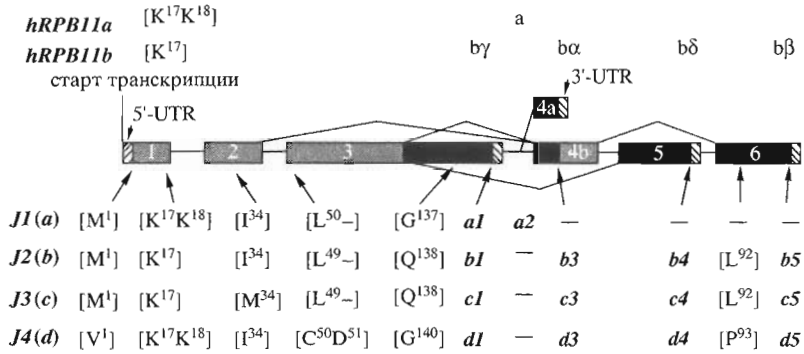

Рис. 2. Схема генерирования и классификация 14 различных видов мРНК *hRPB11*, образующихся в процессе экспрессии генов *POLR2J1–POLR2J4*. Экзоны показаны прямоугольниками, интроны – линиями. Черными прямоугольниками отмечены терминальные экзоны (или варианты экзонов), в которых заканчивается аминокислотная последовательность той или иной изоформы субъединицы hRPB11. В квадратных скобках показаны характеристические особенности (получаемые при трансляции аминокислотные последовательности) экзонов различных генов. Сверху приведены для сравнения обозначения элементов генов и изоформ субъединицы hRPB11 по старой классификации. Прочерк указывает на отсутствие соответствующей мРНК.

hRPB11bα и hRPB11bβ и отделенных от трех первых экзонов достаточно протяженными интронами (см. [11]), ген *POLR2J1* оказался гораздо короче остальных (*POLR2J2–POLR2J4*): соответственно ~5.5 и ~31 т.п.о. (рис. 1). Гены *POLR2J1*, *POLR2J2* и *POLR2J3* образуют в генетической области 7q22.1 длинного плеча хромосомы 7 единый кластер суммарной протяженностью 214530 п. о. (в составе контига NT_007933), в то время как ген *POLR2J4* (31040 п. о.) расположен в цитогенетическом локусе 7p13 короткого плеча хромосомы (контиг NT_007819). Эти позиции хорошо согласуются с определенными нами ранее двумя основными местами локализации на хромосоме 7 участков гибридизации с флуоресцентномеченым зондом (метод FISH), несущим экзоны 1–3 гена *hRPB11a* вместе с располагающимися между ними интронами: в локусах 7p12 и 7q22 [11]. Наличие третьей зоны гибридизации (7q11.23 [11] или 7q11.22 [14]) оставалось загадкой. Поэтому мы проанализировали этот участок на наличие последовательностей, специфических для *hRPB11*. Выяснилось, что область 7q11.22–q11.23 содержит, во-первых, последовательности гена *hPMS2L5*, имеющего родственные экзоны с геном *hRPB11b*, и, во-вторых, нуклеотидные последовательности, пусть в целом и с не очень высокой (64% идентичности), но вполне достаточной для гибридизации (зачастую совпадают от 13 до 24 последовательных нуклеотидов) гомологией с некоторыми участками первого и второго интронов генов *hRPB11* (*POLR2J*). Возможно, в зоне q11.22–q11.23 хромосомы 7, известной высоким содержанием протяженных повторяющихся последовательностей (см., например, сообщение [17] о разнообразных геномных перестройках, вызывающих синдром Вильямса-Бюрена), когда-то существовали последовательности, родственные генам семейств *hRPB11* и *hPMS2*, впоследствии, в основном, утраченные.

Мы также проанализировали способность каждого из четырех обнаруженных нами генов семейства *POLR2J* к экспрессии, прежде всего к синтезу полноразмерных мРНК. Установлено, что при экспрессии четырех генов *POLR2J* человека могут синтезироваться по крайней мере 14 видов зрелых, кодирующих несколько различные изоформы hRPB11, мРНК, 11 из которых обнаружены нами (в виде полноразмерных копий или четко соотносимых фрагментов) в доступных базах EST (expressed sequence tag) и кДНК. В свете этих данных нам пришлось разработать новую классификацию образующихся продуктов экспрессии (мРНК и возможных изоформ субъединицы hRPB11) генов *hRPB11* (*POLR2J*) (рис. 2). По результатам нашего анализа, ген *POLR2J1* способен производить только два вида мРНК (*hRPB11a1* [пока не обнаружена] и *hRPB11a2* [L37127] кодирует основную, характерную для всех живых существ форму субъединицы Rpb11), а с каждого из трех других генов могут синтезироваться (считываться) по четыре вида мРНК: *hRPB11b1* [BI838366], *hRPB11b3* [AJ277739], *hRPB11b4* [пока не обнаружена] и *hRPB11b5* [AJ277740] с гена *POLR2J2*; *hRPB11c1* [AF468111], *hRPB11c3* [AJ277741], *hRPB11c4* [пока не обнаружена] и *hRPB11c5* [R85011] с *POLR2J3*; *hRPB11d1* [BC017250], *hRPB11d3* [BU940313], *hRPB11d4* [BC017341] и *hRPB11d5* [AL699436] с *POLR2J4*. В квадратных скобках указаны номера обнаруженных нами EST, депонированные в базах данных GenBank. В настоящее время мы проводим экспериментальный поиск оставшихся вариантов.

Обнаружение новых генов семейства *POLR2J* = = *hRPB11* и новая классификация продуктов их экспрессии позволили нам проверить и уточнить многие позиции однонуклеотидных полиморфизмов (SNP, single nucleotide polymorphisms) в этих локусах

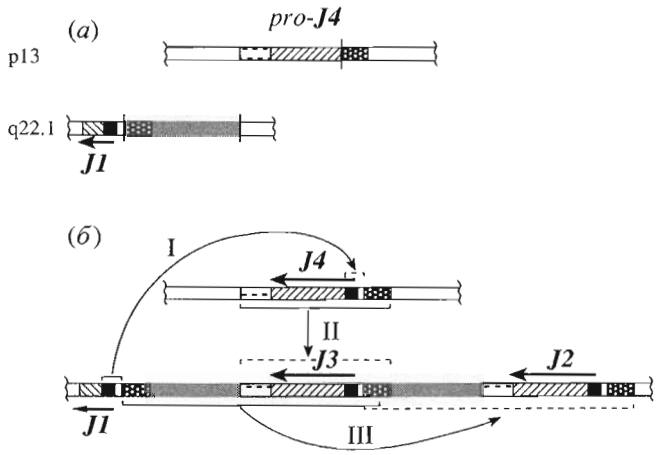

Рис. 3. Схема эволюционного происхождения множественных генов семейства *POLR2J*. Гомологичные участки выделены одинаковым образом. Жирными стрелками показано направление транскрипции и протяженность структурных частей генов, горизонтальными скобками – зоны дупликаций (дупликоны), перемещаемые в ходе трех (I, II и III) последовательных рекомбинационных событий (показаны нежирными стрелками). *а* – исходная организация (структура) цитогенетических локусов p13 и q22.1 хромосомы 7 до первой перестройки генома в этом районе; *б* – схема образования современного состояния хромосомы 7 человека в результате трех последовательных перемещений предварительно дуплицированных областей генома.

ДНК человека. Выяснилось, что примерно половина отмеченных ранее полиморфных замен в составе генов *POLR2J*-семейства [8, 9] таковыми вовсе не являются, а представляют собой характеристические различия нуклеотидных последовательностей разных *POLR2J*-генов. Такие уточнения особенно важны в свете недавно объявленного нового крупномасштабного всемирного проекта по анализу вариабельности генома человеческой популяции [18].

Детальный анализ гомологичных участков в составе генов *POLR2J1*, *POLR2J2*, *POLR2J3* и *POLR2J4* позволил нам также построить вероятную схему эволюционного происхождения множественных генов семейства *POLR2J*. По нашему мнению, такая уникальная структура кластера генов *POLR2J* на хромосоме 7 человека образовалась в результате трех последовательных, увеличивающихся в размерах дупликаций (рис. 3). Исходным следует считать ген *POLR2J1* (*hRPB11a*), кодирующий основную изоформу субъединицы hRPB11, гомологичную на всем протяжении аминокислотной последовательности субъединицам Rpb11 РНК-полимеразы II всех других организмов, включая мышь. Вначале дупликон размером ~6.3 т.п.о., включающий ~1350 п. о. 5'-концевой некодирующей области и около 5 т.п.о. структурной части гена *POLR2J1* (три полных экзона и большую часть третьего интрона), был перенесен из локуса 7q22.1 в проксимальную часть предшественника современного гена *POLR2J4* (*pro-J4*) в составе локуса 7p13 – так была образована генетическая структура генов *hRPB11b*-типа, впервые объединившая экзоны генов *hRPB11* и *hPMS2* (последние изначально присутствовали среди последовательностей *pro-J4*). Именно на такую очередность событий указывает прежде всего гибридная (мозаичная) структура гена *POLR2J4* – он несет в себе элементы как *hRPB11a*- (кодирует два лизиновых остатка на границе первого и второго экзонов и характерную замену Gln138/Gly139, 140 в составе терминального экзона изоформы hRPB11bγ), так и *hRPB11b*-генов (наличие специфических для генов *b*-типа пятого и шестого экзонов) (рис. 2).

Вторым рекомбинационным событием было возвращение уже гораздо более длинного (~68.5 т.п.о.) дупликона из локуса 7p13 в локус 7q22.1 с образованием гена *POLR2J3*. Последний, в свою очередь, в составе дупликона длиной ~108.5 т.п.о. был относительно недавно перемещен в более дальнюю от гена *POLR2J1* область локуса 7q22.1, дав начало современному гену *POLR2J2* (рис. 3). Именно о такой последовательности событий свидетельствует наличие в гене *POLR2J3* более протяженного, чем у *POLR2J2*, участка гомологии с проксимальной частью гена *POLR2J4* (ср. заштрихованные мелкой темной сеткой прямоугольники на рис. 3), а также то, что самые "молодые" из всех четырех, относительно недавно разошедшиеся гены *POLR2J2* и *POLR2J3*, наиболее близки по первичной структуре.

Обнаруженный нами кластер генов из семейства *POLR2J* представляет собой как бы запечатленный моментальный снимок продолжающегося эволюционного процесса. Как хорошо сохранившийся слепок, фиксирующий сразу несколько стадий эволюции генных структур, он в некотором смысле уникален и поэтому особенно интересен для исследований. Дальнейший анализ такой своеобразной генной структуры (кластера генов из семейства *POLR2J*) может прояснить многие важные детали эволюции отдельных человеческих генов и механизмы возникновения белкового разнообразия у высших эукариот.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Для получения реальных физических адресов генов, экзонов и интронов использовался браузер человеческого генома NCBI (http://www.ncbi.nlm.nih.gov).

Поиск участков ДНК, гомологичных последовательностям экзонов и интронов двух уже известных (*POLR2J* и *POLR2J2*), а также двух впервые обнаруженных нами генов (*POLR2J3* и *POLR2J4*) осуществляли с помощью ресурса BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). При анализе протяженных гомологичных участков (более 5 т.п.о.) в используемых программах BLAST варьировали параметр "match" (в сторону увеличения)

и отключали фильтр, что облегчало выявление гомологичных участков на всем протяжении сравниваемых последовательностей.

Множественное сравнение нуклеотидных последовательностей проводили при помощи программы ClustalW, входящей в состав пакета DS Gene компании Accelrys.

Поиск мРНК *H. sapiens* в базах данных GenBank и Human EST осуществляли при помощи программы MEGABLAST. При последующем детальном анализе найденных мРНК "вручную" учитывали приведенные авторами депонированных EST показатели достоверности "чтения" различных участков нуклеотидной последовательности (границы "high quality sequence" в авторских аннотациях).

Настоящая работа частично финансирована грантом № МК-1967.2003.04 Президента Российской Федерации по поддержке молодых кандидатов наук и их научных руководителей и Программой РАН "Физико-химическая биология" (направление "Функциональная геномика").

СПИСОК ЛИТЕРАТУРЫ

1. *Collins F.S., Morgan M., Patrinos A.* // Science. 2003. V. 300. P. 286–290.
2. International Human Genome Sequencing Consortium // Nature. 2001. V. 409. P. 860–921.
3. *Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., et al.* // Science. 2001. V. 291. P. 1304–1351.
4. *Dunham I., Shimizu N., Roe B.A., Chissoe S., Hunt A.R., et al.* // Nature. 1999. V. 402. P. 489–495.
5. *Hattori M., Fujiyama A., Taylor T.D., Watanabe H., Yada T., et al.* // Nature. 2000. V. 405. P. 311–318.
6. *Deloukas P., Matthews L.H., Ashurst J., Burton J., Gilbert J.G., et al.* // Nature. 2001. V. 414. P. 865–871.
7. *Heilig R., Eckenberg R., Petit J.L., Fonknechten N., Da Silva C., et al.* // Nature. 2003. V. 421. P. 601–607.
8. *Hillier L.W., Fulton R.S., Fulton L.A., Graves T.A., Pepin K.H., et al.* // Nature. 2003. V. 424. P. 157–164.
9. *Scherer S.W., Cheung J., MacDonald J.R., Osborne L.R., Nakabayashi K., et al.* // Science. 2003. V. 300. P. 767–772.
10. *Grandemange S., Schaller S., Shpakovski G.V., Kedinger C., Vigneron M.* // Abstracts of Papers Presented at the 4th EMBL Transcription Meeting. Heidelberg, Germany; 26–30 August 2000. P. 129.
11. *Grandemange S., Schaller S., Yamano S., Du Manoir S., Shpakovski G.V., Mattei M.-G., Kedinger C., Vigneron M.* // BMC Molecular Biology. 2001. V. 2. P. 14.
12. *Durrin L.K., Krontiris T.G.* // Genomics. 2002. V. 79. P. 809–817.
13. *Wain H.M., Bruford E.A., Lovering R.C., Lush M.J., Wright M.W., Povey S.* // Genomics. 2002. V. 79. P. 464–470.
14. *Bouffard G.G., Idol J.R., Braden V.V., Iyer L.M., Cunningham A.F., et al.* // Genome Res. 1997. V. 7. P. 673–692.
15. *Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.* // Nucl. Acids Res. 1997. V. 25. P. 3389–3402.
16. *Tailor C.S., Marin M., Nouri A., Kavanaugh M.P., Kabat D.* // J. Biol. Chem. 2001. V. 276. P. 27221–27230.
17. *Osborne L.R., Li M., Pober B., Chitayat D., Bodurtha J., Mandel A., Costa T., Grebe T., Cox S., Tsui L.-C., Scherer S.W.* // Nature Genetics. 2001. V. 29. P. 321–325.
18. *Collins F.S., Green E.D., Guttmacher A.E., Guyer M.S.* // Nature. 2003. V. 422. P. 835–847.

New Genes on Human Chromosome 7: Bioinformation Analysis of a Cluster of Genes from the *POLR2J* Family

D. G. Shpakovski, E. K. Shematorova, and G. V. Shpakovski#

#Phone: +7(095) 330-6583; fax: +7(095) 335-7103; e-mail: gvs@ibch.ru

Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, Moscow, 117997 Russia

Four independent genes encoding various variants of the hRPB11 subunit of *Homo sapiens* RNA polymerase II were revealed in human chromosome 7. Three genes (*POLR2J1*, *POLR2J2*, and *POLR2J3*) form a cluster of total length 214530 bp in the genetic locus 7q22.1 on the long arm of chromosome 7 (contig NT_007933). The fourth gene (*POLR2J4*, 31040 bp) was localized in the cytogenetic locus 7p13 of the short arm of chromosome 7 (contig NT_007819). An analysis enabled us to refine dissimilar experimental data on the mapping of the hRPB11 subunit gene on chromosome 7. In particular, the presence of three sites of its localization according to data on hybridization with fluorescent-labeled probes (the FISH method) was explained. It was established that, upon the expression of the four human *POLR2J* genes, at least 14 types of mature mRNAs encoding somewhat differing hRPB11 isoforms can be synthesized. Eleven of these mRNAs were revealed (as full-length copies or clearly identifiable fragments) in the available databases of expressed sequence tags and cDNAs. The most probable scheme of origination of the multiple genes of the *POLR2J* family, as a result of three consecutive segmented duplications increasing in size, was proposed and substantiated. On the basis of the scheme, some assumptions on the pathways of evolution of separate human genes and the mechanisms of generation of protein diversity in higher eukaryotes were made. The English version of the paper: *Russian Journal of Bioorganic Chemistry*, 2004, vol. 30, no. 6; see also http://www.maik.ru.

Key words: duplications, evolution of genes and proteins, exons, Homo sapiens chromosome 7, nuclear RNA polymerase II, subunit hRPB11