Rozprawa doktorska pt.

Funkcjonalne retrogeny w genomie człowieka

ang. Functional retrogenes in the human genome

Joanna Ciomborowska

Promotor: prof. UAM dr hab. Izabela Makałowska

Pracownia Bioinformatyki
Instytut Biologii Molekularnej i Biotechnologii
Wydział Biologii
Uniwersytet im. Adama Mickiewicza w Poznaniu

Poznań 2014

Składam serdeczne podziękowania:

Pani prof. UAM dr hab. Izabeli Makałowskiej za wszelkie cenne rady, wsparcie, motywację, umożliwienie rozwoju naukowego oraz wspaniałą atmosferę podczas realizacji pracy,

Pani prof. dr hab. Ewie Ziętkiewicz oraz Panu dr hab. Wiesławowi Babikowi za wysiłek włożony w zrecenzowanie niniejszej rozprawy,

Panu prof. dr hab. Bogdanowi Jackowiakowi, Dziekanowi Wydziału Biologii UAM w Poznaniu, szczególnie za wspieranie inicjatyw, zarówno naukowych, jak i organizacyjnych,

Pani prof. dr hab. Zofii Szweykowskiej – Kulińskiej, Dyrektorowi Instytutu Biologii Molekularnej i Biotechnologii oraz Kierownikowi Pracowni Bioinformatyki, szczególnie za stworzenie wspaniałych warunków dla rozwoju młodych naukowców ,

Koleżankom i Kolegom z Pracowni Bioinformatyki: dr Michałowi Szcześniakowi, mgr Michałowi Kabzie, mgr Wojciechowi Rosikiewiczowi, mgr Elżbiecie Kaja oraz mgr Magdalenie Kubiak za pomoc, cenne rady i owocne dyskusje,

Pracownikom i Doktorantom Instytutu Biologii Molekularnej i Biotechnologii oraz wszystkim innym osobom, które okazały mi życzliwość w czasie dotychczasowej pracy naukowej,

Szczególne wyrazy wdzięczności kieruję w stronę mojej Rodziny i Przyjaciół.
Dziękuję, że jesteście!

Dedykuję Rodzicom…

Przedstawiana rozprawa doktorska poświęcona jest funkcjonalnym retrogenom w genomie człowieka, a w szczególności wybranym zagadnieniom związanym z ich ewolucją. Składają się na nią trzy publikacje, w tym jedna praca przeglądowa przedstawiająca dotychczasowy stan wiedzy na temat funkcjonalnych retrogenów. Dwie pozostałe to prace badawcze opisujące nowatorskie analizy i wynikające z nich odkrycia dotyczące powstawania intronów w retrogenach oraz retrogenów, które zastąpiły swoje geny rodzicielskie.

Presented PhD thesis concerns functional retrogenes in the human genome, especially selected aspects of their evolution. It consists of three publications, one review presenting the current state of knowledge in regard to functional retrogenes. Two others are research papers describing innovative analyses and resulting discoveries connected with intron gain in retrogenes and retrogenes that replaced their parental genes.

Spis treści :

I. Streszczenie

1. Wprowadzenie

Odkrycie retrosekwencji było jednym z najbardziej znaczących wydarzeń w badaniach genomu człowieka i innych organizmów. Retrosekwencje, które początkowo były uznawane za nieistotne z biologicznego punktu widzenia elementy, są obecnie szeroko badane i coraz częściej zauważa się ich znaczenie w kształtowaniu genomów i transkryptomów. Obecnie badania retrosekwencji, w tym retrogenów, obejmują m.in. takie aspekty, jak intensywność zjawiska retropozycji, cechy charakterystyczne retrosekwencji, metody ich identyfikacji, ewolucja, funkcjonalność oraz ekspresja.

Retrogeny to kopie genów powstające w wyniku odwrotnej transkrypcji mRNA i wbudowania powstałego cDNA w sekwencję genomową. Opisywany proces nazywany jest retropozycją, a w jego wyniku wieloegzonowy gen rodzicielski daje początek jednoegzonowej retrokopii (Weiner, Deininger et al. 1986). Retrogeny są zwykle nieaktywne i dlatego nazywa się je retropseudogenami lub pseudogenami. Przez wiele lat uznawane były za elementy nieużyteczne, tzw. „śmieciowe" DNA. Od czasu odkrycia pierwszego funkcjonalnego retrogenu w 1985 roku (Soares, Schon et al. 1985) rozpoczęły się badania, które pokazały, że tego typu sekwencje są istotne z punktu widzenia ewolucji genomów.

Pomimo zwiększonego zainteresowania retrogenami, nadal nie wiadomo ile dokładnie jest ich w genomach zwierzęcych, a istniejące dane są zróżnicowane. Przyczyną tych niezgodności są przede wszystkim różnice w metodach ich identyfikacji. Poszukiwanie retrokopii stanowi duże wyzwanie, między innymi ze względu na zróżnicowaną jakość adnotacji poszczególnych genomów, bardzo duże podobieństwo retrosekwencji do genu rodzicielskiego i jego paralogów oraz możliwość zakwalifikowania do retrogenów kopii genów (lub ich fragmentów) powstających w wyniku duplikacji DNA. Uznaje się, że główne cechy retrogenów to brak intronów i elementów regulatorowych, obecność ogona poli(A) oraz powtórzeń otaczających rejon insercji cDNA (Long 2001). W przypadku funkcjonalnych retrogenów dodatkowy wpływ na zróżnicowane wyniki ma samo definiowanie funkcjonalności. Podczas gdy dla jednej z grup badawczych retrogeny były uznawane za funkcjonalne, gdy posiadały co najmniej jedną sekwencję EST (ang. *Expressed Sequence Tag*) lub cDNA jako dowód na ich ekspresję, dla innych ważna była nienaruszona otwarta ramka odczytu (ORF – ang. *Open Reading Frame*) retrogenów (Vinckenbosch, Dupanloup et al. 2006). Często bierze się też pod uwagę tempo zmian pomiędzy retrogenem i genem rodzicielskim określane za pomocą współczynnika Ka/Ks (stosunek liczby mutacji

niesynonimicznych we wszystkich miejscach niesynonimicznych (Ka) do liczby substytucji synonimicznych przypadających na wszystkie miejsca synonimiczne (Ks)) (Betran, Thornton et al. 2002).

Różnice międzygatunkowe kształtują się między innymi poprzez takie zmiany w genomach, jak pojawianie się nowych kopii genów. Ten dodatkowy materiał sprzyja szybkim i intensywnym zmianom ewolucyjnym i dlatego też retropozycję, która jest jednym z głównych źródeł duplikacji genów, uznaje się także jako jeden z najważniejszych mechanizmów sprzyjających powstawaniu zróżnicowania pomiędzy gatunkami. Pomimo, że do tej pory znaleziono stosunkowo niewiele przykładów gatunkowo specyficznych, funkcjonalnych retrogenów, to wiadomo, że ich wpływ na funkcjonowanie organizmu i fenotyp może być znaczący. Bardzo ciekawym przykładem jest retrogen *fgf4* wywołujacy chondrodysplazję u psów. Wszystkie psy ras o krótkich nogach są nosicielami tego retrogenu (Parker, VonHoldt et al. 2009). Inny przykład może stanowić mysi retrogen *Rps23rg1*, który odpowiada za regulację poziomu beta-amyloidu oraz fosforylację białka tau, a więc podstawowe zjawiska towarzyszące chorobie Alzheimera. Odkryto, że mimo intensywnie zachodzącej retropozycji ludzkiego genu rodzicielskiego *RPS23*, *Rps21rg1* występuje tylko u myszy tak jak i drugi funkcjonalny, ulegający ekspresji retrogen *Rps23rg2*. U człowieka natomiast nie zidentyfikowano ortologicznych genów z grupy Rps23rg, a wszystkie powstałe retrokopie są najprawdopodobniej pseudogenami (Zhang, Liu et al. 2009). Inne przykłady to specyficzne dla naczelnych funkcjonalne retrogeny, jak *GLUD2* kodujący dehydrogenazę glutaminianową i ulegający ekspresji w mózgu (Burki and Kaessmann 2004) oraz *CDC14Bretro*, którego gen rodzicielski związany jest z cyklem komórkowym (Rosso, Marques et al. 2008).

Na wczesnych etapach badań nad zduplikowanymi genami uważano, że zazwyczaj jeden z nich akumuluje mutacje i staje się niefunkcjonalny (Haldane 1933, Fisher 1935). W związku z tym, naturalną konsekwencją było powszechne traktowanie wszystkich retrokopii, niefunkcjonalnych w momencie powstania, jako pseudogenów. Jednak z czasem okazało się, że „rozluźniona" selekcja i swoboda ewoluowania jakim podlega większość duplikatów, mogą prowadzić nie tylko do pseudogenizacji, ale i nabywania nowych funkcji (Nei 1969). Z czasem opisano dwa nowe zjawiska związane z ewolucją funkcjonalną po duplikacji: neofunkcjonalizacja, w której jedna kopia zdobywa nowe funkcje, a druga zachowuje dotychczasową oraz subfunkcjonalizacja czyli podział już wykształconej funkcji pomiędzy zaangażowane w duplikację geny (Force, Lynch et al. 1999). Jak pokazały nasze

badania, w przypadku retropozycji istnieje także trzecia opcja, czyli zastąpienie genu rodzicielskiego przez retrogen (Ciomborowska, Rosikiewicz et al. 2013).

W wielu dotychczasowych badaniach postulowano, że retrogeny wykazują wąski, często tkankowo-specyficzny zakres działania i że ulegają ekspresji przede wszystkim w jądrach (Vinckenbosch, Dupanloup et al. 2006), (Bai, Casola et al. 2007), (Pan and Zhang 2009) przy jednoczesnej tendencji genów rodzicielskich do ekspresji w wielu tkankach (Marques, Dupanloup et al. 2005), (Bai, Casola et al. 2007), (Potrzebowski, Vinckenbosch et al. 2008). Jest kilka hipotez, które pomagają wyjaśnić tego typu zjawisko. Po pierwsze tłumaczyć to można stanem tzw. hipertranskrypcji w komórkach spermatogenetycznych, w którym odpowiednio zmodyfikowana chromatyna umożliwia transkrypcję tych fragmentów DNA, które w innych warunkach pozostałyby nieaktywne (Marques, Dupanloup et al. 2005), (Chen, Zou et al. 2011). Druga hipoteza wskazuje na możliwość preferencyjnego wbudowywania retrogenów w rejony aktywnej i otwartej chromatyny, szczególnie w pobliżu genów ulegających ekspresji w komórkach zarodkowych. Takie otoczenie wpływa na zwiększoną ekspresję samych retrogenów, w szczególności w jądrach (Fontanillas, Hartl et al. 2007). Jeszcze inny scenariusz wiąże się z teorią tzw. „ucieczki z chromosomu X". Istnieje dużo przykładów wskazujących na nadreprezentację retrogenów pochodzących od genów rodzicielskich zlokalizowanych na chromosomie X (Betran, Thornton et al. 2002), (Emerson, Kaessmann et al. 2004). Sugeruje się w związku z tym, że „uciekające" po retropozycji na autosomy retrogeny stanowią funkcjonalne odpowiedniki genów źródłowych, które to mogły ulec wyciszeniu w wyniku inaktywacji chromosomu płciowego (Marques, Dupanloup et al. 2005), (Potrzebowski, Vinckenbosch et al. 2008).

Opisane powyżej i inne aspekty badań związanych z retrogenami zostały przeanalizowane i zebrane w rozdziale pod tytułem „Functional retrogenes in animal genomes" w książce *Evolutionary Biology: Mechanisms and Trends.*

2. Utrata i powstawanie intronów w retrogenach

Ogromna większość genów kodujących białka u organizmów eukariotycznych zawiera w swej strukturze introny, których mechanizmy powstawania i wycinania są stosunkowo dobrze poznane i opisane (Chow, Gelinas et al. 1977); (Roy and Gilbert 2006). Wiele dotychczasowych badań wskazało na silny stopień zakonserwowania pozycji intronów w genach nawet daleko spokrewnionych ze sobą organizmów (Rogozin, Wolf et al. 2003); (Carmel, Wolf et al. 2007). Z drugiej strony, metody stosowane w genomice porównawczej

pozwoliły zidentyfikować przykłady utraty, jak i nabywania intronów w toku ewolucji. Zauważono jednak, że powstawanie intronów jest zjawiskiem stosunkowo rzadkim u kręgowców (Loh, Brenner et al. 2007). Co więcej, mimo licznych badań nie zaobserwowano u ssaków żadnego przypadku tzw. „intronizacji", a więc przekształcenia sekwencji egzonowej w intron (Roy, Fedorov et al. 2003); (Coulombe-Huntington and Majewski 2007). Inne opisywane do tej pory mechanizmy uwzględniały jedynie nabywanie intronu poprzez przyłączenie nowego egzonu (O'Neill, Brennan et al. 1998); (Vinckenbosch, Dupanloup et al. 2006); (Fablet, Bueno et al. 2009).

Nasze bioinformatyczne i eksperymentalne badania retrogenów pozwoliły na znalezienie nowych, gatunkowo specyficznych intronów i pokazały, że powstawanie intronów zachodzi także u ssaków. W wyniku analiz procesów retropozycji genów *RNF113* i *DCAF12* znaleźliśmy dwa jednoegzonowe retrogeny u człowieka *RNF113A* i *DCAF12L1*, a także retrokopie z intronami (*RNF113B* oraz *DCAF12L2*). Weryfikacja struktury wyłonionych retrogenów z intronami nastąpiła między innymi poprzez analizę sekwencji EST. Dzięki temu zidentyfikowaliśmy jeden przypadek tzw. „intronizacji" w specyficznym dla naczelnych retrogenie *RNF113B* oraz dwie niezależne intronizacje w retrogenie *DCAF12L2* – jedna z nich nastąpiła u wspólnego przodka naczelnych i gryzoni, a druga tylko u gryzoni. Dodatkowo, jako pierwsi na świecie, znaleźliśmy i potwierdziliśmy eksperymentalnie retrogeny posiadające warianty splicingowe. Co więcej, prześledziliśmy także profile ekspresji zidentyfikowanych retrogenów w kilkunastu organach i tkankach. Pozwoliło to zaobserwować, że wariant bez intronu ulega ekspresji w wielu tkankach, natomiast ewolucyjnie młodszy wariant z intronem wykazuje tendencję do tkankowo specyficznej ekspresji w jądrach. Ograniczona do jąder ekspresja nowego retrogenu jest zgodna z wcześniejszymi doniesieniami wskazującymi na takie właśnie zachowanie wielu retrokopii (Marques, Dupanloup et al. 2005), jednak my pokazaliśmy, że nie zawsze musi ona być charakterystyczna tylko dla retrogenów pochodzących od genów rodzicielskich zlokalizowanych na chromosomie X (Potrzebowski, Vinckenbosch et al. 2008).

Podsumowując, bardzo ciekawym wnioskiem z naszych badań jest to, że z jednej strony retropozycja powoduje utratę intronów a z drugiej, retrogeny są także miejscem intensywnego powstawania nowych intronów u ssaków. Po raz kolejny potwierdza to, że retrogeny stają się źródłem wielu nowości genomowych. Artykuł opisujący powyższe badania ukazał się w czasopiśmie *Molecular Biology and Evolution* pod tytułem „Primate and rodent specific intron gains and the origin of retrogenes with splice variants".

3. „Osierocone“ retrogeny w genomie człowieka

Przez długi czas retrosekwencje opisywane były jako tzw. „śmieci genomowe“ głównie z tego powodu, że uważano je za nieaktywne i z czasem zanikające geny (pseudogeny) (Mighell, Smith et al. 2000). Dopiero niedawno wykazano, że część retrosekwencji może pozostawać w genomie odgrywając istotną rolę, dając na przykład początek nowym genom (Betran, Wang et al. 2002) lub regulatorowym cząsteczkom RNA (Devor 2006). Aspekty pseudogenizacji i neo- lub subfunkcjonalizacji dominowały w dotychczasowych badaniach retrogenów, ale jak dotąd nigdy nie brano pod uwagę tego, że to nie retrogen, ale jego gen rodzicielski może ulec pseudogenizacji. Teoretycznie takiej sytuacji nie można wykluczyć i dlatego podjęliśmy się trudnego zadania poszukania w genomie człowieka przypadków, w których retrogen całkowicie przejął funkcje swojego genu rodzicielskiego. Wykorzystując nowatorskie podejście do tego zagadnienia oraz szereg narzędzi i danych bioinformatycznych przeprowadziliśmy porównawcze analizy genomów człowieka, kury i nicienia, dzięki którym udało się zidentyfikować 25 tzw. „osieroconych” retrogenów w genomie człowieka. Wynik ten jasno pokazał, po raz pierwszy na świecie, że retrogeny nie tylko ulegają pseudogenizacji, neo- lub subfunkcjonalizacji, ale także mogą zastąpić swoje geny rodzicielskie. Prześledzenie historii ewolucyjnej zidentyfikowanych retrogenów pozwoliło z kolei na wykazanie, że większość z nich (14 na 25) powstała i zastąpiła swoje geny rodzicielskie na stosunkowo wczesnych etapach ewolucji zwierząt. Było to zaskakującą informacją, gdyż wcześniejsze doniesienia wskazują na intensywnie zachodzącą retropozycję głównie u ssaków (Moran, Holmes et al. 1996); (Ostlund, Schmitt et al. 2010). Kolejnym ciekawym odkryciem był fakt, że ogromna większość „osieroconych” retrogenów wykazuje szeroki zakres ekspresji w wielu tkankach i organach, co ustalono na podstawie analiz PCR w czasie rzeczywistym. Wspierać to może hipotezę o zastąpieniu genów rodzicielskich, a także stoi w opozycji do wcześniejszych badań wyraźnie sugerujących tkankową specyficzność retrosekwencji, a zwłaszcza silną i często wybiórczą ekspresję w jądrach (Marques, Dupanloup et al. 2005); (Vinckenbosch, Dupanloup et al. 2006); (Potrzebowski, Vinckenbosch et al. 2008). Warto również zaznaczyć, że siedem spośród badanych retrogenów wykazuje powiązania z chorobami u człowieka, takimi jak na przykład rak piersi (Rodriguez, Chen et al. 2007), pląsawica Huntingtona (Carnemolla, Fossale et al. 2009) czy cukrzyca typu 2 (Rosengren, Jokubka et al. 2010).

Nasze szczególne zainteresowanie wzbudził retrogen *CHMP1B*, który zastąpił swój gen rodzicielski w genomie człowieka, podczas gdy u myszy występują obydwa w pełni funkcjonalne geny (zarówno wieloegzonowy, jak i jednoegzonowa retrokopia). Co więcej, poprzez analizę potencjalnych miejsc wiązania czynników transkrypcyjnych pokazaliśmy, że ludzki i mysi retrogen mogą być regulowane w podobny sposób, w przeciwieństwie do genu rodzicielskiego u myszy posiadającego zupełnie inny zestaw elementów regulatorowych. Zidentyfikowaliśmy także silniej zachowany wzór miejsc docelowych dla wiązania miRNA w grupie ortologicznych retrogenów *CHMP1B*, niż w przypadku ich istniejących lub zanikających genów rodzicielskich.

Wszystkie te wyniki wskazują na istotną rolę jaką pełnią retrogeny w kształtowaniu cech gatunkowo specyficznych, a funkcjonalne znaczenie retrogenów podkreśla fakt, że ich mutacje mogą prowadzić do rozwoju poważnych chorób. Warto także zwrócić uwagę na to, że zidentyfikowane przez nas „osierocone" retrogeny nie były wcześniej opisane jako geny powstałe w wyniku retropozycji. Wskazuje to na zasadność badań tych dotąd mało poznanych elementów genomów i pokazuje, że analizy retrogenów znacząco przyczyniają się do lepszego poznania genomów i ich ewolucji oraz zrozumienia różnic międzygatunkowych. Opisywane badania zostały opublikowane w czasopiśmie naukowym *Molecular Biology and Evolution,* w artykule pod tytułem „*Orphan* Retrogenes in the Human Genome".

4. Podsumowanie

Głównym celem mojego projektu doktorskiego było przebadanie zjawiska retropozycji u człowieka pod kątem towarzyszących mu zjawisk ewolucyjnych, takich jak nabywanie nowych intronów czy też zastępowanie genów rodzicielskich przez retrogeny. Przeprowadzone analizy pozwoliły na wykrycie tych zjawisk oraz ich scharakteryzowanie. Wszystkie prowadzone przeze mnie dotychczas badania sprzyjają lepszemu poznaniu ludzkiego genomu, który pomimo zakończenia projektu sekwencjonowania i wielu lat analiz wciąż kryje dużo tajemnic. Nasze i inne badania tzw. „śmieciowego DNA" pokazują, że retrosekwencje mogą być ważne, funkcjonalne, a w wielu procesach wręcz kluczowe. Praca nad tego typu sekwencjami stanowi duże wyzwanie, ale także nieustannie prowokuje do nowych pytań i zadań badawczych. Dlatego też kontynuuję moje badania uczestnicząc w trzech innych projektach związanych z retropozycją. Są to: wielkoskalowa identyfikacja oraz analiza retrogenów i ich genów rodzicielskich w ponad sześćdziesięciu genomach

zwierzęcych; poszukiwanie i charakterystyka kolejnych przykładów intronizacji w retrogenach; a także badanie funkcjonalnych, choć do tej pory uznawanych za pseudogeny, retrogenów specyficznych dla człowieka i istotnych w kształtowaniu różnic międzygatunkowych.

5. Bibliografia

Bai, Y., C. Casola, C. Feschotte and E. Betran (2007). "Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila." Genome Biol 8(1): R11.

Betran, E., K. Thornton and M. Long (2002). "Retroposed new genes out of the X in Drosophila." Genome Res 12(12): 1854-1859.

Betran, E., W. Wang, L. Jin and M. Long (2002). "Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene." Mol Biol Evol 19(5): 654-663.

Burki, F. and H. Kaessmann (2004). "Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux." Nat Genet 36(10): 1061-1063.

Carmel, L., Y. I. Wolf, I. B. Rogozin and E. V. Koonin (2007). "Three distinct modes of intron dynamics in the evolution of eukaryotes." Genome Res 17(7): 1034-1044.

Carnemolla, A., E. Fossale, E. Agostoni, S. Michelazzi, R. Calligaris, L. De Maso, G. Del Sal, M. E. MacDonald and F. Persichetti (2009). "Rrs1 is involved in endoplasmic reticulum stress response in Huntington disease." J Biol Chem 284(27): 18167-18173.

Chen, M., M. Zou, B. Fu, X. Li, M. D. Vibranovski, X. Gan, D. Wang, W. Wang, M. Long and S. He (2011). "Evolutionary patterns of RNA-based duplication in non-mammalian chordates." PLoS One 6(7): e21466.

Chow, L. C., R. E. Gelinas, T. R. Broker and R. J. Roberts (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Rev Med Virol 10(6): 362-371; discussion 355-366.

Ciomborowska, J., W. Rosikiewicz, D. Szklarczyk, W. Makalowski and I. Makalowska (2013). ""Orphan" retrogenes in the human genome." Mol Biol Evol 30(2): 384-396.

Coulombe-Huntington, J. and J. Majewski (2007). "Intron loss and gain in Drosophila." Mol Biol Evol 24(12): 2842-2850.

Devor, E. J. (2006). "Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes." J Hered 97(2): 186-190.

Emerson, J. J., H. Kaessmann, E. Betran and M. Long (2004). "Extensive gene traffic on the mammalian X chromosome." Science 303(5657): 537-540.

Fablet, M., M. Bueno, L. Potrzebowski and H. Kaessmann (2009). "Evolutionary origin and functions of retrogene introns." Mol Biol Evol 26(9): 2147-2156.

Fisher R (1935) The sheltering of lethals. AmNat 69:446-455.

Fontanillas, P., D. L. Hartl and M. Reuter (2007). "Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin." PLoS Genet 3(11): e210.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan and J. Postlethwait (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics 151(4): 1531-1545.

Haldane J (1933) The part played by recurent mutation in evolution. Am Nat 67:5-19.

Loh, Y. H., S. Brenner and B. Venkatesh (2007). "Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon)." Mol Biol Evol 25(3): 526-535.

Long, M. (2001). "Evolution of novel genes." Curr Opin Genet Dev 11(6): 673-680.

Marques, A. C., I. Dupanloup, N. Vinckenbosch, A. Reymond and H. Kaessmann (2005). "Emergence of young human genes after a burst of retroposition in primates." PLoS Biol 3(11): e357.

Mighell, A. J., N. R. Smith, P. A. Robinson and A. F. Markham (2000). "Vertebrate pseudogenes." FEBS Lett 468(2-3): 109-114.

Moran, J. V., S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke and H. H. Kazazian, Jr. (1996). "High frequency retrotransposition in cultured mammalian cells." Cell 87(5): 917-927.

Nei, M. (1969). "Gene duplication and nucleotide substitution in evolution." Nature 221(5175): 40-42.

O'Neill, R. J., F. E. Brennan, M. L. Delbridge, R. H. Crozier and J. A. Graves (1998). "De novo insertion of an intron into the mammalian sex determining gene, SRY." Proc Natl Acad Sci U S A 95(4): 1653-1657.

Ostlund, G., T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings and E. L. Sonnhammer (2010). "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." Nucleic Acids Res 38(Database issue): D196-203.

Pan, D. and L. Zhang (2009). "Burst of young retrogenes and independent retrogene formation in mammals." PLoS One 4(3): e5040.

Parker, H. G., B. M. VonHoldt, P. Quignon, E. H. Margulies, S. Shao, D. S. Mosher, T. C. Spady, A. Elkahloun, M. Cargill, P. G. Jones, C. L. Maslen, G. M. Acland, N. B. Sutter, K. Kuroki, C. D. Bustamante, R. K. Wayne and E. A. Ostrander (2009). "An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs." Science 325(5943): 995-998.

Potrzebowski, L., N. Vinckenbosch, A. C. Marques, F. Chalmel, B. Jegou and H. Kaessmann (2008). "Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes." PLoS Biol 6(4): e80.

Rodriguez, V., Y. Chen, A. Elkahloun, A. Dutra, E. Pak and S. Chandrasekharappa (2007). "Chromosome 8 BAC array comparative genomic hybridization and expression analysis identify amplification and overexpression of TRMT12 in breast cancer." Genes Chromosomes Cancer 46(7): 694-707.

Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin and E. V. Koonin (2003). "Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution." Curr Biol 13(17): 1512-1517.

Rosengren, A. H., R. Jokubka, D. Tojjar, C. Granhall, O. Hansson, D. Q. Li, V. Nagaraj, T. M. Reinbothe, J. Tuncel, L. Eliasson, L. Groop, P. Rorsman, A. Salehi, V. Lyssenko, H. Luthman and E. Renstrom (2010). "Overexpression of alpha2A-adrenergic receptors contributes to type 2 diabetes." Science 327(5962): 217-220.

Rosso, L., A. C. Marques, M. Weier, N. Lambert, M. A. Lambot, P. Vanderhaeghen and H. Kaessmann (2008). "Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein." PLoS Biol 6(6): e140.

Roy, S. W., A. Fedorov and W. Gilbert (2003). "Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain." Proc Natl Acad Sci U S A 100(12): 7158-7162.

Roy, S. W. and W. Gilbert (2006). "The evolution of spliceosomal introns: patterns, puzzles and progress." Nat Rev Genet 7(3): 211-221.

Soares, M. B., E. Schon, A. Henderson, S. K. Karathanasis, R. Cate, S. Zeitlin, J. Chirgwin and A. Efstratiadis (1985). "RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon." Mol Cell Biol 5(8): 2090-2103.

Vinckenbosch, N., I. Dupanloup and H. Kaessmann (2006). "Evolutionary fate of retroposed gene copies in the human genome." Proc Natl Acad Sci U S A 103(9): 3220-3225.

Weiner, A. M., P. L. Deininger and A. Efstratiadis (1986). "Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information." Annu Rev Biochem 55: 631-661.

Zhang, Y. W., S. Liu, X. Zhang, W. B. Li, Y. Chen, X. Huang, L. Sun, W. Luo, W. J. Netzer, R. Threadgill, G. Wiegand, R. Wang, S. N. Cohen, P. Greengard, F. F. Liao, L. Li and H. Xu (2009). "A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation." Neuron 64(3): 328-340.

II. Summary

1. Introduction

The discovery of retrosequences was one of the most significant events in genome analyses. Retrosequences, previously described as completely useless and biologically unimportant elements, now are widely analyzed and their roles in shaping animal genomes and transcriptomes become more and more visible. Currently retrosequences are studied in many aspects, including frequency of retroposition, characteristics of retrosequences, identification methods, their evolution, functionality and expression.

Retrogenes are copies of genes originated from reverse transcription of mRNA and incorporation of cDNA into a genomic sequence. This process is called retroposition and it results in a formation of a single-exon copy from a multi-exon parental gene (Weiner, Deininger et al. 1986). Retrogenes are usually inactive and therefore are commonly called retropseudogenes or just pseudogenes. For many years they have been considered as useless, so called "junk DNA". Nevertheless, since the discovery of the first functional retrogene in 1985 (Soares, Schon et al. 1985) the interest in retrogenes increased and subsequent studies revealed that this type of sequences is important from the evolutionary point of view.

Despite the increased interest in retrogenes, it is still not known how many of them are there in animal genomes and existing data are diversified. This variance may be explained mostly by different methods of retrogenes identification. Searching for retrocopies is a big challenge, mainly because of diverse quality of genomic annotations and high similarity between retrosequences, parental genes and their paralogs. Moreover, there is a strong possibility that genes (or their fragments) duplicated via DNA-based mechanisms can be wrongly qualified as retrogenes. The most distinctive attributes of retrogenes are lack of introns and regulatory elements, presence of poly(A) tail, and repeats located near to the insertion region (Long 2001). In case of functional retrogenes, there is an additional important aspect affecting results, which is the definition of functionality. Whereas for one research group, retrogenes can be considered as functional when their expression can be proved by at least one EST (*Expressed Sequence Tag*) or cDNA sequence, for the others the existence of intact ORF (*Open Reading Frame*) (Vinckenbosch, Dupanloup et al. 2006) or small rate of the evolutionary changes between retrogene and parental gene measured by Ka/Ks ratio (the ratio of non-synonymous substitutions (Ka) to synonymous substitutions (Ks)) (Betran, Thornton et al. 2002) were the most important factors.

Interspecies differences are shaped by various changes in genomes like gain of new gene copies for example. This additional genetic material enhances evolutionary changes and therefore retroposition, as crucial source of duplicated genes, is considered to be one of the most essential processes responsible for interspecies diversification. Although relatively few examples of species-specific functional retrogenes were reported, it is known that their impact on the function of the organism and its phenotype can be significant. A very interesting example is the retrogene *fgf4* responsible for the dogs' chondrodisplasia. All breeds with short legs are carriers of this retrogene (Parker, VonHoldt et al. 2009). Another case is the *Rps23rg1* retrogene in mouse that is responsible for the regulation of beta-amyloid level and tau protein phosphorylation, basic phenomena related to Alzheimer's disease. It was discovered that in spite of the intense retroposition of human parental gene (*RPS23*), *Rps23rg1* and second functional and expressed retrogene *Rps23rg2* exist only in mouse. In human there are no orthologous genes from Rps23rg family and all retrocopies present in the genome are most probably pseudogenes (Zhang, Liu et al. 2009). Other examples of primate-specific functional retrogenes include *GLUD2* coding glutamate dehydrogenase and expressed in brain (Burki and Kaessmann 2004) and *CDC14Bretro*, which originated from parental gene related to cell cycle (Rosso, Marques et al. 2008).

In the early studies of duplicated genes evolution it was postulated that usually one of the duplicates accumulates mutations and becomes nonfunctional (Haldane 1933); (Fisher 1935). Consequently, all retrocopies, which are nonfunctional at the moment of their origin, were considered as pseudogenes. However, it occurred that "relaxed" selection and evolutionary freedom which are characteristic for majority of duplicates, may lead not only to pseudogenization but also to acquisition of new functions (Nei 1969). Over the time two new phenomena related to functional evolution after duplication were described: neofunctionalization, where one copy acquires a new function and the other one keeps the original one and subfunctionalization when maintained function is shared between duplicated genes (Force, Lynch et al. 1999). As our studies showed, there is also another possibility: the retrogene may replace its parent (Ciomborowska, Rosikiewicz et al. 2013).

In many recent studies it has been suggested that retrogenes tend to exhibit narrow, often tissue-specific expression pattern and are expressed mainly in testes (Vinckenbosch, Dupanloup et al. 2006); (Bai, Casola et al. 2007); (Pan and Zhang 2009). At the same time, the general tendency for parental genes to be broadly expressed was shown (Marques, Dupanloup et al. 2005); (Bai, Casola et al. 2007); (Potrzebowski, Vinckenbosch et al. 2008). There are a few possible hypotheses interpreting this phenomenon. Firstly, it can be explained

by so called hypertranscription state in spermatogenic cells, in which modified chromatin enables transcription of DNA that usually remains inactive (Marques, Dupanloup et al. 2005), (Chen, Zou et al. 2011). The second hypothesis is based on the idea of preferential insertion of retrogenes into active and open chromatin, especially near germline-expressed genes. Such surroundings have a big impact on higher expression level of retrogenes, particularly in testes (Fontanillas, Hartl et al. 2007). Another scenario involves so called "out-of-X chromosome escape" theory. There are plenty of examples showing the overrepresentation of retrogenes originated from parental genes located on chromosome X (Betran, Thornton et al. 2002), (Emerson, Kaessmann et al. 2004). It was suggested that retrocopies escaping from chromosome X may work as autosomal counterparts of their source genes, which could be silenced during male meiotic sex chromosome inactivation (Marques, Dupanloup et al. 2005), (Potrzebowski, Vinckenbosch et al. 2008).

These above mentioned, together with other aspects related to retrogenes, were analyzed and described in the chapter entitled "Functional retrogenes in animal genomes" and published in the book *Evolutionary Biology: Mechanisms and Trends.*

2. Gain and loss of introns in retrogenes

Vast majority of protein-coding genes in eukaryotes contains introns, whose origin and splicing machinery are relatively well known and described (Chow, Gelinas et al. 1977); (Roy and Gilbert 2006). Many studies have revealed a strong level of intron position conservation even in distant organisms (Rogozin, Wolf et al. 2003); (Carmel, Wolf et al. 2007). On the other hand, comparative genomics methods allowed to identify many examples of intron loss and gain during the evolution but it was found that intron gain is a very rare event in vertebrates (Loh, Brenner et al. 2007). Moreover, no cases of so called "intronization" - a transformation of exonic sequence into intron - have been discovered in mammals (Roy, Fedorov et al. 2003); (Coulombe-Huntington and Majewski 2007). The only described mechanisms of intron acquisition were associated with new exon capture (O'Neill, Brennan et al. 1998); (Vinckenbosch, Dupanloup et al. 2006); (Fablet, Bueno et al. 2009).

Our bioinformatics and experimental analyses of retrogenes allowed identification of novel, species-specific introns and showed that intron formation process is active also in mammals. As a result of studies of genes *RNF113* and *DCAF12* retroposition, we identified in the human genome two single-exon retrogenes (*RNF113A, DCAF12L1*) as well as retrocopies with introns (*RNF113B* and *DCAF12L2*). Structural verification of obtained

candidates was performed through the analysis of EST sequences. As an outcome we identified one case of "intronization" in primate-specific retrogene *RNF113B* and two independent "intronization" events in the retrogene *DCAF12L2* – one took place in the common ancestor of primates and rodents and another one in the rodent lineage. Additionally, as a first group in the world, we found and experimentally confirmed retrogenes with splicing variants. What is more, we examined expression profiles of those retrogenes in over a dozen of organs and tissues and revealed that a variant without an intron is widely expressed, while a new splicing form containing the intron shows a tendency for tissue-specific expression in testes. Limited to testes expression of the recently originated retrogene variant confirms earlier reports describing such expression pattern of many retrocopies (Marques, Dupanloup et al. 2005). Nevertheless, we demonstrated that not only retrogenes originated from chromosome X show such tendency (Potrzebowski, Vinckenbosch et al. 2008).

Summing up, one of the most interesting conclusions coming from our research is that although retroposition causes intron loss, retrogenes might be regarded as a place of intense intron gain in mammals. This confirms that retrogenes can be viewed as a source of genomic novelties. The article with all described results, entitled "Primate and rodent specific intron gains and the origin of retrogenes with splice variants" was published in the journal *Molecular Biology and Evolution.*

3. „Orphan" retrogenes in the human genome

Retrosequences for a long time have been considered as "genomic junk" mainly because they were regarded as inactive and disappearing over the time genes (pseudogenes) (Mighell, Smith et al. 2000). Only recently it has been showed that some retrosequences may remain in the genome and play a crucial role by giving birth to new genes (Betran, Wang et al. 2002) or regulatory RNAs (Devor 2006). Aspects of pseudogenization together with neo- and sub-functionalization were dominating in all previous analyses and so far nobody took into consideration a situation in which not the retrogene but the parental gene is pseudogenized. Theoretically we cannot exclude such scenario and therefore we decided to take on this challenging task, focusing on identifying in the human genome cases where retrogene overtook the function of its parent. Utilizing innovative approach and a number of bioinformatics tools we performed comparative analyses of human, chicken and nematode genomes, which led us to identification of 25 "orphan" retrogenes in the human genome. These results clearly showed, for the first time, that retrogenes not only can be

pseudogenized, neo- or subfuncionalized, but also are able to replace their progenitors. Analysis of the evolutionary history of identified retogenes showed that majority of them (14 out of 25) originated and replaced their parents in the early stages of animal evolution. It was a surprising result because previous reports suggested very intense retroposition mainly in mammals (Moran, Holmes et al. 1996); (Ostlund, Schmitt et al. 2010). Another fascinating discovery showed that vast majority of "orphan" retrogenes have a broad range of expression, which we demonstrated with real-time PCR experiments. It may support our hypothesis about parental gene replacement and stays in contrary to earlier research suggesting tissue-specificity of retrogenes (specially strong and specific expression in testes) (Marques, Dupanloup et al. 2005); (Vinckenbosch, Dupanloup et al. 2006); (Potrzebowski, Vinckenbosch et al. 2008). It is worth emphasizing that seven of the identified retrogenes are related to human diseases, such as breast cancer (Rodriguez, Chen et al. 2007), Huntington's disease (Carnemolla, Fossale et al. 2009), type 2 diabetes (Rosengren, Jokubka et al. 2010).

A case of retrogene *CHMP1B* aroused our special interest. This gene replaced its source gene in the human genome, while in mouse two fully functional genes (both multi-exon and singe-exon copy) exist. Detailed analyses of potential transcription factor binding sites showed that human and mouse retrogenes may be regulated in a similar way, in contrast to mouse parental gene having completely different set of regulatory elements. We also identified stronger conservation of target binding sites for miRNA among orthologous retrogenes *CHMP1B*, than in case of their existing or disappearing parental genes.

All these results demonstrate a crucial role of retrogenes in shaping species-specific traits. The fact that their mutations are often leading to diseases underlines their functional importance. It is also worth noticing that majority of identified by us retrogenes haven't been earlier described as sequences created via retroposition. This strongly highlights the necessity of further analysis of these still little known genomic elements and shows that retrogene analysis can greatly enrich our knowledge about genomes and their evolution as well as improve the understanding of interspecies differences. Described above results were published in the article "*Orphan* Retrogenes in the Human Genome" in the journal *Molecular Biology and Evolution*.

4. Summary

The main goal of my PhD project was the analysis of retroposition in the human genome with a special focus on accompanying evolutionary phenomena such as intron gain and replacement of parental genes by retrogenes. All performed research led not only to the discovery of abovementioned processes but also to our better understanding of the human genome, which despite the fact of finished sequencing and great number of analyses still hides many secrets. Our and other studies of so called "junk DNA" show that retrosequences might be important, functional and crucial in various processes. Working with such sequences is a big challenge but, at the same time, it constantly gives rise to new questions and research tasks. Therefore, I continue my studies on retrogenes by taking a part in three other projects related to retroposition. These are: identification and large-scale analyses of retrogenes and their progenitors in over sixty animal genomes; searching for next "intronization" events in retrogenes; analysis of human-specific functional retrogenes, considered so far as pseudogenes, which could be involved in formation of interspecies differences.

5. Bibliography

Bai, Y., C. Casola, C. Feschotte and E. Betran (2007). "Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila." Genome Biol 8(1): R11.

Betran, E., K. Thornton and M. Long (2002). "Retroposed new genes out of the X in Drosophila." Genome Res 12(12): 1854-1859.

Betran, E., W. Wang, L. Jin and M. Long (2002). "Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene." Mol Biol Evol 19(5): 654-663.

Burki, F. and H. Kaessmann (2004). "Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux." Nat Genet 36(10): 1061-1063.

Carmel, L., Y. I. Wolf, I. B. Rogozin and E. V. Koonin (2007). "Three distinct modes of intron dynamics in the evolution of eukaryotes." Genome Res 17(7): 1034-1044.

Carnemolla, A., E. Fossale, E. Agostoni, S. Michelazzi, R. Calligaris, L. De Maso, G. Del Sal, M. E. MacDonald and F. Persichetti (2009). "Rrs1 is involved in endoplasmic reticulum stress response in Huntington disease." J Biol Chem 284(27): 18167-18173.

Chen, M., M. Zou, B. Fu, X. Li, M. D. Vibranovski, X. Gan, D. Wang, W. Wang, M. Long and S. He (2011). "Evolutionary patterns of RNA-based duplication in non-mammalian chordates." PLoS One 6(7): e21466.

Chow, L. C., R. E. Gelinas, T. R. Broker and R. J. Roberts (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Rev Med Virol 10(6): 362-371; discussion 355-366.

Ciomborowska, J., W. Rosikiewicz, D. Szklarczyk, W. Makalowski and I. Makalowska (2013). ""Orphan" retrogenes in the human genome." Mol Biol Evol 30(2): 384-396.

Coulombe-Huntington, J. and J. Majewski (2007). "Intron loss and gain in Drosophila." Mol Biol Evol 24(12): 2842-2850.

Devor, E. J. (2006). "Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes." J Hered 97(2): 186-190.

Emerson, J. J., H. Kaessmann, E. Betran and M. Long (2004). "Extensive gene traffic on the mammalian X chromosome." Science 303(5657): 537-540.

Fablet, M., M. Bueno, L. Potrzebowski and H. Kaessmann (2009). "Evolutionary origin and functions of retrogene introns." Mol Biol Evol 26(9): 2147-2156.

Fisher R (1935) The sheltering of lethals. AmNat 69:446-455.

Fontanillas, P., D. L. Hartl and M. Reuter (2007). "Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin." PLoS Genet 3(11): e210.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan and J. Postlethwait (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics 151(4): 1531-1545.

Haldane J (1933) The part played by recurent mutation in evolution. Am Nat 67:5-19.

Loh, Y. H., S. Brenner and B. Venkatesh (2007). "Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon)." Mol Biol Evol 25(3): 526-535.

Long, M. (2001). "Evolution of novel genes." Curr Opin Genet Dev 11(6): 673-680.

Marques, A. C., I. Dupanloup, N. Vinckenbosch, A. Reymond and H. Kaessmann (2005). "Emergence of young human genes after a burst of retroposition in primates." PLoS Biol 3(11): e357.

Mighell, A. J., N. R. Smith, P. A. Robinson and A. F. Markham (2000). "Vertebrate pseudogenes." FEBS Lett 468(2-3): 109-114.

Moran, J. V., S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke and H. H. Kazazian, Jr. (1996). "High frequency retrotransposition in cultured mammalian cells." Cell 87(5): 917-927.

Nei, M. (1969). "Gene duplication and nucleotide substitution in evolution." Nature 221(5175): 40-42.

O'Neill, R. J., F. E. Brennan, M. L. Delbridge, R. H. Crozier and J. A. Graves (1998). "De novo insertion of an intron into the mammalian sex determining gene, SRY." Proc Natl Acad Sci U S A 95(4): 1653-1657.

Ostlund, G., T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings and E. L. Sonnhammer (2010). "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." Nucleic Acids Res 38(Database issue): D196-203.

Pan, D. and L. Zhang (2009). "Burst of young retrogenes and independent retrogene formation in mammals." PLoS One 4(3): e5040.

Parker, H. G., B. M. VonHoldt, P. Quignon, E. H. Margulies, S. Shao, D. S. Mosher, T. C. Spady, A. Elkahloun, M. Cargill, P. G. Jones, C. L. Maslen, G. M. Acland, N. B. Sutter, K. Kuroki, C. D. Bustamante, R. K. Wayne and E. A. Ostrander (2009). "An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs." Science 325(5943): 995-998.

Potrzebowski, L., N. Vinckenbosch, A. C. Marques, F. Chalmel, B. Jegou and H. Kaessmann (2008). "Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes." PLoS Biol 6(4): e80.

Rodriguez, V., Y. Chen, A. Elkahloun, A. Dutra, E. Pak and S. Chandrasekharappa (2007). "Chromosome 8 BAC array comparative genomic hybridization and expression analysis identify amplification and overexpression of TRMT12 in breast cancer." Genes Chromosomes Cancer 46(7): 694-707.

Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin and E. V. Koonin (2003). "Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution." Curr Biol 13(17): 1512-1517.

Rosengren, A. H., R. Jokubka, D. Tojjar, C. Granhall, O. Hansson, D. Q. Li, V. Nagaraj, T. M. Reinbothe, J. Tuncel, L. Eliasson, L. Groop, P. Rorsman, A. Salehi, V. Lyssenko, H. Luthman and E. Renstrom (2010). "Overexpression of alpha2A-adrenergic receptors contributes to type 2 diabetes." Science 327(5962): 217-220.

Rosso, L., A. C. Marques, M. Weier, N. Lambert, M. A. Lambot, P. Vanderhaeghen and H. Kaessmann (2008). "Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein." PLoS Biol 6(6): e140.

Roy, S. W., A. Fedorov and W. Gilbert (2003). "Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain." Proc Natl Acad Sci U S A 100(12): 7158-7162.

Roy, S. W. and W. Gilbert (2006). "The evolution of spliceosomal introns: patterns, puzzles and progress." Nat Rev Genet 7(3): 211-221.

Soares, M. B., E. Schon, A. Henderson, S. K. Karathanasis, R. Cate, S. Zeitlin, J. Chirgwin and A. Efstratiadis (1985). "RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon." Mol Cell Biol 5(8): 2090-2103.

Vinckenbosch, N., I. Dupanloup and H. Kaessmann (2006). "Evolutionary fate of retroposed gene copies in the human genome." Proc Natl Acad Sci U S A 103(9): 3220-3225.

Weiner, A. M., P. L. Deininger and A. Efstratiadis (1986). "Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information." Annu Rev Biochem 55: 631-661.

Zhang, Y. W., S. Liu, X. Zhang, W. B. Li, Y. Chen, X. Huang, L. Sun, W. Luo, W. J. Netzer, R. Threadgill, G. Wiegand, R. Wang, S. N. Cohen, P. Greengard, F. F. Liao, L. Li and H. Xu (2009). "A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation." Neuron 64(3): 328-340.

III. Oświadczenie doktoranta

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

OŚWIADCZENIE DOKTORANTA DOTYCZĄCE JEGO UDZIAŁU W POWSTANIU PRAC NAUKOWYCH STANOWIĄCYCH ROZPRAWĘ DOKTORSKĄ

1. **Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants.**
   (M.W. Szcześniak*, **J. Ciomborowska***, W. Nowak, I.B. Rogozin, I. Makałowska)

Rola doktoranta: dzielone pierwsze autorstwo z dr Michałem Szcześniakiem

Udział doktoranta w badaniach opisanych w tej publikacji obejmował:

- identyfikację opisywanych przypadków tzw. „intronizacji" w retrogenach u naczelnych i gryzoni,
- wielowymiarową analizę bioinformatyczną opisywanych genów,
- potwierdzenie sposobu powstania genów poprzez retropozycję poprzez analizę struktur genów oraz otoczenia genomowego,
- prześledzenie historii ewolucyjnej retrosekwencji w wybranych genomach kręgowców,
- współpracę przy przygotowaniu manuskryptu.

2. **„Orphan" Retrogenes in the Human Genome.**
   (**J. Ciomborowska**, W. Rosikiewicz, D. Szklarczyk, W. Makałowski, I. Makałowska)

Rola doktoranta: Pierwszy autor

Udział doktoranta w badaniach opisanych w tej publikacji obejmował:

- udział w opracowaniu koncepcji badań,
- identyfikację „osieroconych" retrogenów w genomie człowieka,
- sprawdzenie wygenerowanych kandydatów oraz wielowymiarową, kompleksową analizę bioinformatyczną zidentyfikowanych genów,
- analizę ewolucyjną badanych retrogenów oraz ich genów rodzicielskich

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl

- szczegółową analizę genu *CHMP1B* pod kątem miejsc wiązania czynników transkrypcyjnych, miejsc oddziaływania z miRNA,
- zaprojektowanie starterów i przebadanie profili ekspresji „osieroconych" retrogenów metodą PCR w czasie rzeczywistym,
- analizę, wizualizację i opracowanie uzyskanych wyników,
- przygotowanie pierwszej wersji manuskrytpu.

3. Functional Retrogenes in Animal Genomes.

(**J. Ciomborowska**, M. Kubiak, I. Makałowska) – artyku przeglądowy

Rola doktoranta: Pierwszy autor

Udział doktoranta w tworzeniu tej publikacji obejmował:

- udział w opracowaniu koncepcji pracy,
- zebranie i szczegółowe przenalizowanie większości wykorzystanych danych literaturowych,
- opracowanie zebranych materiałów,
- udział w przygotowaniu manuskryptu.

Poznań, dnia 10.01.2014

J. Ciomborowska

mgr Joanna Ciomborowska
Pracownia Bioinformatyki
Wydział Biologii UAM w Poznaniu

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl

IV. Oświadczenia współautorów

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

W związku z wykorzystaniem przez mgr Joannę Ciomborowską poniżej wymienionych publikacji jako rozprawy doktorskiej oświadczam, iż udział mój, jako promotora, polegał przede wszystkim na wspólnym opracowywaniu koncepcji badań oraz przygotowywaniu manuskryptów. Jednocześnie stwierdzam, iż we wszystkich wymienionych publikacjach wkład pracy mgr Joanny Ciomborowskiej był niezwykle duży. Przeprowadzone przez mgr Ciomborowską analizy i opracowania wyników były fundamentalne dla powstania poniżej wymienionych prac.

1. **Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants**
   (M.W. Szcześniak*, J. Ciomborowska*, W. Nowak, I.B. Rogozin, **I. Makałowska**)

2. **„Orphan" Retrogenes in the Human Genome**
   (J. Ciomborowska, W. Rosikiewicz, D. Szklarczyk, W. Makałowski, **I. Makałowska**)

3. **Functional Retrogenes in Animal Genomes**
   (J. Ciomborowska, M. Kubiak, **I. Makałowska**)

Poznań, 14.01.2019

prof. UAM dr hab. Izabela Makałowska
Pracownia Bioinformatyki
Wydział Biologii UAM w Poznaniu

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

Functional Retrogenes in Animal Genomes

(J. Ciomborowska, **M. Kubiak**, I. Makałowska)

Rola : współautor

Oświadczam, że mój udział w tworzeniu tej publikacji obejmował pomoc w zebraniu danych literaturowych związanych z funkcjonalnymi retrogenami u zwierząt.

Poznań, dnia 13.01.2014

Magdalena Kubiak

mgr Magdalena Kubiak

Pracownia Bioinformatyki

Wydział Biologii UAM w Poznaniu

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl

www.biologia.amu.edu.pl

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants

(**M.W. Szcześniak***, J. Ciomborowska*, W. Nowak, I.B. Rogozin, I. Makałowska)

Rola : * dzielone pierwsze autorstwo z mgr Joanną Ciomborowską

Oświadczam, że mój udział w badaniach opisanych w tej publikacji obejmował:

- analizę miejsc splicingowych zidentyfikowanych w retrogenach
- zaprojektowanie starterów do badań eksperymentalnych metodą PCR,
- przebadanie profili ekspresji metodą PCR,
- współpracę przy przygotowaniu manuskryptu.

Poznań, dnia 10.01.2014

Michał Szcześniak

dr Michał Szcześniak
Pracownia Bioinformatyki
Wydział Biologii UAM w Poznaniu

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl

www.biologia.amu.edu.pl

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants

(M.W. Szcześniak*, J. Ciomborowska*, **W. Nowak**, I.B. Rogozin, I. Makałowska)

Rola : współautor

Oświadczam, że mój udział w badaniach opisanych w tej publikacji obejmował:

- pomoc podczas części eksperymentalnej zaplanowanych badań (PCR, elektroforeza w żelu agarozowym),
- przygotowanie i przeprowadzenie reakcji sekwencjonowania wybranych produktów reakcji PCR,

Poznań, dnia 10.01.2014

Witold Nowak

dr Witold Nowak
Wydziałowa Pracownia Technik
Biologii Molekularnej
Wydział Biologii UAM w Poznaniu

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl

www.biologia.amu.edu.pl

Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants

(M.W. Szcześniak*, J. Ciomborowska*, W. Nowak, **I.B. Rogozin**, I. Makałowska)

Role : co-author

I declare that my contribution to this publication covers re-analysis of results to confirm the intronization.

Bethesda, 5/12 2013

.....

Igor Rogozin, PhD
NCBI, Computational Biology Branch

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

„Orphan" Retrogenes in the Human Genome

(J. Ciomborowska, **W. Rosikiewicz**, D. Szklarczyk, W. Makałowski, I. Makałowska)

Rola : współautor

Oświadczam, że mój udział w badaniach opisanych w tej publikacji obejmował udział w identyfikacji „osieroconych" retrogenów w genomie człowieka poprzez przeprowadzenie analizy bioinformatycznej genów ortologicznych *Homo sapiens* i *Caenorhabditis elegans.* Dzięki temu wygenerowano zbiór potencjalnych retrogenów u człowieka i ortologów genów rodzicielskich u nicienia, który został poddany dalszym analizom i ręcznej weryfikacji.

Poznań, dnia 14.11.2013

W. Rosikiewicz

mgr Wojciech Rosikiewicz
Pracownia Bioinformatyki
Wydział Biologii UAM w Poznaniu

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 55 56, fax. +48 61 829 56 36
biologia@amu.edu.pl
www.biologia.amu.edu.pl

„Orphan" Retrogenes in the Human Genome

(J. Ciomborowska, W. Rosikiewicz, **D. Szklarczyk**, W. Makałowski, I. Makałowska)

Rola : współautor

Oświadczam, że mój udział w badaniach opisanych w tej publikacji obejmował pilotażowe analizy bioinformatyczne genomów czlowieka i kury pozwalające na późniejszą optymalizację potoku analitycznego.

Zurich, 11.11.2013

Damian Szklarczyk

Damian Szklarczyk
Institute of Molecular Life Sciences
University of Zurich

WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

medizinische
fakultät
Westfälische
Wilhelms-Universität Münster

Institut für Bioinformatik
Niels-Stensen-Straße 14 | 48149 Münster

Ms. Joanna Ciomborowska
Laboratory of Bioinformatics
Institute of Molecular Biology and Biotechnology
Faculty of Biology
Adam Mickiewicz University in Poznan
Umultowska 89
PL-61-614 Poznan

November 15, 2013

"Orphan" Retrogenes in the Human Genome

(J. Ciomborowska, W. Rosikiewicz, D. Szklarczyk, W. Makałowski, I. Makałowska)

Role: co-author

I declare that my contribution to the listed publication was limited to advising on a scientific context of the project.

Sincerely,

Wojciech Makałowski, Ph.D.
Professor and Director
Institute of Bioinformatics

V. Publikacje wchodzące w skład rozprawy doktorskiej

Ciomborowska J., Kubiak M., Makalowska I.

Functional Retrogenes in Animal Genomes

Evolutionary Biology - Mechanisms and Trends

Springer-Verlag Berlin Heidelberg, 2012, pp 283-300

ISBN: 978-3-642-30424-8

Chapter 16
Functional Retrogenes in Animal Genomes

Joanna Ciomborowska, Magdalena Kubiak and Izabela Makałowska

Abstract The discovery of retrogenes was one of the most surprising breakthroughs of human genomics and had a big impact on other species genomic analyses. Since that moment, retrosequences first considered as useless and unimportant biological elements have been started to be widely studied. Now we know that retrogenes may be functional and can play a crucial role in shaping genomes and transcriptomes, working as sources of new genes or regulatory elements. Here, we describe some insights from RNA-based duplication studies which are focused mainly on numbers of retrogenes in various animal species, methods of functional retrogenes identification, their evolution, and impact on developing interspecies differences.

16.1 Introduction

Duplication is considered to be the most important source of new genes and a major force driving genome evolution (Ohno 1970). The earliest observation about functional importance of duplication was reported in 1936, while it was proved that sequence duplication could reduce eye size in *Drosophila melanogaster* mutant (Bridges 1936). This report, however, did not bring much attention until the late 1960s when papers by Nei (1969) and Ohno (1970), underlining the importance of duplications were published. Throughout the following years, an interest in this topic began to grow and many researchers started to work on problems like the rate of duplication in various genomes, mechanisms involved in duplicates formation and importance of this phenomenon in evolutionary processes.

J. Ciomborowska (✉) · M. Kubiak · I. Makałowska
Laboratory of Bioinformatics, Adam Mickiewicz University,
61-614 Poznan, Poznań, Poland
e-mail: joannac@amu.edu.pl

P. Pontarotti (ed.), *Evolutionary Biology: Mechanisms and Trends*,
DOI: 10.1007/978-3-642-30425-5_16,

Table 16.1 Summary of main features of DNA-based and RNA-based gene duplications

Feature	RNA-based duplicates	DNA-based duplicates
Mechanism of creation	Retroposition	Tandem or segmental duplication, unequal crossing over, genome or chromosome duplication
Changes in gene features	Yes, lack of introns, regulatory elements	No, usually all elements maintain unchanged
Original promoter sequences	Lost	Present
Interchromosomal gene movement	Yes, usually parental gene and retrocopy are on different chromosomes	No, duplicates rather on the same chromosome
Changes in function	Parental gene usually maintains ancestral function while retrocopy acquires new function	Functions among duplicates are often the same, unchanged
Expression profile	Often testis dominant expression, tissue-specific expression pattern	Less frequently transcribed in testes, broad expression pattern
Differences in expresssion	Retrocopy expression can be different from parental gene	DNA duplicates usually exhibit similar expression pattern
Ability to become functional	Less likely to become functional because of lack of regulatory elements but can evolve in unique, new way	Can become more easily functional but limited evolution of new functions

There are four main mechanisms responsible for gene duplication: (i) unequal crossing over, (ii) chromosome or genome duplication, (iii) segmental duplication, and (iv) retroposition (Zhang 2003). However, the main classification of gene duplicates is related to the source of their origin and therefore we distinguish DNA-based duplicates and RNA-based duplicates-retrosequences. Duplicates originated by these mechanisms differ significantly in their nature. The most important features associated with sequence, structure, expression, and localization of these two types of duplicates are summarized in Table 16.1.

This review is focused on covering some insights from RNA-based gene duplication studies, especially on methods of functional retrogenes identification and the impact of retroposition on shaping animal genomes. In this type of duplication, genes get "cloned" via retroposition in which mRNA is reversely transcribed into cDNA (complementary DNA) and reintegrated into a new location in the genome (Weiner et al. 1986) (Fig. 16.1). The key role in this process is played by reverse transcriptase that may originate from different types of retrotransposable elements. In mammals, reverse transcriptase is provided by L1 element as demonstrated by Esnault et al. (2000) and Wei et al. (2001). Retrogenes can be recognized by a few characteristic features: lack of introns and regulatory elements, presence of poly-A tracts, and direct repeats flanking the cDNA insertion area (Long 2001).

First communication about sequences that are nonfunctional copies with high similarity to protein coding genes but containing some genetic defects, like premature stop codons or frameshifts mutations, was published in 1977 (Jacq 1977). These elements were described as pseudogenes and originated from both types of

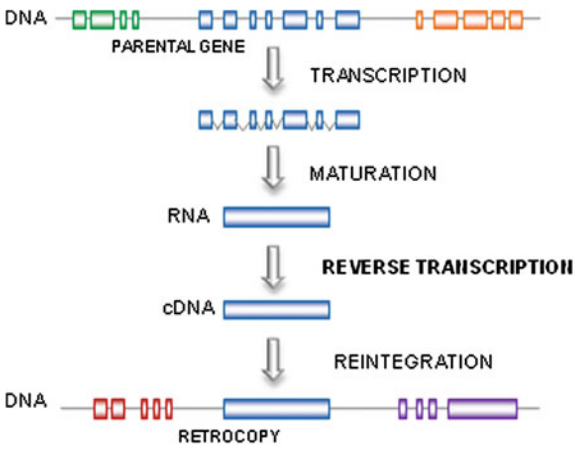


Fig. 16.1 Mechanism of retroposition

duplication events mentioned above. A few years later some well-studied examples of RNA-based duplicates were published (Hollis et al. 1982; Karin and Richards 1982; Ueda et al. 1982). These interesting discoveries of pseudogenes triggered many studies of intergenic regions in order to check whether these sequences are truly representing only so-called "junk DNA" as it was postulated (Balakirev and Ayala 2003). Several analyses pointed out that a great deal of this genetic material can play an important role in creation of new genes and regulatory non-coding RNAs (Makałowski 2000; Bai et al. 2007; Yu et al. 2007).

In retroposition multi-exon genes give birth to single-exon copies. Numerous studies show that most retroposed genes are nonfunctional, inactive, and considered as biologically insignificant sequences. The main reason lies in the fact that these copies, in most cases, lack regulatory elements. The first functional retrogene was discovered in 1985 (Soares et al. 1985). This investigation was followed by other findings of functional retrogenes in mammalian (McCarrey and Thomas 1987; Brosius 1999) and fruit fly genomes (Betran et al. 2002a, b) and recently also in a number of vertebrates and mosquito (Pan and Zhang 2009) as well as chicken and silkworm (Toups et al. 2011). The studies showed that many of these duplicates did recruit regulatory regions (Mighell et al. 2000) and produced new, very often lineage-specific genes (Betran et al. 2002a, b; Marques et al. 2005; Svensson et al. 2006; Sakai et al. 2007; Szczesniak et al. 2011). As latest studies show, these genes may very quickly become essential (Chen et al. 2010). They can also lead to new protein domains through fusion with other genes (Vinckenbosch et al. 2006; Baertsch et al. 2008; Ohshima and Igarashi 2010), new regulatory RNAs (Yano et al. 2004; Devor 2006), or other regulatory elements (Brosius 1999; Nozawa et al. 2005).

Retrogenes, for a long time considered to be not important copies of parental genes are nowadays called "seeds of the evolution", since they made a significant contribution to molecular evolution (Brosius 1991). It has been shown that they play an important role in the diversification of transcriptomes and proteomes and may be responsible for the wealth of species-specific features (Betran et al. 2002a, b; Balasubramanian et al. 2009). As duplicates of their parental genes, they evolve

relatively fast because duplication events allow a relaxed purifying selection; thus these genes may acquire novel functions. A very elegant example of the functional retrogene phenotypic effect was presented by the group of Elaine Ostrander. They showed that retrogene *fgf4* is responsible for dogs' chondrodysplasia. Consequently, carriers of the *fgf4* retrogene have short legs (Parker et al. 2009). Many species-specific traits related to retrogenes are of high importance, not only from the evolutionary point of view but also in medical research as they may be responsible for the fact that results from animal studies cannot be transferred into humans. For example, a functional mouse retrogene *Rps23r1* reduces Alzheimer's beta-amyloid levels and tau phosphorylation (Zhang et al. 2009). However, results of this study cannot be applied to humans since this particular retrogene is rodent specific and does not exist in the human genome. Another interesting case comes from the *PYDC2* retrogene (also known as *POP2*). *PYDC2* is involved in regulating NF-kappaB activity and inflammasome formation (Bedoya et al. 2007). This retrocopy is present and functioning only in the genomes of hominids and Old World primates and is absent from genomes of mice, rats, and other mammals including New World monkeys (Atianand et al. 2011).

Retrocopies of protein coding genes are also known to be involved in many diseases. A good example is the *RHOB* gene, a tumor supressor of the Rho GTPases family, which arose by retroposition in the early stage of vertebrate evolution (Prendergast 2001). Mutation in another retrogene, *TACSTD2*-tumor associated calcium signal transducer 2, causes gelatinous drop-like corneal dystrophy leading to blindness (Tsujikawa et al. 1999). Our studies showed that out of 29 retrogenes, which replaced their progenitor, 9 are associated with human diseases including cancer, diabetes, attention-deficit/hyperactivity disorder, Huntington's disease, and other (Ciomborowska et al. unpublished data).

Duplication by retroposition may also take place in case of other than protein coding genes. For the first time, this possibility was suggested by Brosius in 2003 (Brosius 2003) and in the following years bioinformatics evidence for such events was provided (Weber 2006; Luo and Li 2007). It was proposed that retroposition could be especially feasible for RNAs that are being processed from introns, like snoRNAs or miRNAs (Volff and Brosius 2007).

The discovery that retrosequences, considered as "junk DNA", may be functional and play a crucial role in shaping genome specific features was one of the most surprising breakthroughs in human and other genome analyses. Many studies were recently performed to explore these unique sequences yet, our knowledge about retrogenes evolution, function, and impact on shaping animal genomes is still exceptionally limited.

16.2 Numbers of Functional Retrogenes in Animal Genomes

Despite the growing interest in retrogenes and a fact that a large number of communications reporting functional retrogenes were published, it is still unknown how many of them are actually transcribed in human and other genomes. Currently

Table 16.2 Numbers of identified functional retrogenes in selected publications

	1[a]	2	3	4	5	6	7	8	9	10
Human	131	631	163	77	94	117	7	–	–	–
Chimp	–	476	199	–	–	–	–	–	–	–
Macaca	–	–	275	–	–	–	–	–	–	–
Mouse	–	663	154	147	105	–	–	–	–	–
Rat	–	567	226	–	–	–	–	–	–	–
Cow	–	790	163	–	–	–	–	–	–	–
Dog	–	409	95	103	–	–	–	–	–	–
Opossum	–	–	232	152	–	–	–	–	–	–
Platypus	92	–	–	–	–	–	–	–	–	–
Chicken	51	321	99	–	–	–	–	–	–	–
Frog	140	–	–	–	–	–	–	–	–	–
Lizard	136	–	–	–	–	–	–	–	–	–
Fugu	142	–	–	–	–	–	–	–	–	–
Medaka	131	–	–	–	–	–	–	–	–	–
Stickleback	111	–	–	–	–	–	–	–	–	–
Tetraodon	60	221	–	–	–	–	–	–	–	–
Zebrafish	119	–	140	–	–	–	–	–	–	–
Amphioxus	173	–	–	–	–	–	–	–	–	–
Sea squirt	96	–	–	–	–	–	–	–	–	–
Fruit fly	–	–	212	–	–	–	–	94	23	21
Mosquito	–	–	108	–	–	–	–	–	–	–

[a] The list of publications is as follows: 1. Chen et al. 2011, 2. Yu et al. 2007, 3. Pan and Zhang 2009, 4. Potrzebowski et al. 2008, 5. Emerson et al. 2004, 6. Vinckenbosch et al. 2006, 7. Marques et al. 2005, 8. Bai et al. 2007, 9. Betran et al. 2002a, b, 10. Metta and Schlotterer 2010

available data comprise 19 chordate species and two insects (Table 16.2), but even for the most studied genomes, like human or fruit fly, results are quite dispersed. The main reason for these differences lies in methods applied for retrocopies identification and criteria used for defining a given copy as functional. For example, Vinckenbosch et al. (2006) considered as a functional retrogenes only those copies, which had at least one EST (expressed sequence tag) and an intact ORF (open reading frame). Therefore, retrogenes, which change their functions and serve as regulatory RNA for instance, were excluded from their set as those usually do not have an intact ORF. Emerson et al. (2004) for a retrocopy to be evaluated as functional, in addition to the expression evidence, required a nonsynonymous to synonymous substitution rate ratio (Ka/Ks) to be lower than 0.5. The conservation of the ORF was not necessary.

It was estimated that the human genome contains about 8,000 (Zhang et al. 2003) to over 10,000 (Harrison et al. 2002) retrogenes. In following studies they found that some 4–6 % of them were widely expressed (Harrison et al. 2005). Using bioinformatic tools Vinckenbosch et al. (2006) identified over 1,000 transcribed retrogenes, out of which 117 evolved in bona fide genes. As mentioned above, they considered a retrogene as functional only if its ORF inherited from parental gene remained intact. Another research group identified 631 functional retrogenes in the

human genome and suggested that 2–3 % of all human genes belong to this category (Yu et al. 2007). One of the first screenings of other than human mammalian genomes, in order to identify expressed retrogenes, was done by Emerson and coworkers (Emerson et al. 2004). They obtained 94 functional retroposed and transcribed genes in the human and 105 in the mouse genome. Most recent studies, performed on 10 vertebrate and 2 insect genomes, substantially increased those numbers (Pan and Zhang 2009). Estimated by this group number of functional retrogenes in the human genome is 163 and in mice 154. They also identified considerable number of functional retrogenes in other species (see Table 16.2). Differences in the number of these genes in various mammals are, at this point, difficult to interpret and therefore, we cannot draw any general conclusions. Results may be affected, for instance, by the quality and level of genome sequencing. For example, dogs have relatively small number of identified functional retrogenes. However, the dog genome was not sequenced with the same coverage as human or mouse and there is much less transcripts available for this species.

Marques et al. (2005) postulated that there was an exceptional burst of retroposition in the human lineage and many young retrogenes significantly contributed to the origin of new human-specific genes. Nevertheless, the study by Pan and Zhang (2009) showed that retroposition gave birth to more new genes in mice than in humans. Therefore, rapid emergence of retrogenes might be a common phenomenon in mammals.

It is noticeable that chicken has overall low level of retrogenes (Pan and Zhang 2009; Chen et al. 2011). This is explained by the fact that the reverse transcriptase in chicken is encoded by unique LINE-like elements (LINE-long interspersed elements) that are unlikely to copy poly-A mRNA (Burch et al. 1993; Haas et al. 2001). The number of retrogenes, however, does not only depend on the type of LINE elements present in a given genome. It was observed that fruit fly has a high level of retroposition and it seems that this level is much higher than in humans. Nevertheless, it also has a different response to retroposition events. It has been shown that euchromatic inserts of retroposons are under much stronger purifying selection and therefore are quickly eliminated (Eickbush and Furano 2002). In another insect, anopheles, there is also a big loss of retrocopies. In this case it was postulated that the rapid disappearance of retrotransposons is just a stochastic process (Crainey et al. 2005). Chen et al. (2011) explained the difference in the functional retrogenes number between mammalian and non-mammalian chordate species by the dissimilar path of their evolution. While in mammalian genomes the majority of retrocopies become retropseudogenes and never gain functionality, in non-mammalian chordates most of these sequences have intact open reading frame and could be functional. This finding reflects actually the previous above-mentioned findings suggesting a higher rate of retroposons turnover in non-mammalian species. Therefore, mammals possess a large number of old retrogenes, which had enough time to accumulate mutations and their open reading frames are not intact anymore. In species where retrogenes are eliminated at a high rate we observe more young copies, which still have undamaged ORF.

The number of functional retrogenes also seems to vary among gene families. Some of them appear to have higher than average rate of retroposition and gain of function. Zhang et al. (2002) identified 2090 retrocopies of ribosomal protein genes in the human genome. Out of them 12.3 % were intact. Yu et al. (2007) found that ribosomal-protein genes are statistically overrepresented among retrogenes and Pan and Zhang (2009) showed that copies of ribosomal protein genes compose 28 % of all retrogenes with intact ORF in mammalian species. Our studies on ribosomal protein genes revealed, based on RNA-Seq and transcription start site data, that 17 % of identified retrocopies is transcribed in at least one of 30 screened libraries (unpublished data). All these reports confirmed earlier studies showing that house-keeping genes in general have more retrocopies (Goncalves et al. 2000). The higher level of mRNA and therefore higher likelihood of retroposition could explain this phenomenon. However, Balasubramanian et al. (2009) compared the expression level of ribosomal protein mRNA to the number of retrocopies and did not observe any correlation. They concluded that expression level is not the only major component determining the number of retrocopies arising from a gene. Their analyses indicated that sequence composition could be an important factor influencing the activity of the retroposition. Also, Goncalves et al. (2000) and Zhang et al. (2002) suggested that reverse-transcription and transposition might depend on the sequence composition since mRNAs of genes with higher number of retrocopies were GC-poor.

16.3 Methods of Retrogenes Identification

As mentioned above, definitions of functional retrogenes differ within the literature. As a consequence, also methods of retrogenes identifications are diverse. There are postulates that these methods should be carefully revised taking into account the evaluation of DNA- and RNA-based duplicates classification (Zhang et al. 2011).

The first and obvious step in functional retrogenes identification is to find pairs of putative retrogene and its parental counterpart within a given genome. The most common method of retrogenes identification is based on BLAST (Basic Local Alignment Search Tool) analyses in which proteins encoded by multi-exon genes serve as a query against genomic sequence (TBLASTN) (Altschul et al. 1990). All groups studying retroposition of genes at the whole genome level used this approach, regardless of analyzed species. Differences between strategies undertaken by various groups lie in filtering the results, although some requirements were shared by most groups, like the alignment covering majority of parental gene exon–exon junctions (Marques et al. 2005; Vinckenbosch et al. 2006; Meisel et al. 2009; Zhang et al. 2011). However, while Marques et al. (2005) required that the alignment have minimum length of 50 amino acids and covered more than 70 % of the sequences, Chen et al. (2011) accepted alignments covering minimum of 60 % of query and subject and more than 40 % identity. Other investigators considered for further analysis only hits with 50 % identity and the overlap level between two

proteins at least 70 % (Betran et al. 2002a, b; Bai et al. 2007; Toups et al. 2011). In some studies BLAST e-value was also used as a cutoff for initial results. For example, Svensson et al. (2006) set this value at the level of 10^{-10}.

Selection of criteria that can be helpful in obtaining the most reliable results is probably the most important step in the process of retrogene-parent pair identification. Summarizing published works, we may say that the prevalent set of such criteria consist of: (i) sequence similarity and sequence coverage in an alignment, (ii) coverage of parental gene exon–exon junctions, (iii) difference in the genomic localization. In addition, in case of young retrogenes, criteria may include, (iv) traces of poly-A tail and (v) insertion site repeats.

Pinpointing retrocopies in the genome is a starting step in the way to identify those that are functional. The most obvious method for searching functionality is to look for the evidence of the expression. Performing this step most of the research groups were utilizing only computational approaches and mapped identified retrogenes to ESTs and mRNAs (Emerson et al. 2004; Vinckenbosch et al. 2006; Bai et al. 2007; Baertsch et al. 2008) or microarray data (Potrzebowski et al. 2008). These analyses are quite challenging because of high sequence similarity between parental gene and its copy. It means that in some cases mRNAs or ESTs sequences of parental or paralogous genes can be wrongly assigned to retro-duplicates and vice versa (Zheng and Gerstein 2007) which can lead to incorrect interpretation of the results. Other methods relied on available annotations in databases and considered only these genes, which already were annotated as functional (Betran et al. 2002a, b; Pan and Zhang 2009). In some cases, Gene Ontology categories were used to confirm functionality and expression of identified retrogenes (Bai et al. 2007; Yu et al. 2007). Another widely used measure of functionality is the Ka/Ks ratio calculated for parental-retrogene pairs (Betran et al. 2002a, b; Vinckenbosch et al. 2006). Usually, retrocopy to be considered as functional should have Ka/Ks ratio less than 0.5 (Betran et al. 2002a, b; Vinckenbosch et al. 2006; Chen et al. 2011). Some researchers used, as an indicator of retrogene functionality, conservation of the open reading frame (ORF) (Potrzebowski et al. 2008; Chen et al. 2011). These parameters were used based on the postulate that retrocopies might be mainly a result of subfunctionalization, i.e., they perform the same function as parents but at different time or in different tissues (Force et al. 1999) and therefore there should be some evolutionary constrains on changes in the coding region. This approach is very limiting in finding all functional retrogenes since, as already mentioned by us, retrogenes may undertake entirely new functions, as regulatory elements for example, and as such do not necessarily need the ability to code for the protein. Excellent examples are here retrogenes serving as microRNA sponges. These functional retrocopies do not have conserved ORFs as they do not code for proteins. The main role of these gene transcripts is to regulate protein-coding mRNAs transcribed from parental genes by competing for microRNAs (Ebert and Sharp 2010; Poliseno et al. 2010).

Identification of novel, uncharacterized elements, such as new genes should be confirmed experimentally. Unfortunately, most researchers limit their genome-wide studies to computational analyses and functionality of the retrocopy was very rarely confirmed by direct experiments using molecular biology techniques. One of

the rare examples of experimental validation is the determination of seven retrogene-parental gene pairs expression patterns using RT-PCR (reverse transcription PCR) in human (Marques et al. 2005) where results provided strong evidences for testis-specific expression patterns for retrogenes, while parental genes presented almost ubiquitous expression. Another experimental investigation of retrogenes using RT-PCR was performed for several genes in *Drosophila melanogaster* and it revealed that most new retrogenes are expressed in one or more analyzed tissues (Betran et al. 2002a, b).

16.4 Role of Retrogenes in Shaping Interspecies Differences

The differences between species cannot be explained just by point mutations and small indels as the evolution via these mechanisms is relatively slow. By contrast, the high number of retroposition observed in many genomes is causing quite rapid evolutionary changes. Therefore, retroposition has to be considered as one of the major players in formation of interspecies differences. Nevertheless, the number of systematic studies evaluating the impact of gene retroposition on species evolution is relatively low. However, even these selected studies show that the processes of retroposition, to a big extent, are species specific. Studies performed on kinases' retrogenes indicated that 97 kinase copies found in mice are all distinct from 107 retrocopies identified in the human genome (Caenepeel et al. 2004). The lack of orthologous retrogenes, demonstrated in this study, may not be very convincing as analysis considered only a tiny fraction of all retrocopies. However, Svensson et al. (2006) performed a genome-wide survey of functional pseudogenes in the human, mouse, and chimpanzee and found only two functional retrogenes conserved in the human and mouse genomes. The first large-scale comparative analysis of ribosomal protein pseudogenes in four mammalian genomes showed that among around 1500 retrocopies of ribosomal protein genes identified in chimpanzee genomes, 13 % are species specific. The same scientific group also discovered that only six ribosomal retrogenes are common for human and mouse (Balasubramanian et al. 2009). In another studies, performed on primates genomes, it was estimated that 57–76 functional retrogenes are specific for primate lineage and seven of them arose in the ancestor of hominoids (Marques et al. 2005).

Evidence for species-specific functional retrogenes comes not only from genome-wide analyses but most of all from single gene studies. We already mentioned a case of the mouse-specific retrogene *Rps23r1*, which reduces Alzheimer's beta-amyloid levels and tau phosphorylation (Zhang et al. 2009) and the primate-specific *PYDC2* retrogene involved in regulating NF-kappaB activity and inflammasome formation (Bedoya et al. 2007; Atianand et al. 2011). Other examples of primate-specific functional retrogenes include brain-specific isotype of the glutamate dehydrogenase (*GLUD2*) gene (Burki and Kaessmann 2004) and brain- and testis-specific *CDC14Bretro* gene, which originated from *CDC14B* cell cycle gene (Rosso et al. 2008). Recently, a unique mechanism of functional

retrocopy origination was described by Babushok et al. (2007). A gene called *PIPSL* arose from the combination of functional domains at the RNA level from distinct genes. The resulting chimera was then reverse transcribed and integrated into the genome. The *PIPSL* gene, present only in hominoids, encodes a protein combining the lipid kinase domain of *PIP5K1A* and the ubiquitin-binding motifs of *PSMD4* and is transcribed specifically in the testis in humans and chimpanzees.

Important contribution of retrogenes to organismal differentiation is also visible at the population level. An elegant example of retrogene diversifying dogs was mentioned above in a study performed by Parker et al. (2009) who showed that all short legged breeds of dogs carry *fgf4* retrogene. Robertson et al. (2006) found strain-specific retrogenes of *Nanog* in mouse. While *NanogPc* is present in 129/Ola and 129/Sv but not in C57/B16 or CBA, *NanogPd* exist in 129/Ola, 129/Sv, and CBA but not in C57/B16. A recent study on North American *Drosophila melanogaster* inbred lines revealed the first ever set of polymorphic retrogenes (Schrider et al. 2011). They found 34 retroCNVs (copy number variants) and estimated that any two gametes in the North American population of fruit fly differ in the presence or absence of six retrogenes, which accounts for approximately 13 % of gene copy-number heterozygosity.

16.5 Retrogenes Evolution and Gain of Function

For a long time it was assumed that retroposed gene copies are nonfunctional because in the process of duplication they do not inherit parental regulatory elements and that is why they lack expression potential. Therefore, it was expected that molecular evolution of retrogenes is selectively neutral and these genes evolve relatively quickly. This assumption is confirmed by some empirical data showing comparison with DNA-based duplicates (Cusack and Wolfe 2007). The degree and type of selection can be measured by the ratio of non-synonymous substitutions (K_A) to synonymous substitutions (K_S). Under neutral evolution $K_A = K_S$, deviation of K_A from K_S may be due to positive selection when the K_A/K_S is >1, or purifying selection when $K_A/K_S < 1$. The majority of retrogenes are in the state of "relaxed" selection as it was shown by Yu and coworkers (2007). However, they also discovered that some human retrogenes are undergoing a non-neutral evolution. Retrogenes under a strong purifying selection were also identified by several other groups (Betran et al. 2002a, b; Svensson et al. 2006; Vinckenbosch et al. 2006; Chen et al. 2011). First reports of functional retrogenes were published in the 1980s (Soares et al. 1985; McCarrey 1987) and a number of genome-wide studies showed that many genes are under purifying or positive selection and therefore may be functional. Nowadays, this so-called "junk DNA" is considered to be important for the evolution of species-specific phenotypes as it provides raw material for the emergence of genes with new functions.

A new gene needs to acquire a core promoter and other regulatory elements to become expressed. One way of obtaining such would be to "hitch-hike" on

regulatory elements of other genes in their vicinity. A number of cases have been reported in which retrogenes are located in the intron of another gene and are transcribed together with a host gene (Long and Langley 1993; Bradley et al. 2004; Vinckenbosch et al. 2006). It was also observed that transcribed retrocopies are often at a very short distance from other genes. This suggests that their transcription may be facilitated by the open chromatin and regulatory machinery of these neighboring genes (Vinckenbosch et al. 2006). Retrogenes may also be transcribed from CpG-rich promoters or CpG enriched sequences located at a substantial distance. In this case the gap between retrogene and the promoter can be brigded by new 5' untranslated exons that arose during the process of promoter acquisition (Kundu and Rao 1999; Vinckenbosch et al. 2006; Makalowska and Szczesniak, unpublished data). Similarly, nearby or remote promoters from retrotransposable elements can be captured and directly used (Zaiss and Kloetzel 1999; Makałowski 2000). Interestingly, although retrogenes are not expected to inherit parental promoters, there is an evidence for such events (Soares et al. 1985; McCarrey 1987). This may happen when parental gene is transcribed from promoters, which have multiple transcriptional start sites (TSSs). If the retrogene arise from a transcript with a TSS located upstream, the mRNA from which retrogene originated may carry downstream promoter and TSS with capacity to stimulate transcription.

In the early studies of duplicates' evolution it has been postulated that it is natural that one of the duplicates from a pair, after accumulating mutation always become nonfunctional (Haldane 1933; Fisher 1935). Consequently, all retrocopies would be expected to transform into pseudogenes. However, gene duplication is also thought to be an important evolutionary process as it relaxes some constrains and opens new evolutionary pathways. Indeed, although a majority of gene duplicates are in the state of a "relaxed" selection and remain "dormant", many become functional. Nei was the first to propose that gene duplication could promote adaptation and while one of the copies keeps the original function of the gene, the other one is free to examine the sequence space and acquire new function (Nei 1969). This process is called "neofunctionalization". Alternatively, after duplication two genes would maintain the ancestral function; however, they would demonstrate different spatio-temporal expression patterns. This process was named "subfunctionalization" (Force et al. 1999). As recent studies on *Drosophila* (Krasnov et al. 2005) and our study on human genes (Ciomborowska et al., unpublished data) showed, there is also another possibility, the retrogene may replace its parent, which gets deleted or pseudogenized.

It was hypothetized that functional retrocopies might be mainly the result of subfunctionalization (Force et al. 1999) and there is a wealth of examples of retrogenes sharing the function with their parents. Nevertheless, there is a growing evidence for retrocopies obtaining brand new functions. Recently, a non-coding RNA expressed from human retrogene was reported to regulate transcript of its parental gene by acting as a decoy for miRNA that binds to common sites in the 3' untranslated region (Poliseno et al. 2010). It was proposed that the general activity of this retrogene is competing for miRNAs and therefore, the level of retrogene

expression regulates the level of protein encoded by target mRNA. Lately, Rosso et al. (2008) demonstrated a novel mode for the emergence of a new gene function. They showed that *CDC14Bretro* that originated through retroposition in the hominoid ancestor, by virtue of amino acids substitutions in distinct protein regions, shifted the subcellular localization from the association with microtubules to an association with endoplasmic reticulum. This process of subcellular adaptation was termed neolocalization.

16.6 Expression Pattern and "out of the X" Hypothesis

It has been suggested that retrogenes tend to exhibit an expression bias toward the testes and a number of studies confirmed this testis-specific expression patterns in both mammals and *Drosophila* (Betran et al. 2002a, b; Vinckenbosch et al. 2006; Bai et al. 2007; Yu et al. 2007; Pan and Zhang 2009). In contrast, their parental genes have a general tendency to be broadly expressed (Marques et al. 2005; Bai et al. 2007; Potrzebowski et al. 2008). There are a few possible hypotheses interpreting this phenomenon. The first explanation links this specific expression pattern to the fact that in meiotic and post-meiotic spermatogenic cells are in a state of hypertranscription. This hypertranscription, possible due to the various modifications of chromatin, enables transcription of DNA that usually is not transcribed; therefore, it may also facilitates the transcription of retrogenes in testis (Marques et al. 2005; Chen et al. 2011). Promoters of these genes, or some of them, could be later enhanced and retrocopy could evolve in new gene with testis-specific expression pattern, which potentially could adopt functions in somatic tissues (Kaessmann 2010). The second hypothesis is based on the idea of preferential insertion of retrocopies into open and actively transcribed chromatin (Fontanillas et al. 2007). Considering that retroposition occurs in the germline, it may be expected that retrocopies are mostly located near to germline expressed genes and this would make possible transcription of the retrocopy in the germ line (Kaessmann et al. 2009).

Another hypothesis links the testis-specific expression of retrogenes with the "escape" of genes from the X chromosome. Chromosomal gene movements have been studied in various taxonomic groups including mammalian genomes (Emerson et al. 2004; Marques et al. 2005; Vinckenbosch et al. 2006; Potrzebowski et al. 2008), vertebrates (Pan and Zhang 2009), chordates (Chen et al. 2011), fruit fly, (Betran et al. 2002a, b; Bai et al. 2007; Metta and Schlotterer 2010) and recently in mosquito, (Baker and Russell 2011) chicken, and silkworm (Toups et al. 2011). Most of the studies considering XX/XY system show that among functional retrogenes there is an excess of those originated from genes located on the X chromosome (Betran et al. 2002a, b; Emerson et al. 2004). It was proposed that in mammals retrocopies originated from X-linked genes work as autosomal counterparts of their parents which can be inactivated during male meiotic sex chromosome inactivation (MSCI) (Marques et al. 2005; Vinckenbosch et al. 2006;

Potrzebowski et al. 2008; Kaessmann et al. 2009). This phenomenon was probably shaped by natural selection in order to compensate for silenced parental genes (Marques et al. 2005; Vinckenbosch et al. 2006) and expression analyses seem to support this hypothesis (Potrzebowski et al. 2008).

Out-of-X retroposition was originally identified in fruit fly (Betran et al. 2002a, b) but this phenomenon was explained in this species by different hypotheses. The first hypothesis suggests that there is a disproportion of X-linked genes that causes nonrandom generation of retrogenes. The second explanation postulates negative selection as driving force against new genes inserted in X chromosome and, at the same time, positive selection can play a significant role in favoring retrogenes moved to autosomes. The third option is related to possible differences in recombination rate between autosomes and sex chromosomes (Betran et al. 2002a, b). However, postulate about advantages coming from having required functions independently on parents inactivation, was also noticed for *Drosophila* (Bai et al. 2007). It is worth mentioning that excess of movement was also detected in the opposite direction, i.e., onto X chromosome. However, this phenomenon was observed in mammals (Emerson et al. 2004; Potrzebowski et al. 2010) but not in *Drosophila* (Betran et al. 2002a, b; Meisel et al. 2009) or mosquito (Toups and Hahn 2010).

The excess of retrogenes originated from sex chromosomes and their testis-specific expression patterns are most probably specific for XX/XY systems only. Toups et al. (2011) studied retrogenes expression in chicken and silkworm and were unable to identify any overabundance of retroposed genes that had testis-biased expression. One explanation for this observation may be related to the fact identified by them that retrogenes were relatively old and previous analyses have found that testis-biased expression is lost over time (Vinckenbosch et al. 2006). Another possibility is that since they studied ZZ/ZW systems it could be anticipated that genes would be moving out of chromosome Z and would exhibit ovary-biased expression. However, they did not find either ovary-biased expression of studied retrogenes or any excess of movement out of Z chromosomome in either birds or lepidopterans.

There is also another type of selective pressure, which may have an impact on retrogenes exportation-sexual antagonism. It means that some genetic changes are preferred by only one sex, so for example genes that are meaningful for males would be more often found on autosomes than on X chromosome (which can be described as more "feminized") (Wu and Xu 2003; Emerson et al. 2004). Nevertheless, this mechanism previously considered as an alternative for the MSCI, plays a rather less important role in mammals and more significant in fruit fly. To sum up, driving forces related to the out-of-X phenomenon seems to work in a different way in insects and in other animals (Potrzebowski et al. 2008).

Testis-specific expression pattern, however, even in mammals is not uniform. Some researchers observed that in selected primate species there is statistically significant bias for retrogenes to be expressed in brain (Marques et al. 2005; Chen et al. 2011). McCole et al. (2011) analyzed four imprinted retrogenes and all of them had broad expression patterns. The results obtained in our laboratory also do not confirm testis-specific preferences. We performed a real-time PCR for 29 human

retrogenes which replaced their parental gene. Expression analysis in 16 human cDNA libraries including testis showed that a vast majority of investigated retrogenes, 27 out of 29, were detected in all samples and not a single retrogene revealed a testis-specific expression pattern (Ciomborowska et al. unpublished data).

16.7 Conclusions

Retrogenes for a long time were thought to be not functional and evolutionarily not important. However, multiple reports show that many mRNA-derived duplicates gain the function and become not only important but also essential bona fide genes. Investigations of retrogenes and their evolution brought a lot of compelling results. Based on these studies, we learned about very unique ways of gain of functionality and new genes origination, specific mechanisms of promoter recruitment, gene movements, and even the evolution of sex chromosomes. We also discovered how big contribution retrogenes had in the process of speciation, and in the process of acquiring a function retrogenes can move toward subfunctionalization, neofunctionalization, or neolocalization. All these discoveries made mRNA-based gene duplicates even more exciting subject of studies. There is still a lot to uncover about these puzzling retrocopies and although studies of retrogenes are quite challenging we believe that they are worth much of undertaking and that they will bring a lot of fascinating discoveries.

Glossary

Retrogene Expressed and functional retrocopy; product of multi-exon parental gene mRNA retroposition

Retrocopy product of multi-exon parental gene mRNA retroposition

Retropseudogene Non-functional retrocopy, usually with degenerative mutations and genetic defects which become silenced short after retroposition

Parental Gene Multi-exon gene gene which gives birth to retrocopy, works as a source of mRNA during retroposition

Duplication Appearance of gene copies which are heritable

Retroposition A mechanism in which mRNA of parental gene is reversely transcribed and new retrocopy is incorporated in new genomic positions (also known as RNA-based duplication or retroduplication)

Homologs Genes which have common origin

Paralogs Homologous genes related because of duplication

Orthologs Homologous genes originating from a single ancestral gene in the last common ancestor of the compared genomes, genes related through speciation

Subfunctionalization Subdivision of function between retrocopy and parental gene as a result of accumulation of degenerative mutation in duplicate

Neofunctionalization The development of new function in duplicated gene as a result of the accumulation of neutral mutations

MSCI Meiotic sex chromosome inactivation-process in which genes related to sex development are transcriptionally silenced

Ka/Ks ratio Ratio between two values-a (the rate of substitution at non-synonymous sites in nucleotide sequence) and Ks (the rate of substitution at synonymous sites). Ka/Ks is often used to deduce type of the selection. Ka/Ks < 1 functional constraint, Ka/Ks $= 1$ lack of functional constraint; Ka/Ks > 1 positive Darwinian selection.

References

Altschul SF, Gish W et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
Atianand MK, Fuchs T et al (2011) Recent evolution of the NF-kappaB and inflammasome regulating protein POP2 in primates. BMC Evol Biol 11:56
Babushok DV, Ohshima K et al (2007) A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. Genome Res 17:1129–1138
Baertsch R, Diekhans M et al (2008) Retrocopy contributions to the evolution of the human genome. BMC Genomics 9(1):466
Bai Y, Casola C et al (2007) Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol 8(1):R11
Baker DA, Russell S (2011) Role of testis-specific gene expression in sex-chromosome evolution of Anopheles gambiae. Genetics 189(3):1117–1120
Balakirev ES, Ayala FJ (2003) PSEUDOGENES: Are They "Junk" or functional DNA? Annu Rev Genet 37:123–151
Balasubramanian S, Zheng D et al (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biol 10(1):R2
Bedoya F, Sandler LL et al (2007) Pyrin-only protein 2 modulates NF-kappaB and disrupts ASC:CLR interactions. J Immunol 178(6):3837–3845
Betran E, Thornton K et al (2002a) Retroposed new genes out of the X in Drosophila. Genome Res 12(12):1854–1859
Betran E, Wang W et al (2002b) Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. Mol Biol Evol 19(5):654–663
Bradley J, Baltus A et al (2004) An X-to-autosome retrogene is required for spermatogenesis in mice. Nat Genet 36(8):872–876
Bridges C B (1936) The bar "Gene" a duplication science 83(2148):210–211
Brosius J (1991) Retroposons–seeds of evolution. Science 251(4995):753
Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115–134
Brosius J (2003) The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118(2–3):99–116

Burch JB, Davis DL et al (1993) Chicken repeat 1 elements contain a pol-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. Proc Natl Acad Sci U S A 90(17):8199–8203

Burki F, Kaessmann H (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet 36(10):1061–1063

Caenepeel S, Charydczak G et al (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci U S A 101(32):11707–11712

Chen S, Zhang YE et al (2010) New genes in Drosophila quickly become essential. Science 330(6011):1682–1685

Chen M, Zou M et al (2011) Evolutionary patterns of RNA-based duplication in non-mammalian chordates. PLoS ONE 6(7):e21466

Crainey JL, Garvey CF et al (2005) The origin and evolution of mosquito APE retroposons. Mol Biol Evol 22(11):2190–2197

Cusack BP, Wolfe KH (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol 24(3):679–686

Devor EJ (2006) Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes. J Hered 97(2):186–190

Ebert MS, Sharp PA (2010) Emerging roles for natural microRNA sponges. Curr Biol 20(19):R858–R861

Eickbush TH, Furano AV (2002) Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev 12(6):669–674

Emerson JJ, Kaessmann H et al (2004) Extensive gene traffic on the mammalian X chromosome. Science 303:537–540

Esnault C, Maestre J et al (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24(4):363–367

Fisher R (1935) The sheltering of lethals. Am Nat 69:446–455

Fontanillas P, Hartl DL et al (2007) Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. PLoS Genet 3(11):e210

Force A, Lynch M et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151(4):1531–1545

Goncalves I, Duret L et al (2000) Nature and structure of human genes that generate retropseudogenes. Genome Res 10(5):672–678

Haas NB, Grabowski JM et al (2001) Subfamilies of CR1 non-LTR retrotransposons have different 5'UTR sequences but are otherwise conserved. Gene 265(1–2):175–183

Haldane J (1933) The part played by recurrent mutation in evolution. Am Nat 67:5–19

Harrison PM, Hegyi H et al (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 12(2):272–280

Harrison PM et al (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res 33:2374–2383

Hollis GF, Hieter PA et al (1982) Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature 296(5855):321–325

Jacq (1977) A pseudogene structure in 5S DNA of Xenopus laevis. Cell 12: 109–120

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20(10):1313–1326

Kaessmann H, Vinckenbosch N et al (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10(1):19–31

Karin M, Richards RI (1982) Human metallothionein genes–primary structure of the metallothionein-II gene and a related processed gene. Nature 299(5886):797–802

Krasnov AN, Kurshakova MM et al (2005) A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic Acids Res 33(20):6654–6661

Kundu TK, Rao MR (1999) CpG islands in chromatin organization and gene expression. J Biochem 125(2):217–222

Long M (2001) Evolution of novel genes. Curr Opin Genet Dev 11(6):673–680
Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260(5104):91–95
Luo Y, Li S (2007) Genome-wide analyses of retrogenes derived from the human box H/ACA snoRNAs. Nucleic Acids Res 35(2):559–571
Makałowski W (2000) Genomic scrap yard: how genomes utilize all that junk. Gene 259:61–67
Marques AC, Dupanloup I et al (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 3(11):e357
McCarrey JR (1987) Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. Gene 61(3):291–298
McCarrey JR, Thomas K (1987) Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature 326(6112):501–505
McCole RB, Loughran NB et al (2011) A case-by-case evolutionary analysis of four imprinted retrogenes. Evolution 65(5):1413–1427
Meisel RP, Han MV et al (2009) A complex suite of forces drives gene traffic from Drosophila X chromosomes. Genome Biol Evol 1:176–188
Metta M, Schlotterer C (2010) Non-random genomic integration: an intrinsic property of retrogenes in Drosophila? BMC Evol Biol 10:114
Mighell AJ, Smith NR et al (2000) Vertebrate pseudogenes. FEBS Lett 468(2–3):109–114
Nei M (1969) Gene duplication and nucleotide substitution in evolution. Nature 221(5175):40–42
Nozawa M, Aotsuka T et al (2005) A novel chimeric gene, siren, with retroposed promoter sequence in the Drosophila bipectinata complex. Genetics 171(4):1719–1727
Ohno S (1970) Evolution by gene duplication. Springer, Berlin
Ohshima K, Igarashi K (2010) Inference for the initial stage of domain shuffling: tracing the evolutionary fate of the PIPSL retrogene in hominoids. Mol Biol Evol 27(11):2522–2533
Pan D, Zhang L (2009) Burst of young retrogenes and independent retrogene formation in mammals. PLoS ONE 4(3):5040
Parker HG, VonHoldt BM et al (2009) An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325(5943):995–998
Poliseno L, Salmena L et al (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465(7301):1033–1038
Potrzebowski L, Vinckenbosch N et al (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6(4):e80
Potrzebowski L, Vinckenbosch N et al (2010) The emergence of new genes on the young therian X. Trends Genet 26(1):1–4
Prendergast GC (2001) Actin' up: RhoB in cancer and apoptosis. Nat Rev Cancer 1(2):162–168
Robertson M et al (2006) Nanog retrotransposed genes with functionally conserved open reading frames. Mamm Genome 17:732–743
Rosso L, Marques AC et al (2008) Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. PLoS Biol 6(6):e140
Sakai H, Koyanagi KO et al (2007) Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene 389(2):196–203
Schrider DR, Stevens K et al (2011) Genome-wide analysis of retrogene polymorphisms in Drosophila melanogaster. Genome Res 21(12):2087–2095
Soares MB, Schon E et al (1985) RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol 5(8):2090–2103
Svensson O, Arvestad L et al (2006) Genome-wide survey for biologically functional pseudogenes. PLoS Comput Biol 2(5):e46
Szczesniak MW, Ciomborowska J et al (2011) Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol Biol Evol 28(1):33–37
Toups MA, Hahn MW (2010) Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. Genetics 186(2):763–766
Toups MA, Pease JB et al (2011) no excess gene movement is detected off the avian or lepidopteran z chromosome. Genome Biol Evol 3:1381–1390

Tsujikawa M, Kurahashi H et al (1999) Identification of the gene responsible for gelatinous drop-like corneal dystrophy. Nat Genet 21(4):420–423

Ueda S, Nakai S et al (1982) Long terminal repeat-like elements flank a human immunoglobulin epsilon pseudogene that lacks introns. EMBO J 1(12):1539–1544

Vinckenbosch N, Dupanloup I et al (2006) Evolutionary fate of retroposed gene copies in the human genome. PNAS 103(9):3220–3225

Volff JN, Brosius J (2007) Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. Genome Dyn 3:175–190

Weber MJ (2006) Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet 2(12):e205

Wei W, Gilbert N et al (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21(4):1429–1439

Weiner AM, Deininger PL et al (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661

Wu CI, Xu EY (2003) Sexual antagonism and X inactivation–the SAXI hypothesis. Trends Genet 19(5):243–247

Yano Y, Saito R et al (2004) A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. J Mol Med (Berl) 82(7):414–422

Yu Z, Morais D et al (2007) Analysis of the role of retrotransposition in gene evolution in vertebrates. BMC Bioinformatics 8:308

Zaiss DM, Kloetzel PM (1999) A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. J Mol Biol 287(5):829–835

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18(6):292–298

Zhang Z, Harrison P et al (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 12(10):1466–1482

Zhang Z, Harrison PM et al (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13(12):2541–2558

Zhang YW, Liu S et al (2009) A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation. Neuron 64(3):328–340

Zhang YE, Vibranovski MD et al (2011) A cautionary note for the retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. Bioinformatics 27(13):1749–1753

Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 23(5):219–224

Szczesniak M., Ciomborowska J., Nowak W., Rogozin I.B., Makałowska I.

Primate and rodent specific intron gains and the origin of retrogenes with splice variants

Molecular Biology and Evolution 2011 Jan; 28(1):33-7

Molecular Biology and Evolution

Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants

Michał W Szcześniak,†[,1] Joanna Ciomborowska,†[,1] Witold Nowak,[2] Igor B Rogozin,[3] and Izabela Makałowska*[,1]

[1]Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

[2]Laboratory of Molecular Techniques, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

†These authors contributed equally

***Corresponding author**: E-mail: izabel@amu.edu.pl.

Associate editor: Helen Piontkivska

Abstract

Retroposition, a leading mechanism for gene duplication, is an important process shaping the evolution of genomes. Retrogenes are also involved in the gene structure evolution as a major player in the process of intron deletion. Here, we demonstrate the role of retrogenes in intron gain in mammals. We identified one case of "intronization," the transformation of exonic sequences into an intron, in the primate specific retrogene *RNF113B* and two independent "intronization" events in the retrogene *DCAF12L2*, one in the common ancestor of primates and rodents and another one in the rodent lineage. Intron gain resulted from the origin of new splice variants, and both genes have two transcript forms, one with retained intron and one with the intron spliced out. Evolution of these genes, especially *RNF113B*, has been very dynamic and has been accompanied by several additional events including parental gene loss, secondary retroposition, and exaptation of transposable elements.

Key words: intron gain, gene structure evolution, splice variant, RNF113, DCAF12.

The majority of protein-coding genes in eukaryotes are interrupted by introns that are removed from the pre-mRNA by a RNA–protein complex called the spliceosome (Cavalier-Smith 1985; Crick 1979). Introns and the splicing machinery have been found in all eukaryotic species with fully sequenced genomes (Chow et al. 1977; Roy and Gilbert 2006). Comparative genomic studies have revealed striking conservation of intron positions in distant eukaryotes such as animals and plants (Fedorov et al. 2002; Rogozin et al. 2003; Carmel et al. 2007). On the other hand, many genome-wide comparisons of eukaryotic species demonstrated multiple intron losses and intron gains (Roy et al. 2003; Cho et al. 2004; Qiu et al. 2004; Coulombe-Huntington and Majewski 2007b; Li et al. 2009). However, it was found that intron gain is a very rare event in vertebrate evolution (Loh et al. 2007) and no intron gains into intact conserved coding regions of mammalian genes are known (Roy et al. 2003; Coulombe-Huntington and Majewski 2007a).

Comparative gene structure studies have not revealed any intron gain into existing exons in mammals. The only reported new introns were acquired, by and large, by either a fusion of retrogene with host genes or de novo from the genomic environment as a result of new exon capture (O'Neill et al. 1998; Vinckenbosch et al. 2006; Sela et al. 2007; Baertsch et al. 2008; Fablet et al. 2009). Here, we report two retrogenes, *RNF113B* and *DCAF12*, where the exon sequence was split by creation of a new intron as the result of mutations and emergence of new splice sites. The introns discovered by us represent cases of intron creation via recruitment of exonic sequence (intronization) proposed by Irimia et al. (2008) and Lahn and Page (1999).

Evolution of Introns

RNF113A is a retrogene encoding a ring finger protein of unknown function and is present in the genomes of all vertebrates. Interestingly, in mammalian genomes, only intronless copy exist, whereas in all other vertebrates, a ten-exon parental gene is present and no retrogenes were detected. Genomic sequence analysis showed that there are two copies of *RNF113* in primates, rodents, carnivores, and even-toed ungulates and only one in the genomes of the other mammals we studied. The first copy of *RNF113* was retroposed into the intronic region of *NDUFA1* gene in the genome of the mammalian ancestor. Following the retroposition, the parental gene was lost. This likely took place before the divergence of Prototheria (Monotremes) and Theria (Marsupials and Placentals) because in the genomes of all species representing these lineages, the multiexon form of *RNF113* is absent. After the mammalian radiation the *RNF113A* retrogene was duplicated, by retropositions or segmental duplications, in several lineages. Analysis of genomic locations of these copies suggests that the duplication events were independent in each lineage. For example, in rodents, the *RNF113* copy (*RNF113A2*) was inserted into an intron of

Letter

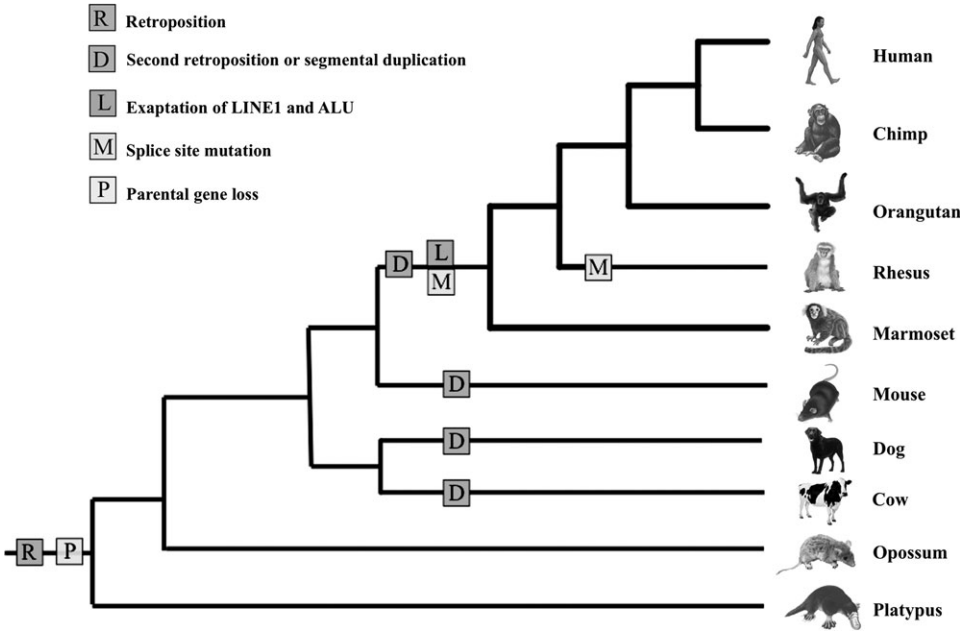


Fig. 1. Schematic tree representing major events during evolution of *RNF113* gene in mammalian lineage. Color version of the figure can be found in supplementary Data (Supplementary Material online).

the *2900006K08Rik* gene, whereas the primate specific gene, *RNF113B*, was copied into an intron of the *FARP1* gene. The primate specific duplication happened before Old World Monkeys and New World Monkeys diverged (fig. 1).

After the retroposition/duplication, the primate specific *RNF113B* gene underwent rapid evolution including intron gain. The presence of the intron is surprising, however, it is supported by several GenBank mRNA sequences (accession numbers: AF539427, BC025388, and BC017585). To confirm the existence of the intron and learn about its origin, we compared *RNF113B* sequences from available primate genomes (human, marmoset, macaque, orangutan, and chimpanzee) with sequences of other mammalian *RNF113A* genes. Sequence alignment revealed that the intron of *RNF113B* is not a de novo insertion but rather originated from the exonic sequence (fig. 2*a*). A double point mutation, AG → GT, generated the donor site (fig. 2*a*). The origin of acceptor site is not so clear. One possibility is that a point mutation, GG → AG, created acceptor site. Another option is that the acceptor site was brought during the exonization of L1 element, merged at the 3′ end of *RNF113B* (fig. 2*b*). The newly generated splice sites together with the branch site and the polypyrimidine tract likely enabled recognition of the new intron by the U2 spliceosome (fig. 2*a*). The 105 bp intron contains 59 nucleotides of previously coding sequence and 46 nucleotides from the 3′ UTR.

Generation of splice sites most probably occurred in the primate specific *RNF113B* copy since neither human *RNF113A* gene, which gave a rise to primate *RNF113B*, nor *RNF113A* genes from other mammals have AG or GT at the donor and acceptor positions. Splicing signals were formed before the Old World Monkeys and New Monkeys split. Interestingly, loss of the splicing boundaries subsequently converted the intron into a "retained intron" in some primates. In rhesus, for example, acceptor was lost due to a point mutation (AG → AA change) (fig. 1*b*).

The creation of splicing signals was accompanied not only by exaptation of an L1 element but also by exonization of an Alu element. The L1 element inserted within the 3′ end of the gene could have contributed the acceptor site and provided a new polyA signal used for the new splice variant (fig. 2*a*). The complete AluSx element transposed upstream the gene was exapted at the 5′ end and most probably delivered some regulatory elements.

Sequencing of the human *RNF113B* cDNA using primers flanking the intronic sequence revealed that *RNF113B* produces two variant transcripts. One variant has two exons, as described above, and the other one is a single exon transcript similar to *RNF113A*. Consequently, most primates have three transcripts of *RNF113*: one from the *RNF113A* retrogene and two from the *RNF113B* (fig. 2*b*). Rodents, cow, and dog have two transcripts, each coming from different copy of *RNF113*, and all other mammals have only one *RNF113* transcript. The presence of the splice variants in the retrogene is very surprising and has only been reported once before (Lahn and Page 1999).

A second case involves *DCAF12* (DDB1 and CUL4 associated factor 12), which encodes a WD repeat-containing protein that interacts with the COP9 signalosome (Jin et al. 2006). Although the gene is present in vertebrate and insect genomes, only placental mammals have retrocopies of this gene. One copy, *DCAF12L2*, has the same location in all placental mammals and therefore most likely was retroposed in the placental mammals ancestor. Another copy, *DCAF12L1*, is present only in Euarchontoglires (a clade which includes rodents and primates). It likely emerged as a result of tandem duplication of *DCAF12L2* as it is located next to the *DCAG12L1* gene. There were two events that changed the

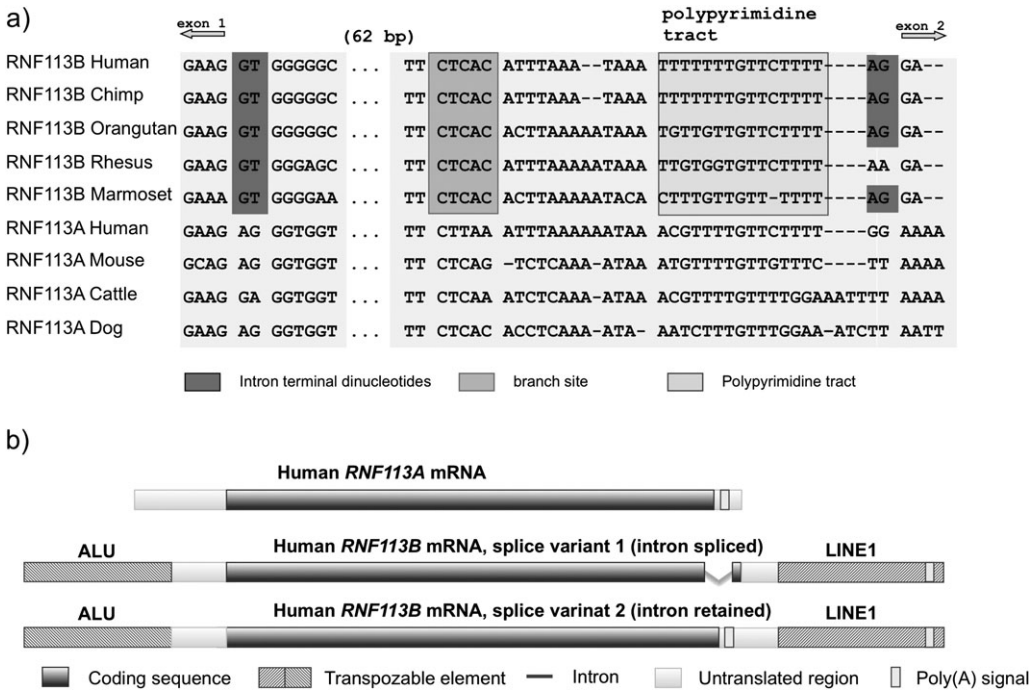


Fig. 2. (*a*) Alignment of mammalian *RNF113A* and primate *RNF113B* genomic sequences at the acceptor and donor sites. (*b*) Structure of human *RNF113A* mRNA and two splice variants of *RNF113B*. Color version of the figure can be found in supplementary Data (Supplementary Material online).

splicing pattern in *DCAF12L2*. First, an intronization event occurred in the common ancestor of primates and rodents. Second, an alternative donor site emerged in rodents only (fig. 3). The limited available data and sequence divergence make any conclusions in regard to the exact pattern of splice site evolution infeasible. However, there is convincing experimental evidence confirming both splicing events (fig. 3): splicing at the shared rodent–primate intron, boundaries are confirmed by two expressed sequence tags (ESTs) (AK034343 and AK047360), and usage of the rodent

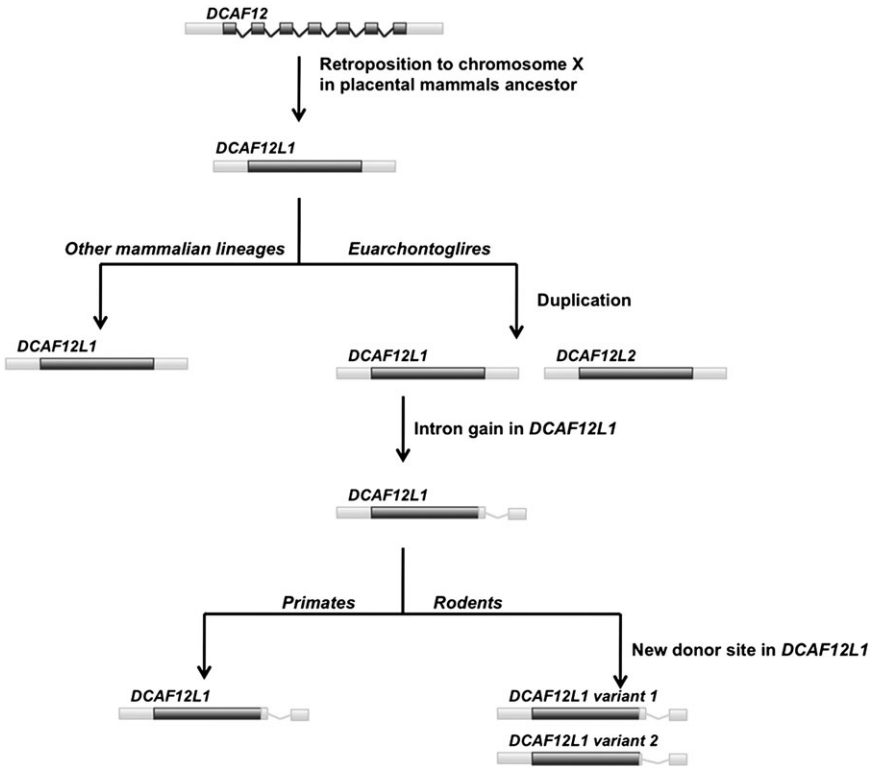


Fig. 3. Pattern of *DCAF12* duplication and "intronization" events in mammalian genomes. Color version of the figure can be found in supplementary Data (Supplementary Material online).

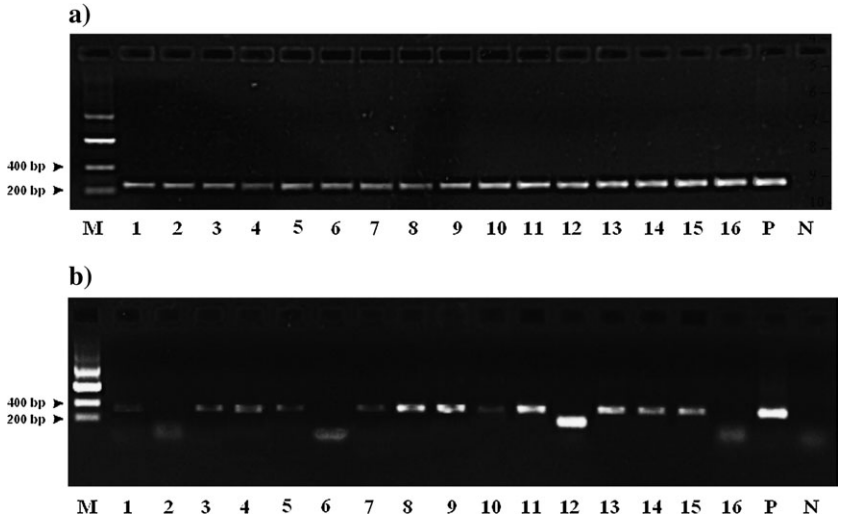


Fig. 4. Expression pattern of *RNF113A* and two forms of *RNF113B* (195 bp product with intron spliced; 295 bp product-form with intron retained) in 16 human tissues: 1: heart, 2: brain, 3: placenta, 4: lung, 5: liver, 6: skeletal muscle, 7: kidney, 8: pancreas, 9: spleen, 10: thymus, 11: prostate, 12: testis, 13: ovary, 14: small intestine w/o mucosal lining, 15: colon, 16: peripheral leukocytes, P: genomic DNA, and N: water.

alternative donor site is confirmed by four ESTs (AK038557, BC068319, AK034472, and AK039767).

Retrogene Expression

Numerous studies revealed a tendency of retrogenes to be expressed exclusively in testis. It was suggested that the hypertranscription present in the meiotic and postmeiotic spermatogenic cells makes possible transcription of DNA that is usually not transcribed. This may facilitate transcription of retrocopies in the testis during their early evolution (reviewed in (Kaessmann et al. 2009). Another hypothesis explains the high expression of retrogenes in testis by the fact that these are, in most cases, retrocopies of spermatogenesis-related genes located on the X chromosome. Because the X chromosome is inactivated during meiosis, retroposition to autosomes enables escape from inactivation and expression during spermatogenesis (Turner 2007).

The retroposition of both genes studied here, *RNF113* and *DCAF12*, was in the opposite direction, from autosomes to chromosome X. In the case of *RNF113*, the parental gene is detectable by sequence similarity as an apparent pseudogene on chromosome 9. The parental multiexon *DCAF12* gene is coincidentally also located on chromosome 9. *RNF113A* and both *DCAF12* retrogenes are on chromosome X. We surveyed the expression pattern of all human *RNF113* transcripts (one from *RNF113A* and two from *RNF113B*) in 16 human tissues (fig. 4) (for methods, see Supplementary Material online). *RNF113A* was expressed in all studied tissues, including testes. Interestingly, *RNF113B* exhibited tissue-specific splicing; while the unspliced form of *RNF113B* was expressed in all tissues but testis, the spliced variant was expressed in testis, prostate, thymus, and lung. Both *RNF113B* splice variants were present in thymus, prostate, and lung, but in all of these tissues, the form with the intron spliced out had much lower expression level than the single exon primary form. Relatively high expression of the new form of *RNF113B*, form with the intron spliced out, was observed only in testis.

According to the EST data, the human *DCAF12* gene is widely expressed. EST sequences present in the dbEST database represent almost 40 libraries and show the highest expression in testis and trachea. The retrogene *DCAF12L1* is expressed only in kidney and testis and a second human retrogene, *DCAF12L2*, is expressed in eye and testis. Therefore, both retrogenes show very different expression patterns than their parental genes, with very limited and low expression level and notable expression in testis.

Conclusions

Retroposition, a major mechanism for gene duplication, is an important process shaping the evolution of genomes (Brosius 1991; Marques et al. 2005). Our study confirms the unusual role of retrogenes in shaping the genomes and underscores the importance of mobile elements in evolution. It also reveals that retrogenes may be responsible for a wealth of species-specific features including species-specfic introns and splice variants.

Previous analyses of introns in the vertebrate genomes did not uncover any intron gain in mammals (Roy et al. 2003). Our study clearly shows that creation of introns has occurred during mammalian evolution. The failure of previous studies to find intron gains can be explained by the fact that they were focused on different intron gain mechanisms and did not consider exon intronization. In addition, they looked at conserved among studied species genes, while we focused on young and in many cases lineage-specific retrogenes.

Interestingly, the retrogenes studied here exhibit testis-specific expression typically associated with genes escaping from the X chromosome despite their opposite history (retroposition from autosome to X). This biased expression pattern may not be exclusively related to meiotic genes, sex chromosome inactivation, and dosage compensation

(Marques et al. 2005; Vinckenbosch et al. 2006; Potrzebowski et al. 2008). The same pattern of high expression level in testis is observed in young, primate-specific splice variant of retrogene *RNF113B* as well as in both retroposed copies of *DCAF12* retroposed on the human X chromosome. The older, unspliced variant of *RNF113B*, as well as an earlier retrocopy *RNF113A*, displays more diverse expression patterns. Therefore, testis-specific expression could be a common feature of all newly evolved transcripts regardless of their chromosomal localization and may reflect a transcriptional noise due to "hypertranscription" in testis, facilitating the activation of new transcripts (Kleene et al. 1998).

The small number of observed intron gain in retrogenes may reflect that this is a rare event. Alternatively, the low number of observations could reflect the difficulties in identification of such events. One major complication lies in annotation problems and the common expectation that retrogenes do not have introns. Genome-wide comparative studies currently underway have already showed that intron gain in retrogenes could be more frequent than we expected but that annotations remain a major obstacle in uncovering this phenomenon.

Supplementary Material

Supplementary Data are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jurgen Brosius and two reviewers for their comments and insightful suggestions. I.B.R. was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/U.S. Department of Health and Human Services.

References

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics*. 9:466.

Brosius J. 1991. Retroposons–seeds of evolution. *Science* 251:753.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct models of intron dynamics in the evolution of eukaryotes. *Genome Res*. 17:1034–1044.

Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. *Nature* 315:283–284.

Cho S, Jin SW, Cohen A, Ellis RE. 2004. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Res*. 14:1207–1220.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12:1–8.

Coulombe-Huntington J, Majewski J. 2007b. Characterization of intron loss events in mammals. *Genome Res*. 17:23–32.

Coulombe-Huntington J, Majewski J. 2007a. Intron loss and gain in *Drosophila*. *Mol Biol Evol*. 24:2842–2850.

Crick F. 1979. Split genes and RNA splicing. *Science* 204:264–271.

Fablet M, Bueno M, Potrzebowski L, Kaessmann H. 2009. Evolutionary origin and functions of retrogene introns. *Mol Biol Evol*. 26:2147–2156.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci U S A*. 99:16128–16133.

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. 2008. Origin of introns by 'intronization' of exonic sequences. *Trends Genet*. 24:378–381.

Jin J, Arias EE, Chen J, Harper JW, Walter JC. 2006. A family of diverse Cul4-Ddb1-interacting proteins includes Cdt2, which is required for S phase destruction of the replication factor Cdt1. *Mol Cell*. 23:709–721.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet*. 10:19–31.

Kleene KC, Mulligan E, Steiger D, Donohue K, Mastrangelo MA. 1998. The mouse gene encoding the testis-specific isoform of Poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. *J Mol Evol*. 47:275–281.

Lahn BT, Page DC. 1999. Retroposition of autosomal mRNA yielded testis specific gene family on human Y chromosome. *Nat Genet*. 21:429–433.

Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in Daphnia populations. *Science* 326:1260–1262.

Loh Y-H, Brenner S, Venkatesh B. 2007. Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). *Mol Biol Evol*. 25:526–535.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*. 3:e357.

O'Neill RJ, Brennan FE, Delbridge ML, Crozier RH, Graves JA. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc Natl Acad Sci U S A*. 95: 1653–1657.

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol*. 6:e80.

Qiu WG, Schisler N, Stoltzfus A. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol*. 21:1252–1263.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*. 13:1512–1517.

Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A*. 100:7158–7162.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*. 7:211–221.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol*. 8:R127.

Turner JM. 2007. Meiotic sex chromosome inactivation. *Development* 134:1823–1831.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*. 103:3220–3225.

Szczesniak M., Ciomborowska J., Nowak W., Rogozin I.B., Makałowska I.

Primate and rodent specific intron gains and the origin of retrogenes with splice variants

Molecular Biology and Evolution 2011 Jan; 28(1):33-7

- supplementary materials -

Online Supplementary Material

Primate and rodent specific intron gains and the origin of retrogenes with splice variants

Michał Szcześniak[1#], Joanna Ciomborowska[1#], Witold Nowak[2], Igor Rogozin[3], and Izabela Makałowska[1]*

[1] Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
[2] Laboratory of Molecular Techniques, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
[3] National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA
[#] these authors contributed equally
* corresponding author: Izabela Makałowska (izabel@amu.edu.pl)

Materials and Methods

Retrogenes identification

Single exon retrogenes, *RNF113A* and *DCAF12L1,* were identified by genome wide comparison of chicken multi-exon genes with transcripts of single exon genes in human. Forms with introns were singled out during screening the entire human genome in order to identify all copies of a given gene. Both analyses were performed using BLAST and MegaBLAST. To make sure that annotated genes structures are not artifacts of genome assembly genomic sequences were aligned to raw data from Trace Archives Database at NCBI. In addition we identified EST sequences confirming all splice forms. We also verified expression of all *RNF113* forms by RT-PCR experiments.

Evolutionary analysis

To estimate time of retroposition and intron creation we screened available mammalian genomes using protein sequences coded by chicken multi-exon genes and human retrogenes. Presence of parental (multiexon) genes, orthologous retrogenes or copies created independently was established based on sequence similarity, gene structure (multiexon/single exon) and genomic location.

cDNA

Two normalized cDNA panels derived from male and female Caucasians (Clontech Laboratories, Inc) were used. First one, Human MTC™ Panel I included cDNA preparations from RNA from heart, brain, placenta, lung, liver, skeletal muscle, kidney and pancreas. The second panel, Human MTC™ Panel II, contained total cDNA from spleen, thymus, prostate, testis, ovary, small intestine w/o mucosal lining, colon with mucosa and peripheral leukocytes.

Primers

Two pairs of fluorescent PCR primers were used, one per gene. In the case of *RNF113B* the primers were designed to amplify a region containing the whole intron, while the primers for *RNF113A* (with no intron) were designed to give a product of similar size, with no special restraints for the amplified gene region. For *RNF113A* primers used were: 5'-TTTGAGCGCAGCCAGAAGATCC-3' (forward, labeled with VIC), 5'-AAGCCGCAGAAGCCAGTCTC-3' (reverse) and for *RNF113B*: 5'-GCTACATCTGTGACCAGCCAACC-3' (forward, labeled with NED), 5'-CAGGCATGGGATTGCAGGAAGAC-3' (reverse).

PCR reaction

To amplify cDNA fragments of interest of both *RNF113A* and *RNF113B* PCR reaction was performed, using HiFi Polymerase (Novazym). The PCR mixture contained the following components for 10 μl reactions: forward primer: 0.25 mM, reverse primer: 0.25 mM , dNTP: 80 μM, $MgCl_2$: 1.5 mM and HiFi Polymerase (1 U). In the case of RNF113B betaine (0.4 M) was added to increase PCR efficiency, as amplified region is GC-rich. PCR profile was as follows: 2 min at 98°C, followed by 38 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 20 s and elongation at 72°C for 30 s. Reactions were terminated by a final elongation at 72°C for 5 min. 2 ml of PCR product were run on 1.4% agarose gel for 30 min at 45 mA. The analysis of the size of PCR products was also performed on an automated sequencing apparatus (ABI 3130xl; POP-7 gel, filter set G5, array length 36 cm; GeneMapper software version 3.7) and the size of PCR products was determined in comparison with the internal GS600LIZ size standard

(Applied Biosystems). Sample PCR products (from liver and testis) were sequenced as well using the above sequencing apparatus. Further analysis was carried out on Peak Scanner™ Software v1.0.

Ciomborowska J., Rosikiewicz W., Szklarczyk D.,

Makałowski W., Makałowska I.

“Orphan” Retrogenes in the Human Genome

Molecular Biology and Evolution 2013 Feb; 30(2): 384-96

MBE Advance Access published November 20, 2012

"Orphan" Retrogenes in the Human Genome

Joanna Ciomborowska,[1] Wojciech Rosikiewicz,[1] Damian Szklarczyk,[‡,1] Wojciech Makałowski,[2] and Izabela Makałowska[*,1]

[1]Laboratory of Bionformatics, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

[2]Institute of Bioinformatics, University of Muenster, Muenster, Germany

[‡]Present address: Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

***Corresponding author:** E-mail: izabel@amu.edu.pl.

Associate editor: Helen Piontkivska

Abstract

Gene duplicates generated via retroposition were long thought to be pseudogenized and consequently decayed. However, a significant number of these genes escaped their evolutionary destiny and evolved into functional genes. Despite multiple studies, the number of functional retrogenes in human and other genomes remains unclear. We performed a comparative analysis of human, chicken, and worm genomes to identify "orphan" retrogenes, that is, retrogenes that have replaced their progenitors. We located 25 such candidates in the human genome. All of these genes were previously known, and the majority has been intensively studied. Despite this, they have never been recognized as retrogenes. Analysis revealed that the phenomenon of replacing parental genes with their retrocopies has been taking place over the entire span of animal evolution. This process was often species specific and contributed to interspecies differences. Surprisingly, these retrogenes, which should evolve in a more relaxed mode, are subject to a very strong purifying selection, which is, on average, two and a half times stronger than other human genes. Also, for retrogenes, they do not show a typical overall tendency for a testis-specific expression. Notably, seven of them are associated with human diseases. Recognizing them as "orphan" retrocopies, which have different regulatory machinery than their parents, is important for any disease studies in model organisms, especially when discoveries made in one species are transferred to humans.

Key words: retrogene, gene duplication, gene expression, human genetic disease.

Research article

Introduction

Despite advances in molecular biology and plethora of genomic and transcriptomic data, understanding genetic basis of diseases and turning basic science discoveries into therapies remains challenging. Animal experiments have contributed a lot to decoding the mechanisms of diseases. However, the value of animal studies in predicting the effectiveness of treatment is often controversial (Hackam 2007; Perel et al. 2007; van der Worp et al. 2010). Inconsistency between animal models and clinical trials may be explained by inadequate animal data or simply because animal models do not reflect disease in humans in a satisfactory way.

The key in deciphering this disparity is in understanding interspecies differences and translating genomes into phenotypes. Phenotypic diversity, beside environmental factors, is generated through changes in the genomic sequence. Without knowing which genomic features result in phenotypic differences between species, we will not be able to predict functional consequences of transferring model organism research results to medical treatment of humans. One of the fundamental factors in the evolution of lineage-specific and species-specific traits is the birth of new genes. Gene duplication is the major process contributing to the origin of these genes. There are two mechanisms for gene duplication: DNA-based creating copies with genetic features similar to their parental genes and RNA based. In RNA-based duplication, mRNA is reverse-transcribed into cDNA and reintegrated into a new location in the genome (Vanin 1984; Weiner et al. 1986; Brosius 1991). Although the mechanism of this process has not been widely studied, there is experimental evidence that in humans the machinery of long interspersed repeats is used (Esnault et al. 2000). In this type of duplication, multi-exon genes give birth to single-exon copies which, in most cases, lack regulatory elements and are commonly believed to be pseudogenes (Mighell et al. 2000). However, many of them are known to produce new, very often lineage-specific genes (Betran, Wang, et al. 2002; Marques et al. 2005; Svensson et al. 2006). They can also lead to new protein domains through fusion with other genes (Vinckenbosch et al. 2006; Baertsch et al. 2008), regulatory RNAs (Yano et al. 2004; Devor 2006), or other regulatory elements (Nozawa et al. 2005).

Soares et al. (1985) discovered for the first time a functional retrosequence in the rodent genome in 1985. They found that the rat insulin I gene is a functional retrocopy of the insulin II gene. This finding was followed by the number of discoveries of functional retrogenes in mammalian genomes (McCarrey and Thomas 1987; Ashworth et al. 1990) (for review see Brosius 1999) as well as in the fruit fly (Long and Langley 1993; Betran, Thornton, et al. 2002). Although several genome-wide surveys have been performed over the last

decade, it is still unknown how many retrogenes are actually transcribed in human and other genomes. It is estimated that the human genome contains approximately 8,000 retrogenes (Zhang et al. 2003). Harrison et al. (2005) found that some 4–6% of them are abundantly expressed. Utilizing in silico assays Vinckenbosh et al. (2006) identified over 1,000 transcribed retrogenes, out of which 120 evolved into bona fide genes. Other investigators reported that only 2–3% of processed pseudogenes are transcribed in the human genome (Yano et al. 2004; Yu et al. 2007) and an even lower number of functional retrogenes in the human genome come from the studies of Sakai et al. (2007). Only 79 of retrogenes studied by them had evidence for transcription and they estimated that 1.08% of all processed pseudogenes are transcribed. In the most recent studies, Pan and Zhang (2009) identified 163 functional human retrogenes.

Retrogenes, for a long time considered being "dead on arrival" copies of parental genes, are nowadays often called "seeds of evolution" (Brosius 1991) because they made a significant contribution to molecular evolution. As duplicates of their parental genes, these retrocopies evolve fast because duplication events allow a relaxed purifying selection, so that these genes may acquire novel functions. They are important source of functional innovations and species-specific traits. For example, retrogene *fgf4* is responsible for the dogs' chondrodisplasia. All breeds with short legs are carriers of the *fgf4* retrogene (Parker et al. 2009). Another example of retrogenes contribution in shaping interspecies differences is retrogene *RNF113B*, which gained an intron in primates and has two splicing forms with distinct expression patterns while in other mammals it has only one single-exon form (Szczesniak et al. 2011).

Retrogenes are also known to be involved in many diseases. A good example is the *RHOB* gene, a tumor suppressor of the Rho GTPases family (Prendergast 2001), which arose by retroposition in the early stage of vertebrate evolution (Sakai et al. 2007). Mutation in another retrogene, *TACSTD2* (tumor-associated calcium signal transducer 2) causes gelatinous drop-like corneal dystrophy leading to blindness (Tsujikawa et al. 1999).

Although several efforts have been made to detect functional retrogenes, their number remains unclear. A genome-wide study showed that 20% of mammalian protein encoding genes lack introns in their coding sequence (Sakharkar et al. 2002). Therefore, it is conceivable that many genes lacking introns arose by retroposition. In published studies, the identification of retrogenes was always based on the assumption that both, the parental gene and its retrocopy, are present in the genome. Therefore, only genomic sequence loci that were homologous to multi-exon genes were considered and single-exon genes without close paralogs were automatically eliminated from the set of putative retrogenes. However, we cannot exclude the possibility that the parental gene was lost or pseudogenized after the duplication and the retrogene, which took over its function, does not have any multi-exon homologs. Here, we present a comparative analysis of human, chicken, and worm genes leading to the identification of 25 "orphan" retrogenes, which likely replaced their progenitors, in the human genome. All of them are functional and although most were studied more intensively, none of them were ever recognized as a retrogene.

Materials and Methods

Identification of "Orphan" Retrogenes

The sequence collection used in this study consisted of 5,342 human transcripts encoded by single exon genes, and 60,922 human and 4,613 chicken mRNAs encoded by multi-exon genes as annotated in the UCSC Genome Browser database (Fujita et al. 2011), assemblies hg18 and galGal3, respectively. We deliberately used all human transcripts encoded by single-exon genes to avoid the exclusion of transcribed retrogenes annotated as noncoding due to the frameshift, premature stop codons, missing 3′- or 5′-end of coding sequence, and annotation errors. In addition to human and chicken genes, sequences of 4,649 human–worm orthologs were downloaded from the InParanoid database (Ostlund et al. 2010).

"Orphan" retrogenes in the human genome, that is, retrocopies without their parental genes present in the genome, were identified using three approaches. The first two were based on the analysis of sequence similarity between human and chicken genes. Furthermore, in the second approach, the genomic location was taken into consideration. The third approach relied on the gene structure analysis of already predefined human and *Caenorhabditis elegans* orthologs.

Method I

mRNA sequences from single-exon and multi-exon human genes and chicken multi-exon genes were downloaded using the UCSC Table Browser. The set of human single-exon genes was next filtered to exclude out histone sequences, which are known to be intronless in all vertebrates, as well as all sequences equal or shorter than 200 bp to eliminate putative small RNAs. In this step, we removed 79 and 2006 sequences, respectively. The remaining 3,257 sequences were used as a query in translated similarity searches, using TBLASTX (Altschul et al. 1997), against mRNAs of multi-exon chicken genes and against mRNAs of human multi-exon genes. Following the similarity searches, results were filtered based on three criteria: 1) identity percentage, 2) score in the BLAST searches, and 3) query coverage in the alignment with chicken mRNAs. Approved for further analysis were single-exon human genes that showed a higher alignment score and a higher similarity to chicken multi-exon genes than to human multi-exon gene and with an alignment covering at least 35% of the chicken mRNA sequence. After filtering, the resulting set of sequences was manually checked and all cases with an uncertain status were removed.

The manual checking included BLASTX searches against human and other genomes, synteny analysis of a retrogene and the parental gene orthologs, analysis of annotations in several resources such ENSEMBL, UCSC Genome Browser, NCBI genomic maps, as well as alignment analysis to confirm that alignment of retrogene and its parental gene ortholog covers more than two exons. The main reasons for rejecting candidates were incorrect annotations in the chicken

genome, gaps in the sequence creating artificial introns, and the alignment spanning only one exon of the parental gene ortholog. In few cases, the candidate was discarded due to the presence of parental gene paralogs and uncertainty, which of the gene was a progenitor of a given retrogene.

Method II

In the second approach, filtered transcripts from human intronless genes were used for a BLAST search against chicken multi-exon genes. Sequences with no hits to the chicken mRNAs and those with alignments to chicken transcripts shorter than 100 bp were removed from the set. The remaining pairs, a human single-exon gene and its matching chicken multi-exon gene, were analyzed in regard to their chromosomal localization and surrounding genomic sequence. We compared, by BLAST searches, genes in the nearest vicinity of candidate retrogene in the human genome and in the region near the multi-exon gene in the chicken genome. Based on the assumption that a retroposed gene will have different neighbors than its parental gene, all pairs that have as neighbors orthologous genes at one or both sides were eliminated from the data set. All gene pairs that passed this filtering were manually examined and, similarly to method I, all cases with an uncertain status were removed.

Method III

In the last approach, identifiers of human and *C. elegans* proteins coded by orthologous genes were downloaded from the InParanoid database (version 7.0) (Ostlund et al. 2010). All proteins identifiers were converted into nucleotide accession numbers using Galaxy (Goecks et al. 2010) and for each gene the exon number was obtained using the UCSC Table Browser (Karolchik et al. 2004). All pairs where a human gene had only one exon and the matching *C. elegans* gene had two or more exons were selected and manually inspected.

In the search for "orphan" retrogenes, we intentionally did not use a standard practice applied in the retrogenes identification studies, which is mapping all multi-exon genes to the genomic sequence. This approach, although very efficient in identifying retrocopies, would return a lot of pseudoretrogenes, which were beyond our interests.

Identification of Orthologous Genes in Other Species of Animals

To determine the evolutionary history of identified human "orphan" retrogenes, we looked for their orthologs and/or orthologs of their parental genes in seven vertebrate species: *Mus musculus* (house mouse), *Bos Taurus* (cattle), *Monodelphis domestica* (opossum), *Ornithorhyncus anatinus* (platypus), *Gallus gallus* (chicken), *Xenopus tropicalis* (western clawed frog), and *Danio rerio* (zebrafish) as well as in one insect species: *Drosophila melanogaster* (fruit fly). Orthology relations between genes were established based on the annotations in the NCBI Gene database (Maglott et al. 2011) and the Ensembl database (Flicek et al. 2011) as well as BLAST (Sayers et al. 2011) similarity searches.

Gene Expression Analysis

Expression of identified "orphan" retrogenes was analyzed in MTC Multiple Tissue cDNA Panels, Human I and Human II, from Clontech. The selected panels represented together cDNA libraries from 16 human tissues and organs: heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, small intestine w/o mucosal lining, colon, and peripheral leucocytes. As a positive and a negative control, *GAPHD* and *GYS2*, respectively, were used as recommended by the cDNA libraries provider.

Forward and reverse primers for all genes were designed using Primer-BLAST (Sayers et al. 2011) with the following parameters: product length 120–160 bp; primers melting temperature (T_m) 58–62°C; GC content between 40% and 60%.

The expression of analyzed genes was determined by a real-time polymerase chain reaction (PCR) method (Kubista et al. 2006) performed in Applied Biosystems 7900HT System with Power SYBR Green PCR Master Mix (Applied Biosystems) and the results were interpreted using SDS Software 2.3. The cut-off value for C_T (cycle threshold) was established as 32 based on the optimal cut-off for real-time PCR experiments obtained in other studies. Results were visualized through the construction of a heatmap in the R software environment (version 2.11.1).

Identification of MicroRNA Target Sites and TFBS Analysis

Information about microRNA target sites was obtained from TargetScan Release 5.1, a database of target site predictions (Friedman et al. 2009). Identification of potential binding sites for transcription factors in DNA sequences was performed using Match^{TM} – 1.0 Public (Alamanova et al. 2010). We analyzed 1,000 nt upstream sequence for each gene and looked for transcription factor binding sites with the highest two most important parameters: the matrix similarity score and the core similarity score. Identification was limited to vertebrate-specific weight matrices.

Calculation of K_A/K_S Ratio

The K_A/K_S ratio for human retrogenes and their orthologs in mice was calculated using the K_AK_S_Calculator, which uses the MYN method (modified version of the Yang–Nielsen method) (Zhang et al. 2006).

Results

Identification of Retrogenes without Parents

As proposed in several papers by Nei and coworkers (Ota and Nei 1994; Nei et al. 2000; Nikolaidis et al. 2005) gene families may evolve by the "birth-and-death process." Therefore, after the speciation event, the divergence between two resultant species may be shaped by the gradual accumulation of gene gains and losses. Retroposition provides a wealth of gene duplicates. These so-called processed pseudogenes are considered to have little evolutionary significance as they are "dead on arrival" and represent disabled copies of functional

parental gene (Li et al. 1981; Lynch and Conery 2000). However, some of them gain a function and become functional paralogs (Soares et al. 1985; McCarrey and Thomas 1987; Ashworth et al. 1990; Long and Langley 1993; Brosius 1999). Thus, according to the "birth-and-death evolution," we may expect that after divergence in one lineage both copies may be retained, in another the retrocopy may be lost, and yet in another the parental gene will lose its function and the retrogene will be left as the only functional copy.

Zhang et al. (2010) described what they called unitary pseudogenes in the primate lineage. They identified 87 unprocessed pseudogenes without functioning counterparts. These genes, although well established in the vertebrate lineage, are extinct in humans and/or other primates. In this study, we also looked for well-established genes that were lost, for example, due to deletion, or pseudogenized in the human genome. However, the function of these genes was undertaken by their duplicates—retrocopies. These presumed "orphan" retrogenes were identified based on the comparative analysis of human, chicken, and worm genes using three different approaches as described in the Materials and Methods section. In the first one, putative orphan retrogenes were selected based on similarity searches, in which human single-exon genes were run against human and chicken multi-exon gene transcripts. The results of both BLAST searches were compared and sequences showing higher similarity to chicken genes than to human genes were selected. Seventeen single-exon human genes met these rigorous filtering criteria. However, after manual checking only four pairs of human retrogenes and chicken orthologs of their parental genes remained.

In the second approach, the results of a similarity search for human single-exon genes versus chicken multi-exon genes were filtered and pairs of human–chicken sequences with at least 100 bp alignments were selected for further studies. Only 915 pairs met this criterion. For further data processing, considering the mechanism of retroposition, we made a rather obvious assumption that a retrogene and its parental gene, or in this case the ortholog of parental gene, should have different genomic locations. Based on this deduction, we analyzed sequences surrounding genes from each human–chicken pair and removed those that had orthologous genes at one or both sides. This analysis returned 260 potential pairs of "orphan" retrogenes in the human genome and orthologs of its parental gene in the chicken genome. Nevertheless, only nine pairs were confirmed after manual examination, out of which four were identified in the previous approach.

It is noticeable that the ratio of false-positives in methods I and II was relatively high. This may imply inaccuracy in the methodology. However, majority of false positives come from incorrect annotations of the chicken genome. In addition, gaps in the chicken genomic sequence were generating artificial introns and often single-exon chicken genes would appear, according to annotations, as multi-exon.

The third strategy relied on the orthology relationships established in the InParanoid database (Ostlund et al. 2010). 4649 human–*Caenorhabditis elegans* orthologous groups were identified in the database. After filtering followed by an exon number comparison, as described in Material and Methods, 58 pairs were selected. Twenty pairs passed manual verification and four of them were already identified by methods I and II. This gave 16 new "orphan" retrogenes. Therefore, overall we identified 25 unique retrogenes, which do not have their parental gene in the human genome. All of these genes are listed in table 1. Interestingly, only for one retrogene, *CHMP1B*, we were able to find traces of the parental gene in the human genome. In other cases, the region where the parental gene was located was either deleted or mutated to the degree in which no similarity can be found.

Zhang et al. (2011) pointed out that partial DNA-level duplications of intron containing genes can make a significant contribution to the existence of intronless genes. Therefore, even relatively long alignments between single-exon genes and intron-containing parents may not be sufficient to define a new copy as retrogene. Keeping this in mind, in the process of manual evaluation, we looked not only at the alignment length but also checked whether the alignment covers exon–exon junctions of putative parental gene ortholog. The graphical representation of this comparison is shown in supplementary figure S1, Supplementary Material online. It is visible that in all identified by us retrogene–parental ortholog pairs alignments cover all or majority of introns located in the coding region.

Retroposition and Loss of Parental Gene

Each pair of genes, either human–chicken or human–*C. elegans*, was further examined in selected animal species: house mouse, cattle, opossum, platypus, zebrafish, frog, and fruit fly. In addition, genes identified in method III were investigated in the chicken genome. Using genome annotations and similarity searches, we looked for orthologs of retrogenes as well as orthologs of multi-exonic parental genes. The main goal of this analysis was to estimate the time when the retroposition took place and when the parental gene was lost or pseudogenized. We were able to identify the time of these events for all genes. Interestingly, the loss of the parental gene occurred, in most cases, almost simultaneously with retroposition, before the next major phylogenetic split (fig. 1). The exceptions are genes *CHMP1B* and *TRMT12* in the mammalian lineage. The first of these, retrogene *CHMP1B*, arose in a common ancestor of placental mammals but the parental gene is still functioning in some mammals, for example, in rodents. In other species, such as humans and cattle, the parental gene was pseudogenized. This loss of function in the human and cow genomes occurred independently. *TRMT12* was also retroposed in the genome of the placental mammals' ancestor but the parental gene was lost after the divergence of *Metatheria* and *Eutheria* (fig. 1).

We cannot exclude that in some cases, the parental gene is not observed in the genomic sequence due to the sequencing gaps. However, this is not very likely in the case of the human genome and genomes of model organisms such as mouse, fruit fly, and *C. elegans*, which were sequenced with high coverage and are well annotated. For other genomes used

Table 1. "Orphan" Retrogenes in the Human Genome.

	Gene Symbol	Gene Name	Chromosomal Localization	K_a	K_s	K_a/K_s
1	*MAB21L1*	Mab-21-like 1	13	0	0.74	0
2	*MAB21L2*	Mab-21-like 2	4	0.001	0.806	0.001
3	*PURA*	Purine-rich element binding protein A	5	0.001	0.29	0.004
4	*ADRA2A*[a]	Adrenergic, alpha-2A-, receptor	10	0.036	2,112	0.017
5	*CHMP1B*[a]	Chromatin modifying protein 1B	18	0.009	0.398	0.022
6	*IMP3*[a]	U3 small nucleolar ribonucleoprotein	15	0.017	0.681	0.024
7	*EXOC8*	Exocyst complex component 8	1	0.03	1.214	0.024
8	*B3GALT6*	UDP-Gal:betaGal beta 1,3-galactosyltransferase polypeptide 6	1	0.073	1.79	0.041
9	*RRS1*[a]	RRS1 ribosome biogenesis regulator	8	0.042	0.963	0.043
10	*TTC30B*	Tetratricopeptide repeat domain 30B	2	0.037	0.594	0.063
11	*PIGM*[a]	Phosphatidylinositol glycan anchor biosynthesis, class M	1	0.051	0.698	0.073
12	*MOCS3*	Molybdenum cofactor synthesis 3	20	0.117	1.391	0.084
13	*TBCC*	Tubulin folding cofactor C	6	0.126	1.489	0.085
14	*CH25H*	Cholesterol 25-hydroxylase	10	0.11	1.151	0.095
15	*CEBPB*	CCAAT/enhancer binding protein (C/EBP), beta	20	0.068	0.687	0.099
16	*ADRA2B*	Adrenergic, alpha-2B-, receptor	2	0.079	0.769	0.103
17	*MARS2*	Methionyl-tRNA synthetase 2	2	0.073	0.697	0.105
18	*UTP3*	Small subunit (SSU) processome component	4	0.063	0.589	0.108
19	*KTI12*	KTI12 homolog, chromatin associated	1	0.129	1.165	0.111
20	*MGAT2*[a]	Mannosyl (alpha-1,6-)-glycoprotein beta-1,2-*N*-acetylglucosaminyltransferase	14	0.058	0.407	0.144
21	*RNF113A*	Ring finger protein 113A	X	0.066	0.423	0.156
22	*SFT2D3*	SFT2 domain containing 3	2	0.129	0.822	0.157
23	*ZNF830*	Zinc finger protein 830	17	0.09	0.459	0.197
24	*TRMT12*[a]	tRNA methyltransferase 12 homolog	8	0.107	0.515	0.208
25	*LCMT2*	Leucine carboxyl methyltransferase 2	15	0.131	0.54	0.242

[a]Gene associated with human disease.

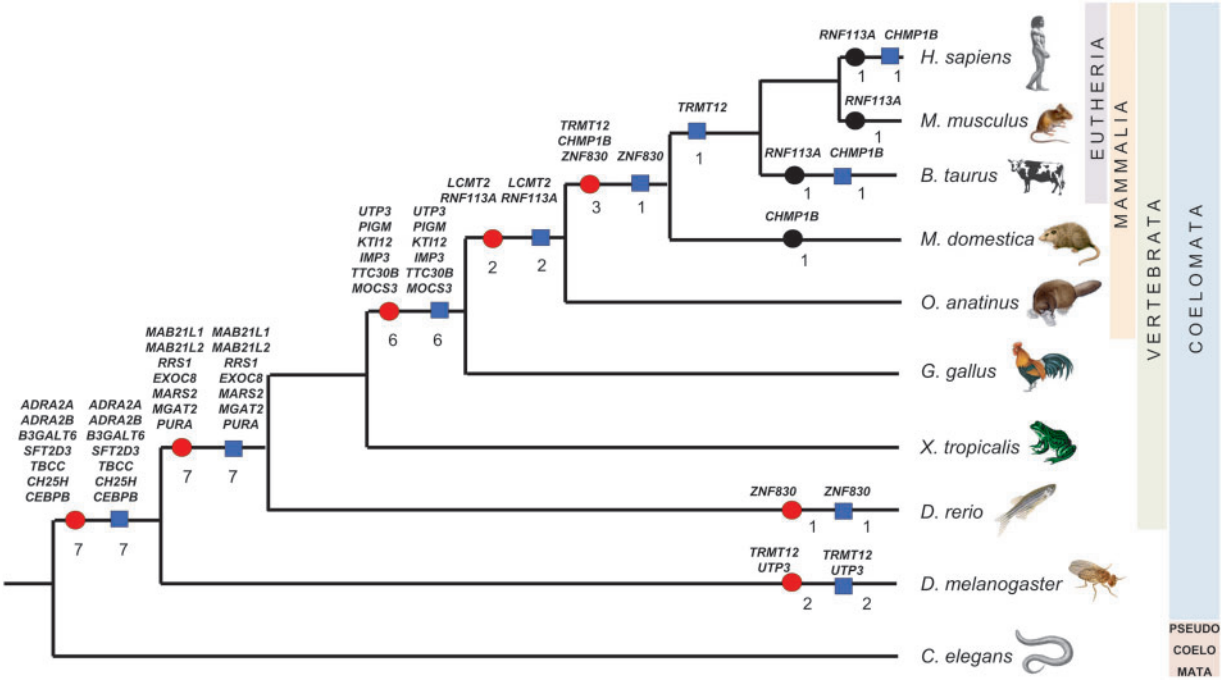


Fig. 1. Phylogenetic tree showing points of retroposition and parental gene loss for each retrocopy. Red circle represents retroposition; blue square, parental gene loss; black circle, retrogene duplication or retroposition.

in the analysis, we cannot completely rule out the possibility that the parental gene exists but was not sequenced.

It is known that retroposition has a remarkably high rate in placental mammals (Moran et al. 1996; Ostlund et al. 2010), and therefore we expected that the turnover between the parental gene and its retrocopy will be especially intensive in this taxonomic group. Surprisingly, the highest rate of parental gene loss subsequent to the retroposition was before the divergence of vertebrates. Seven genes were retroposed and eventually lost right after the divergence of *Pseudocelomata* and *Celomata* and also seven retrogenes replaced their parental genes in the common ancestor of vertebrates (fig. 1). The next wave of the birth of "orphan" retrogenes started in the genome of the warm-blooded animals' predecessor. Six retrogenes substituted parental genes at this point of the evolution and two parental genes were lost in the genome of the mammalian ancestor. Only three retrogenes took the place of their progenitors in placental mammals, out of which two in *Eutheria*.

Our analyses also revealed that four parental genes, which are lost in the human genome, independently vanished in other species (fig. 1). It was already mentioned in this article that the progenitor of the *CHMP1B* retrogene was pseudogenized in the human as well as in the cattle genome. In addition *ZNF830* was replaced by its retrocopy in *Danio rerio*. Two retrogenes, *TRMT12* and *UTP3*, took the place of their parents in the *D. melanogaster* genome.

Disease Association

As we have already mentioned, retrogenes can be involved in human diseases (Tsujikawa et al. 1999; Prendergast 2001; Zemojtel et al. 2010). Identified by us "orphan" retrogenes are not the exception in this matter. However, in all previously described cases both genes, a retrocopy and its parent, were present. Here, we identified disease-associated retrogenes, which functionally replaced their parental genes. These genes, although coding for the same protein as the pseudogenized parent, have different regulatory machinery, as promoter regions are not inherited in the process of retrotransposition. There is an evidence for functional evolution of retrogenes and differences in the expression scheme between the parental gene and its functional retrocopy (Zhang et al. 2002; Marques et al. 2005; Vinckenbosch et al. 2006; Zemojtel et al. 2010). Therefore, we may anticipate that "orphan" retrogenes are not necessarily regulated in the same way as their parents were. This should be kept in mind in any disease studies in model organisms, where discoveries made in one species are transferred to humans, especially when one organism has functional parental gene and the other only its retrocopy.

Among 25 "orphan" retrogenes identified by us, seven are involved in human diseases, which corresponds to 28% of all identified genes. Two of these genes are linked to cancer. The *IMP3* gene is expressed in tumors and its expression level is associated with metastasis in renal cell carcinomas and patient's survival rate (Jiang, Chu, et al. 2008; Jiang, Lohse, et al. 2008). Overexpression of another "orphan" retrogene, *TRMT12*, may lead to translation errors in breast tumor cells (Rodriguez et al. 2007). A high expression level of *ADRA2A* can increase type 2 diabetes risk (Rosengren et al. 2010). The same gene is also involved in attention-deficit/hyperactivity disorder (Roman et al. 2006). Other examples include *MGAT2* responsible for defective brain development (Tan et al. 1996), mutation of *ADRB1* is associated with congestive heart failure and beta-blocker response (Mason et al. 1999), *RRS1* is involved in endoplasmic reticulum stress response in Huntington's disease (Carnemolla et al. 2009), and *PIGM* is linked to glycosylphosphatidylinositol deficiency (Almeida et al. 2006).

It is expected that molecular evolution of retrogenes is selectively neutral and therefore these genes evolve relatively quickly, although there is evidence for retrogenes under strong purifying selection (Vinckenbosch et al. 2006; Yu et al. 2007). The degree and type of selection can be measured by the ratio of nonsynonymous substitutions (K_A) to synonymous substitutions (K_S). Under neutral evolution $K_A = K_S$, deviation of K_A from K_S may be due to positive selection when the K_A/K_S is >1, or purifying selection when $K_A/K_S < 1$. Nevertheless, genes are considered to be under strong purifying selection when K_A/K_S ratio is $\ll 1$ (Hurst 2002). We calculated the K_A/K_S ratio for all "orphan" human retrogenes and their orthologs in mouse (table 1). As the results show, none of these genes are evolving neutrally and the K_A/K_S ratio is <0.25 for all of them, strongly indicating that retrogenes, which replaced their parents, are under purifying selection. The average ratio for all 25 genes is 0.088 and it is much lower than the average for human–mouse genes, which was estimated as 0.180 (Makalowski and Boguski 1998). An even stronger purifying selection is observed in the case of seven disease-associated "orphan" retrogenes. The average ratio for this group is 0.076. Interestingly, this value is lower than previously published. Tu et al. (2006) analyzed the evolutionary rate for human disease genes and obtained, for human–mouse orthologs, average K_A/K_S ratio 0.12. Another group (Thomas et al. 2003) analyzed 121 human genes implicated in cancer and calculated the average ratio to be 0.079, which is close to the value obtained by us. It is intriguing that the retrogenes studied by us, disease related or not, are under a similarly strong pressure as cancer-related genes.

Although we did not apply any minimum similarity filtering, it is possible that methods used by us led to the enrichment of slow evolving genes in our set. On the other hand, these genes represent single-copy or two-copy genes, which are known to be slowly evolving (Waterhouse et al. 2011).

A Study Case of *CHMP1B* Gene

An interesting case represents *CHMP1B*, a retrogene associated with hereditary spastic paraplegia (Reid et al. 2005). This gene was retroposed before the divergence of *Theria*. The retrogene was then either tandemly duplicated or retroposed in *Metatheria* as opossum has two single-exon genes and one multi-exon gene. In the *Eutherian* lineage, the retrogene and its parent coexist in the majority of the taxa. However, in the human and cattle genomes the parental genes do not

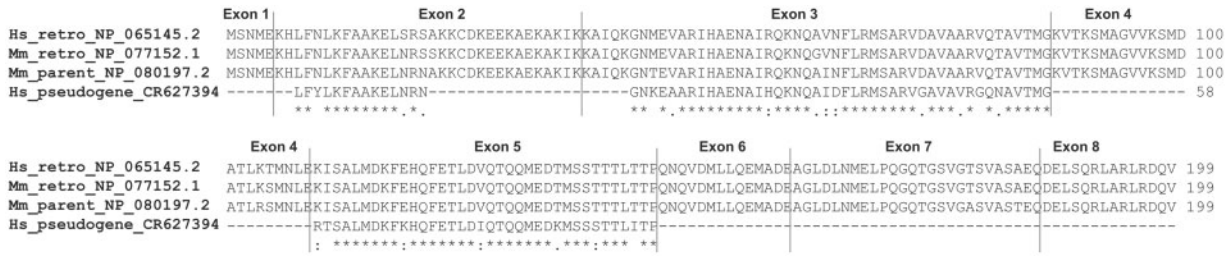


Fig. 2. Alignment of proteins coded by human and mouse *CHMP1B* retrogenes and their parental genes (functional gene in mouse and pseudogene in human genome).

function anymore. Pseudogenization of the *CHMP1B* parental gene was independent in both lineages since in mice and rats, which like humans belong to *Euarchontoglires*, the parental gene is intact and expressed in various tissues. In the primate lineage, the *CHMP1B* parent was pseudogenized in the genome of the ancestor of Old World and New World monkeys because this gene is fragmentary in all available primate genomes: marmoset, macaque, orangutan, chimpanzee, and human.

Proteins coded by the *CHMP1B* retrogene and its functional parents are highly conserved (fig. 2), which may indicate that retrogene gained its function shortly after the retroposition and immediately became subjected to purifying selection. The strong pressure to conserve protein sequences confirms the K_A/K_S ratio, which is 0.012 for the mouse retrogene and its parent and 0.022 for human and mouse retrogenes. This is an order of magnitude lower than average K_A/K_S ratio (0.18) for human–mouse coding sequences (Makalowski and Boguski 1998). The human parental gene, although pseudogenized, does get expressed; there is one mRNA sequence, CR627394, and two EST sequences deposited in the GenBank. Nevertheless, from the very low number of ESTs, we may conclude that the expression level of this gene is very low. Also, this gene is significantly different from its ortholog in mice. It contains only parts of exons coding for the prototype protein: fragment of exon 2 and most of exons 3 and 5 (fig. 2). In addition, there is a frameshift since a fragment of exon 2 is in frame +1 and the other two exons are in frame +3. Interestingly, nearly all the coding exons present in the mouse gene can be detected in the human genomic sequence but they are not used in any transcript.

Retroposed genes need to recruit regulatory elements to become transcribed and usually, as a consequence of hiring transcription regulation factors different from their parent, acquire a new function. We performed analysis of 1,000 bp upstream sequences of human and mouse *CHMP1B* retrogenes and the mouse parental gene. Indeed, regulatory elements present in upstream sequences of retrogenes differ from elements observed in parental gene's regulatory region. Three transcription factor binding sites (TFBS): CREB, CRE-BP1, and E2F are specific for human and mouse retrogenes and are not found in the regulatory region of the mouse parental gene. On the other hand, the mouse parental gene has two unique TFBS: HNF-1 and Evi-1. There is no single TFBS shared between all three genes (fig. 3). However, the transcript level is not regulated exclusively by the transcription factors. Short RNA molecules like microRNA may bind to the complementary sequence on target transcripts leading to translational repression and gene silencing (Ambros 2004). MicroRNA target sites are located in 3′-UTR sequences and therefore, unlike transcription factor binding sites, are inherited by retrogenes. It is known that the conservation of 3′-UTRs is much lower than conservation of coding sequence (Makalowski and Boguski 1998). Nevertheless, most microRNA targets are well conserved in mammalian mRNAs (Friedman et al. 2009). Employing TargetScan (Friedman et al. 2009), we identified microRNA target sites in *CHMP1B* retrogenes and their parental genes, functional or pseudogenized, in several mammalian species. The TargetScan identified only one microRNA target site, site for miR-743ab/743b-3p, conserved in all functional parental genes. The target sequence for this microRNA, present in rodent, horse, and elephant genes, was clearly deleted in human and chimpanzee where the gene was pseudogenised (fig. 4*A*). None of the other target sites recognized by the program were conserved in all functional genes. For example, sites for miR-155 and miR-669f are conserved in rodent and elephant functional genes but not in horse genes. On the other hand, the target site for miR-9 is conserved in mouse, rat, and horse but not in elephant. All these four target sites are conserved in the human pseudogene and three of them in the chimpanzee pseudogene.

CHMP1B retrogenes have two highly conserved microRNA target sites, miR-9 and miR-182, which are present in all available transcripts from placental mammals (fig. 4*B*). Interestingly, only one of them, target site for miR-9, is also present in some but not all functional parental genes. In addition, this site has a different location in parental genes and in retrogenes and the microRNA–mRNA pairing type is also different. Although in retrogenes the site for miR-9 is 7mer-1A type, in parental genes it is type 7mer-m8 (Friedman et al. 2009).

It is quite interesting that retrogenes, which are expected to evolve under a more relaxed selective pressure, have conserved microRNA target sites to a greater extent than that of parental genes. However, considering the pseudogenization of parental gene in some genomes, the lack of high conservation of microRNA target sites in the remaining functional genes may indicate that retrogenes took over the function in all genomes and the parental gene is an "unnecessary copy,"

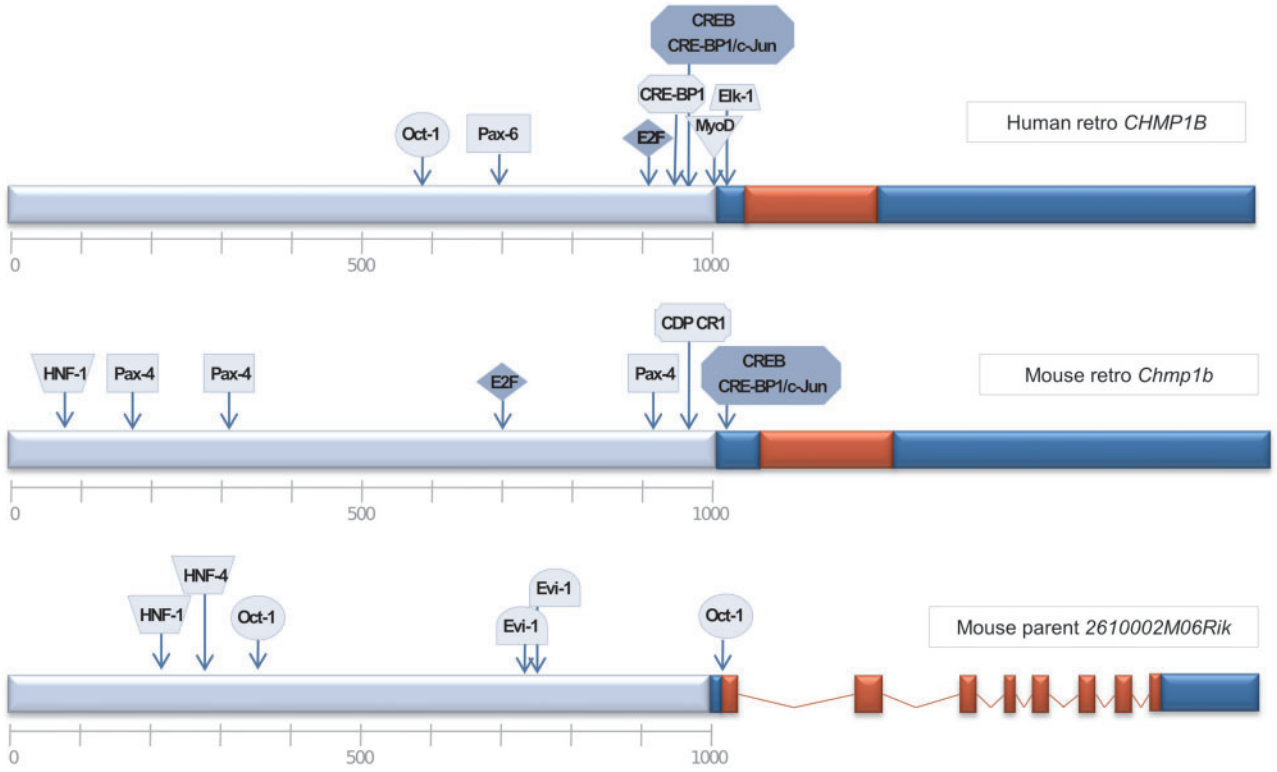


Fig. 3. Upstream regions of human and mouse *CHMP1B* retrogenes and mouse parental gene with annotated positions of identified transcription factor binding sites. TFBS which are shared by retrogenes but not present in upstream sequence of parental gene have darker background.

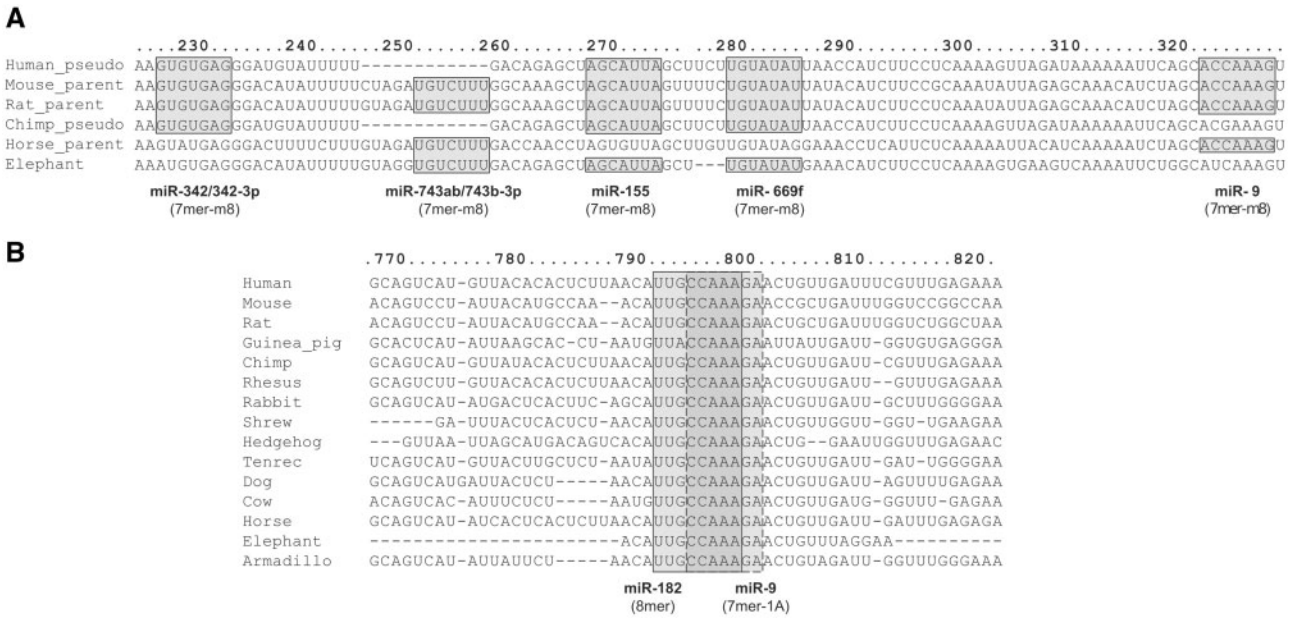


Fig. 4. microRNA target sites in 3′-UTR sequences of *CHMP1B* mammalian retrogenes (*A*) and available functional or pseudogenized parental genes (*B*).

which eventually may also lose its function in other mammalian genomes.

Expression Pattern

Gene retroposition, together with segmental duplication, belongs to the central mechanisms responsible for the creation of species-specific traits (Brosius 1991, 1999; Marques et al. 2005). Duplication of chromosomal segments tends to produce daughter copies that inherit features of their parental genes. Therefore, these copies show not only the same protein functions but also similar expression patterns. On the contrary, the retroposed cDNA is generally expected to lack regulatory elements and duplicated genes are considered to be

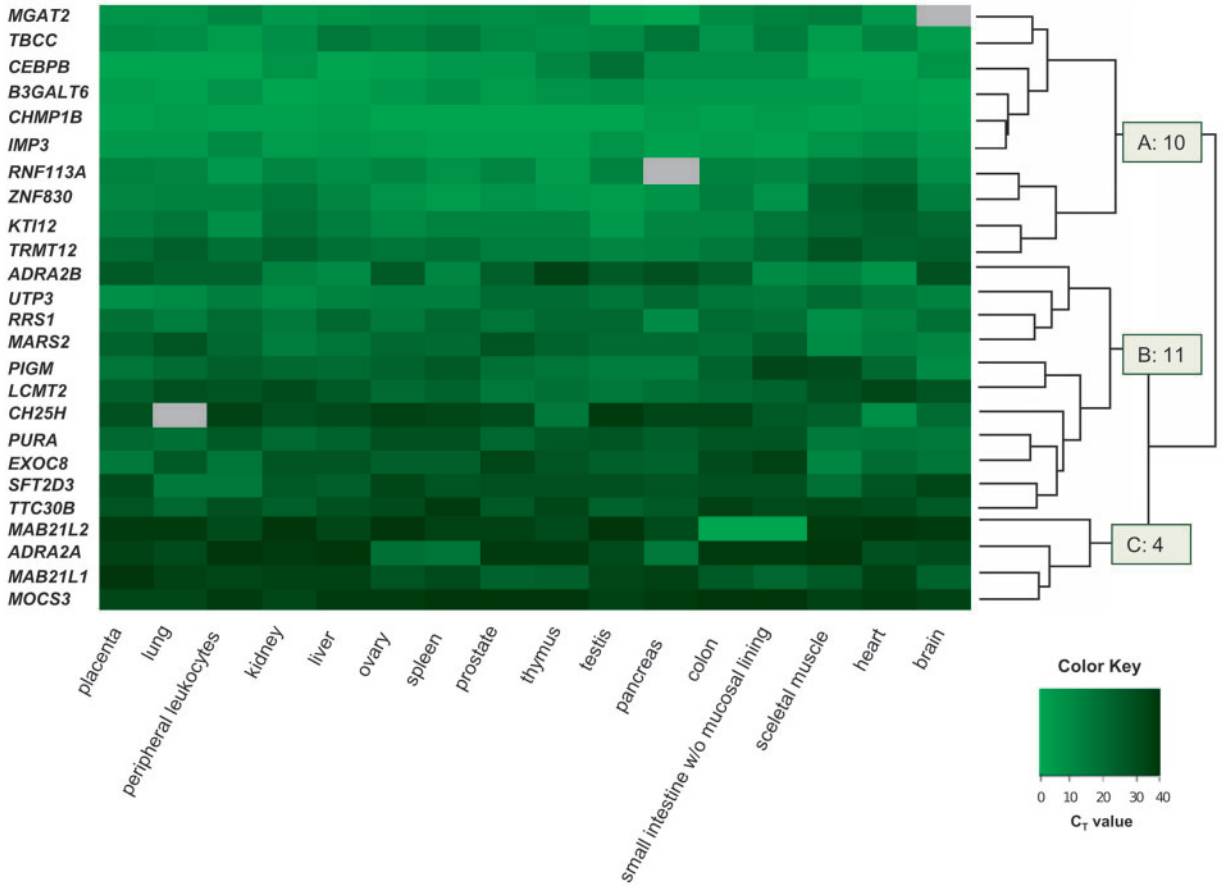


Fig. 5. Heat map representing expression pattern of all identified human "orphan" retrogenes. Gray color indicates undetermined CT values.

"dead on arrival." However, as a number of studies shows, many of them do acquire new functions (Burki and Kaessmann 2004; Krasnov et al. 2005; Sakai et al. 2007; Kaessmann et al. 2009). These new functions, usually different from the functions of parental genes, may come from the gain of new spatiotemporal expression patterns, imposed by the content of the genomic sequence surrounding inserted cDNA. Numerous studies revealed a tendency of retrogenes to be expressed in the testis (Marques et al. 2005; Vinckenbosch et al. 2006; Potrzebowski et al. 2008) and a significant excess of autosomal testis-expressed retrogenes were identified as duplicates of X-linked parental genes (Betran, Thornton, et al. 2002). This specific transcription of retrocopies may be resulting from the hypertranscription state observed in meiotic and postmeiotic spermatogenic cells (Kleene 2001). An alternative explanation may come from the hypothesis that retrocopies are preferentially inserted into actively transcribed, and therefore open chromatin (Fontanillas et al. 2007). As the retroposition occurs in the germ line, retrocopies may primarily be inserted into, or nearby genes expressed in the germ line. This could enable and/or enhance their expression in testis. Yet another hypothesis, based on the fact that there is an excess of retrogenes originated from the X chromosome, links this testis-specific expression with an escape from the male meiotic sex chromosome inactivation (Emerson et al. 2004; Wang 2004).

Preferential expression of retrogenes in testis was previously reported for retrocopies for which functional parent genes prevail in a given genome (Brosius 1991, 1999; Marques et al. 2005). To test if this specific pattern is also observable in "orphan" retrogenes we performed a real-time PCR for all 25 retrogenes in 16 human cDNA libraries including a cDNA library from testis. Real-time PCR C_T values referring to the number of cycles during reaction in which product (dsDNA) appeared, with cut-off C_T 32, were used to construct a heat map of expression profiles with a dendrogram (fig. 5). A majority of investigated retrogenes, 19 out of 25, was detected in all libraries. Five genes were expressed in 15 libraries and 1 in 14. No single retrogene revealed a testis-specific expression, including those that originated from genes located on chromosome X, like *CHMP1B* or *TRMT12*; both of them are ubiquitously expressed.

Dai et al. (2006) found that new genes seem to be expressed in fewer tissues or organs in comparison with parental genes. From the presented data, obviously we cannot make any conclusions as for the change in the expression pattern in comparison with these genes progenitors because parental genes are not present in the human genome and comparison with other species would be questionable. However, we made one interesting observation. The expression pattern of studied retrogenes is related to their age. Younger retrocopies tend to be expressed in all tissues and have a higher expression level. Cluster A represents retrogenes with the strongest

and broadest expression. Out of 10 genes in this cluster, six were retroposed in the ancestor of warmblooded animals or later. Clusters B (moderate expression) and C (lowest expression) are build in majority from genes retroposed before vertebrates. This is quite intriguing since, according to a previous study (Wolf et al. 2009), we should rather expect that retrogenes slowly gain functions as they get older and their regulatory regions "mature." Apparently, it seems to be the opposite in the case of "orphan" retrogenes where younger copies have, on average, a broader and higher expression.

Discussion

Gene duplicates generated via retroposition were long thought to be pseudogenized and consequently decayed. However, a significant number of these genes escaped their evolutionary destiny and evolved into functional genes. The function of the retrogenes was usually discussed in the aspects of neofunctionalization and/or subfunctionalization (Kaessmann et al. 2009). Here, we presented the first genome wide analysis aimed at the identification of retrogenes which replaced their progenitors and took over their functions. We identified 25 functional retrogenes, for which parental genes do not exist or do not function anymore in the human genome. None of these genes were considered earlier as retrogenes. One of the most surprising discoveries was the fact that many of these genes have ancient origins dating back even more than 900 million years and are common for all *Coelomata*. Obviously, we cannot exclude that these intronless copies originated via other than retroposition mechanism of intron loss; however, retroposition is the most parsimonious and most plausible in the case where all introns from a given gene have disappeared. Unexpectedly, despite a very intensive retroposition in placental mammals (Moran et al. 1996), a relatively low number of retrogenes replaced their parent in the mammalian lineage. One explanation could be that they just need a long time to do so but the data does not verify this. The replacement of the parental gene, in the majority of cases, was in the same lineage, before the next major divergence.

It is postulated that molecular evolution of retrocopies is selectively neutral, whereas their parental genes are subject to purifying selection. Indeed, Yu et al. (2007) found that the majority of retrogenes are in the state of a "relaxed" selection. Nonetheless, they also discovered that some human retrogenes are undergoing a nonneutral evolution. Retrogenes under a strong purifying selection were also identified by Vinckenbosch et al. (2006). Apparently, all the identified here "orphan" retrogenes are under a strong purifying selection. We showed that the *CHMP1B* protein is highly conserved between mouse parental genes and retrogenes as well as between human and mouse retrogenes. This strong conservation and low K_A/K_S values are characteristic for all analyzed by us genes. As shown in table 1, the ratio of nonsynonymous to synonymous substitution for all but three genes is below the average value estimated for human–mouse genes, which is 0.18 (Makalowski and Boguski 1998) and the average for all "orphan" retrogenes is about two times lower: 0.088. Therefore, this particular group of retrogenes is not only, without any exception, under a strong purifying selection but also evolves at a lower than average rate. This rate is even lower for disease associated "orphan" retrogenes: 0.076. The high conservation level is in concordance with the observation that these genes replaced their parents soon after the retroposition. Consequently, they became the only functional copy of the gene and their evolution was immediately constrained by a purifying selection.

Large-scale analyses of retrogenes in mammals and fruit flies revealed the overall tendency to testis-specific expression (Marques et al. 2005; Vinckenbosch et al. 2006; Potrzebowski et al. 2008). This trend was observed independently of the parental gene expression pattern. Shiao et al. (2007) showed that mouse retrogenes are expressed at more restrictive pattern than parental paralogs and all of them were expressed predominantly in testis. Similar observation was made by Dai et al. (2006) based on the *Drosophila* retrogenes study. Our study does not confirm this bias. The majority of "orphan" retrogenes was expressed in all examined 16 tissues/organs. Not a single gene showed a testis-specific expression pattern. The simple explanation of this disparity may be in the fact that analyzed by us retrogenes naturally mimic the parental expression pattern and therefore, have much broader expression than expected. It was also suggested that the propensity to be expressed in testis observed in other studies might be related to the fact that in meiotic and postmeiotic spermatogenic cells chromosomes are in the state of hypertranscription. This state enables transcription of DNA that is usually not transcribed and therefore facilitates the transcription of retrocopies (Kleene 2001). Subsequently, these retrocopies could evolve into bona fide genes, enhance their regulatory elements, and broaden the range of tissues they get expressed in. If this would be a scenario for "orphan" retrogenes evolution we would see a limited expression in younger retrogenes and a wider expression in older copies. Evidently the picture is quite the opposite, younger genes from our set tend to be ubiquitously expressed at relatively high level and the older ones have more limited expression. These results are in disagreement with the studies of Wolf et al. (2009) who found that among human genes those that are eukaryote specific, "old" ones, are expressed at a higher level than younger, mammalian-specific genes.

It has been shown that many retrogenes, also those that are functional, are species-specific and contribute to interspecies differences. Some of these differences are of a high importance in medical research and may be responsible for the fact that results from animal studies cannot be transferred to humans. For example, the functional mouse retrogene *Rps23r1* reduces Alzheimer's beta-amyloid levels and tau phosphorylation (Zhang et al. 2009). However, results of this study cannot be applied to humans because this particular retrogene is rodent specific and does not exist in the human genome. Recognizing which retrogene is species specific, which replaced its parental gene, and which coexists with its progenitor is of high importance. In each of these scenarios genes would behave differently. If parental genes and retrogenes function as a single copy (i.e., parental only or retrogene only), they would code for the same protein but

their expression regulation would be different. Therefore, it would be crucial to check if genes that seem to be very similar from the protein comparison level are truly orthologous before transferring animal studies to humans. If both copies exist, we may expect that there will be either subfunctionalization and functions previously carried out by parental genes will be divided between these two copies or alternatively a retrocopy could develop completely new functions. In the described example of the *CHMP1B* gene, the human retrogene was associated with hereditary spastic paraplegia (Reid et al. 2005). Mice are the most likely species of choice when one would like to study this gene in a model organism. However, mice have both a functional retrogene and its parent, coding for almost identical protein. In the human genome, the parental gene got pseudogenized and does not code for a functional protein anymore. Although the parental gene could compensate mutation in the *CHMP1B* retrogene in mice, in humans it could not. Therefore, studies on the *CHMP1B* gene in mice may not be, by any means, comparable with what is taking place in humans.

Here, we presumed that analyzed retrogenes functionally replaced pseudogenized parental genes. To consider these evolutionary events as perfect "replacement," the retrogene would need to have the same regulatory sequences as parental gene and exhibit identical expression pattern. Because retrogenes, in most cases, do not inherit regulatory regions (the exception is the case when parental gene has alternative regulatory motifs in the 3′-UTR region), they need to acquire new regulatory machinery. This could happen either by mutations and positive selection leading to the origination of appropriate regulatory elements or by the "hitchhiking" of the existing elements regulating nearby gene. Without assurance that newly developed or adopted elements are the same as possessed by parental gene we cannot, in unquestionable way, determine whether the events described by us illustrate "replacement" or neofunctionalization. Because for the majority of retrogenes, there is no detectable trace of their parents in the human genome we cannot perform any considerable comparative studies. However, it would be interesting to see how evolutionary processes change the genomic sequence into the regulatory elements and to what degree these sequences mimic sequences of parental genes. To comprehend these processes a large-scale comparative analysis of functional retrogenes and their progenitors are required and such studies were recently launched in our laboratory.

Before the final conclusions, it is necessary to point out that the number of 25 "orphan" retrogenes in the human genome may seem to be low and not very appealing. At this point, it is impossible to form the opinion whether the number of such genes simply is so low or maybe the methodology needs to be worked out for better results as there are no studies to compare with. However, identifying retrogenes that lost their progenitors is very challenging due to the fact that many genes underwent multiple, and sometimes partial, duplications followed by significant changes in the gene structure, which often are difficult to trace. In addition, poorly annotated genomes likely produce false positives. Moreover, many retrogenes are known to gain exons and introns and in this particular study, we focused only on single exon genes. Nevertheless, we are currently conducting analyses concentrated on functional retrocopies, which acquired new exons and/or gain introns. It is quite conceivable that this study will reveal additional examples of human "orphan" retrogenes.

In summary, we may say that "orphan" retrogenes represent a very specific group of genes. They not only replaced their parental gene but also "behave" in unexpected ways. Although previous studies suggested that retrogenes evolve neutrally or under a relaxed functional constraint, they are actually more conserved than the average gene. They also seem to have a reversed expression pattern, that is, younger genes have higher expression and older ones are more limited. In addition, many of them are involved in serious human diseases. Altogether, these facts make this class of genes extremely interesting.

Supplementary Material

Supplementary figure S1 is available at *Molecular Biology and Evolution* online (http:www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Jakub Dolata for technical support during real-time PCR experiments. This work was supported by Ministry of Science and Higher Education grant no. N303 320 437 (to I.M.), National Science Centre grant no. 2011/01/N/NZ2/01701 (to J.C.), and Seventh Frame Work Programme of the European Union, International Research Staff Exchange Scheme grant no. PIRSES-GA-2009-247633 (to I.M, W.M., and J.C).

References

Alamanova D, Stegmaier P, Kel A. 2010. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics* 11:225.

Almeida AM, Murakami Y, Layton DM, et al. (17 co-authors). 2006. Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. *Nat Med.* 12:846–851.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Ambros V. 2004. The functions of animal microRNAs. *Nature* 431: 350–355.

Ashworth A, Skene B, Swift S, Lovell-Badge R. 1990. Zfa is an expressed retroposon derived from an alternative transcript of the Zfx gene. *EMBO J.* 9:1529–1534.

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* 9:466.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12:1854–1859.

Betran E, Wang W, Jin L, Long M. 2002. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol Biol Evol.* 19:654–663.

Brosius J. 1991. Retroposons—seeds of evolution. *Science* 251:753.

Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 238: 115–134.

Burki F, Kaessmann H. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet.* 36: 1061–1063.

Carnemolla A, Fossale E, Agostoni E, Michelazzi S, Calligaris R, De Maso L, Del Sal G, MacDonald ME, Persichetti F. 2009. Rrs1 is involved in endoplasmic reticulum stress response in Huntington disease. *J Biol Chem.* 284:18167–18173.

Dai H, Yoshimatsu TF, Long M. 2006. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. *Gene* 385:96–102.

Devor EJ. 2006. Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes. *J Hered.* 97:186–190.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* 303:537–540.

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet.* 24:363–367.

Flicek P, Amode MR, Barrell D, et al. (52 co-authors). 2011. Ensembl 2011. *Nucleic Acids Res.* 39:D800–D806.

Fontanillas P, Hartl DL, Reuter M. 2007. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. *PLoS Genet.* 3:e210.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19: 92–105.

Fujita PA, Rhead B, Zweig AS. 2011. The UCSC genome browser database: update 2011. *Nucleic Acids Res.* 39:D876–D882.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86.

Hackam DG. 2007. Translating animal research into clinical benefit. *BMJ.* 334:163–164.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.* 33:2374–2383.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486–487.

Jiang Z, Chu PG, Woda BA, Liu Q, Balaji KC, Rock KL, Wu CL. 2008. Combination of quantitative IMP3 and tumor stage: a new system to predict metastasis for patients with localized renal cell carcinomas. *Clin Cancer Res.* 14:5579–5584.

Jiang Z, Lohse CM, Chu PG, Wu CL, Woda BA, Rock KL, Kwon ED. 2008. Oncofetal protein IMP3: a novel molecular marker that predicts metastasis of papillary and chromophobe renal cell carcinomas. *Cancer* 112:2676–2682.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet.* 10: 19–31.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 32:D493–D496.

Kleene KC. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech Dev.* 106: 3–23.

Krasnov AN, Kurshakova MM, Ramensky VE, Mardanov PV, Nabirochkina EN, Georgieva SG. 2005. A retrocopy of a gene can functionally displace the source gene in evolution. *Nucleic Acids Res.* 33:6654–6661.

Kubista M, Andrade JM, Bengtsson M, et al. (12 co-authors). 2006. The real-time polymerase chain reaction. *Mol Aspects Med.* 27:95–125.

Li WH, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* 292:237–239.

Long M, Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* 260:91–95.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2011. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.* 39:D52–D57.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A.* 95: 9407–9412.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* 3:e357.

Mason DA, Moore JD, Green SA, Liggett SB. 1999. A gain-of-function polymorphism in a G-protein coupling domain of the human beta1-adrenergic receptor. *J Biol Chem.* 274:12670–12674.

McCarrey JR, Thomas K. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* 326:501–505.

Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *FEBS Lett.* 468:109–114.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* 87:917–927.

Nei M, Rogozin IB, Piontkivska H. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A.* 97:10866–10871.

Nikolaidis N, Makalowska I, Chalkia D, Makalowski W, Klein J, Nei M. 2005. Origin and evolution of the chicken leukocyte receptor complex. *Proc Natl Acad Sci U S A.* 102:4057–4062.

Nozawa M, Aotsuka T, Tamura K. 2005. A novel chimeric gene, siren, with retroposed promoter sequence in the *Drosophila bipectinata* complex. *Genetics* 171:1719–1727.

Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38: D196–D203.

Ota T, Nei M. 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol Biol Evol.* 11:469–482.

Pan D, Zhang L. 2009. Burst of young retrogenes and independent retrogene formation in mammals. *PLoS One* 4:e5040.

Parker HG, VonHoldt BM, Quignon P, et al. (17 co-authors). 2009. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* 325:995–998.

Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. 2007. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ.* 334:197.

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* 6:e80.

Prendergast GC. 2001. Actin' up: RhoB in cancer and apoptosis. *Nat Rev Cancer.* 1:162–168.

Reid E, Connell J, Edwards TL, Duley S, Brown SE, Sanderson CM. 2005. The hereditary spastic paraplegia protein spastin interacts with the ESCRT-III complex-associated endosomal protein CHMP1B. *Hum Mol Genet.* 14:19–38.

Rodriguez V, Chen Y, Elkahloun A, Dutra A, Pak E, Chandrasekharappa S. 2007. Chromosome 8 BAC array comparative genomic hybridization and expression analysis identify amplification and overexpression of TRMT12 in breast cancer. *Genes Chromosomes Cancer* 46:694–707.

Roman T, Polanczyk GV, Zeni C, Genro JP, Rohde LA, Hutz MH. 2006. Further evidence of the involvement of alpha-2A-adrenergic receptor gene (ADRA2A) in inattentive dimensional scores of attention-deficit/hyperactivity disorder. *Mol Psychiatry.* 11:8–10.

Rosengren AH, Jokubka R, Tojjar D, et al. (16 co-authors). 2010. Overexpression of alpha2A-adrenergic receptors contributes to type 2 diabetes. *Science* 327:217–220.

Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T. 2007. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene* 389:196–203.

Sakharkar MK, Kangueane P, Petrov DA, Kolaskar AS, Subbiah S. 2002. SEGE: a database on "intron less/single exonic" genes from eukaryotes. *Bioinformatics* 18:1266–1267.

Sayers EW, Barrett T, Benson DA, et al. (42 co-authors). 2011. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 39:D38–D51.

Shiao MS, Khil P, Camerini-Otero RD, Shiroishi T, Moriwaki K, Yu HT, Long M. 2007. Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. *Mol Biol Evol.* 24: 2242–2253.

Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A. 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Mol Cell Biol.* 5: 2090–2103.

Svensson O, Arvestad L, Lagergren J. 2006. Genome-wide survey for biologically functional pseudogenes. *PLoS Comput Biol.* 2:e46.

Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I. 2011. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Mol Biol Evol.* 28:33–37.

Tan J, Dunn J, Jaeken J, Schachter H. 1996. Mutations in the MGAT2 gene controlling complex *N*-glycan synthesis cause carbohydrate-deficient glycoprotein syndrome type II, an autosomal recessive disease with defective brain development. *Am J Hum Genet.* 59: 810–817.

Thomas MA, Weston B, Joseph M, Wu W, Nekrutenko A, Tonellato PJ. 2003. Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes. *Mol Biol Evol.* 20:964–968.

Tsujikawa M, Kurahashi H, Tanaka T, Nishida K, Shimomura Y, Tano Y, Nakamura Y. 1999. Identification of the gene responsible for gelatinous drop-like corneal dystrophy. *Nat Genet.* 21:420–423.

Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. 2006. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* 7:31.

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. 2010. Can animal models of disease reliably inform human studies? *PLoS Med.* 7:e1000245.

Vanin EF. 1984. Processed pseudogenes: characteristics and evolution. *Biochim Biophys Acta.* 782:231–241.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A.* 103:3220–3225.

Wang PJ. 2004. X chromosomes, retrogenes, and their role in male reproduction. *Trends Endocrinol Metab.* 15:79–83.

Waterhouse RM, Zdobnov EM, Kriventseva EV. 2011. Correlating traits of gene retention, sequence divergence, duplicability, and essentiality in vertebrates, arthropods, and fungi. *Genome Biol Evol.* 3: 75–86.

Weiner AM, Deininger PL, Efstratiadis A. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu Rev Biochem.* 55: 631–661.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A.* 106:7273–7280.

Yano Y, Saito R, Yoshida N, Yoshiki A, Wynshaw-Boris A, Tomita M, Hirotsune S. 2004. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J Mol Med.* 82:414–422.

Yu Z, Morais D, Ivanga M, Harrison PM. 2007. Analysis of the role of retrotransposition in gene evolution in vertebrates. *BMC Bioinformatics* 8:308.

Zemojtel T, Duchniewicz M, Zhang Z, Paluch T, Luz H, Penzkofer T, Scheele JS, Zwartkruis FJ. 2010. Retrotransposition and mutation events yield Rap1 GTPases with differential signalling capacity. *BMC Evol Biol.* 10:55.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet.* 30:411–415.

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2011. A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. *Bioinformatics* 27:1749–1753.

Zhang YW, Liu S, Zhang X, et al. (17 co-authors). 2009. A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation. *Neuron* 64:328–340.

Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* 13: 2541–2558.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4:259–263.

Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.* 11:R26.

Ciomborowska J., Rosikiewicz W., Szklarczyk D.,
Makałowski W., Makałowska I.

“Orphan” Retrogenes in the Human Genome

Molecular Biology and Evolution 2013 Feb; 30(2): 384-96

- supplementary materials -

ADRA2A

Caenorhabditis elegans 90%

Homo sapiens 87%

ADRA2B

Caenorhabditis elegans 91%

Homo sapiens 95%

B3GALT6

Caenorhabditis elegans 94%

Homo sapiens 94%

CEBPB

Caenorhabditis elegans 72%

Homo sapiens 21%

CH25H

Caenorhabditis elegans 87%

Homo sapiens 96%

CHMP1B

Mus musculus 100%

Homo sapiens 100%

EXOC8

Drosophila melanogaster 96%

Homo sapiens 96%

IMP3

Danio rerio 95%

Homo sapiens 93%

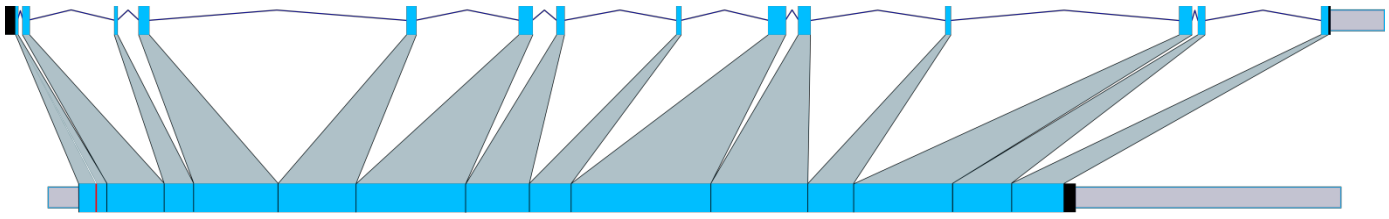
KTI12

Danio rerio 97%

Homo sapiens 76%

LCMT2

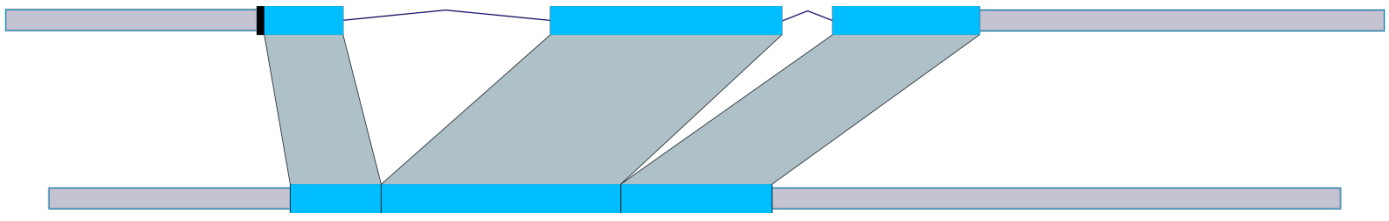
Gallus gallus 91%



Homo sapiens 98%

MAB21L1

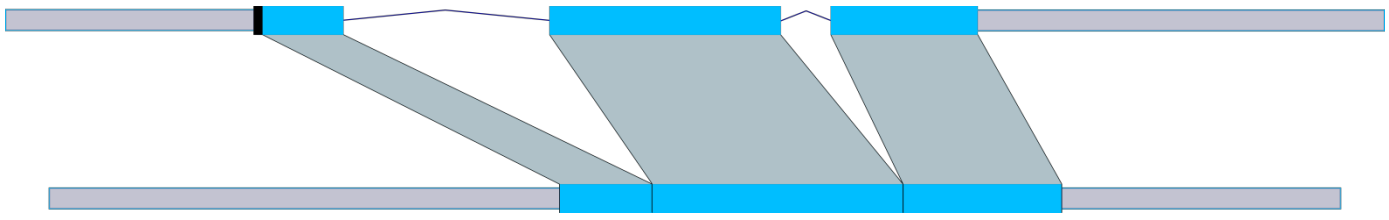
Drosophila melanogaster 98%



Homo sapiens 100%

MAB21L2

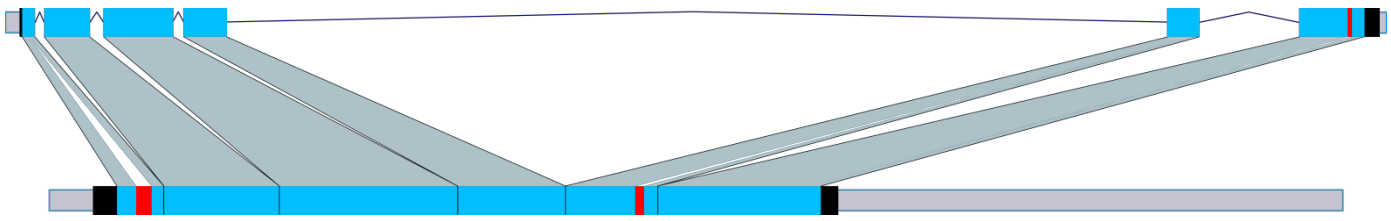
Drosophila melanogaster 98%



Homo sapiens 100%

MARS2

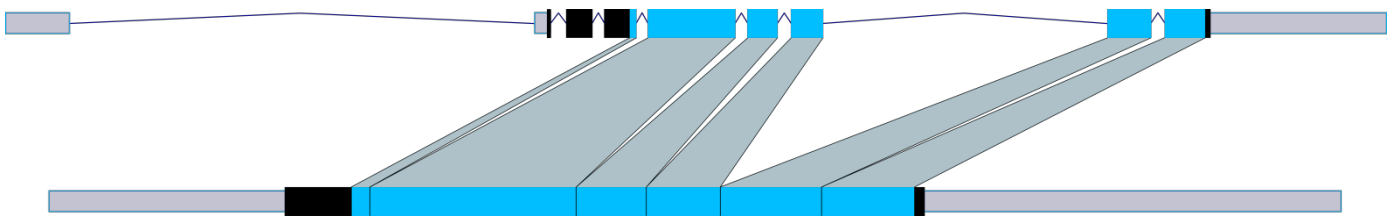
Drosophila melanogaster 94%



Homo sapiens 95%

MGAT2

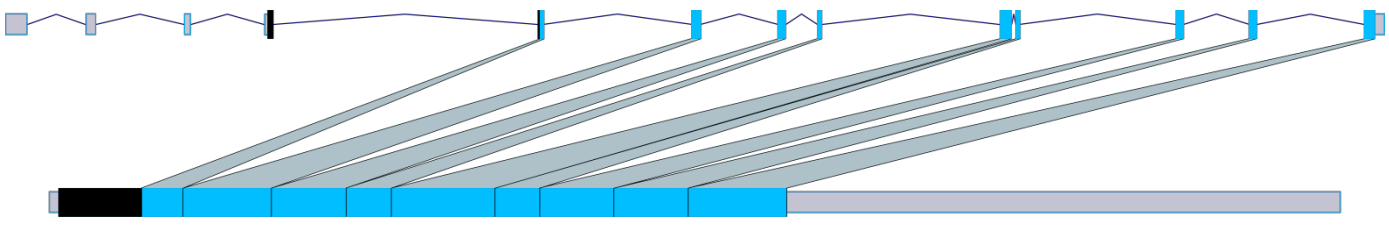
Drosophila melanogaster 80%



Homo sapiens 88%

MOCS3

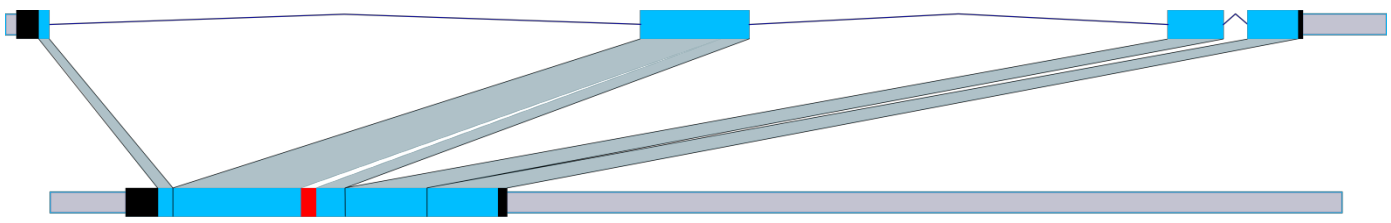
Xenopus laevis 90%



Homo sapiens 89%

PIGM

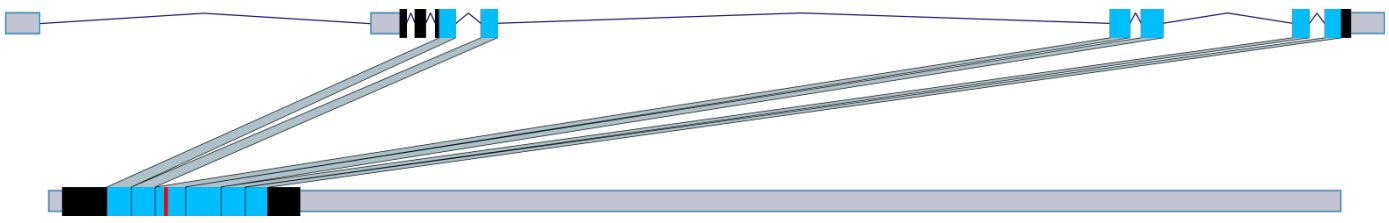
Danio rerio 96%



Homo sapiens 87%

PURA

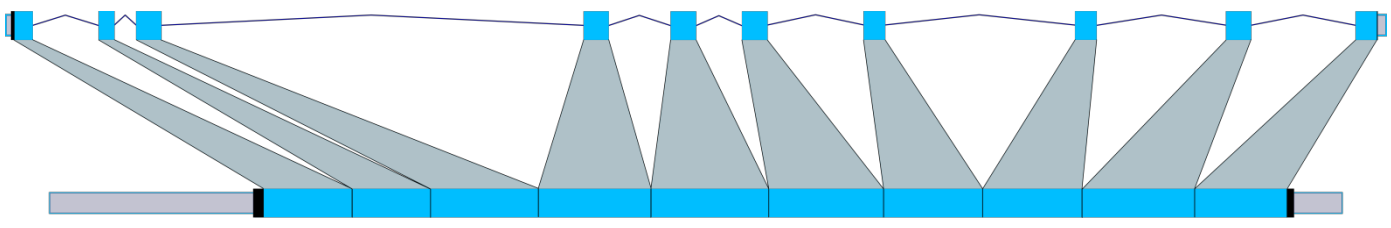
Drosophila melanogaster 78%



Homo sapiens 72%

RNF113A

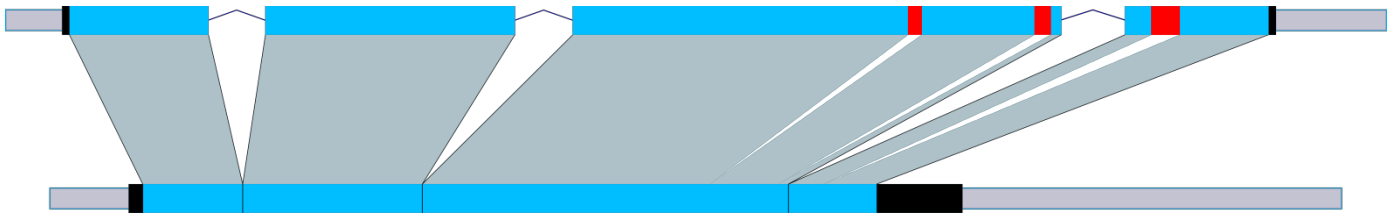
Gallus gallus 98%



Homo sapiens 94%

RRS1

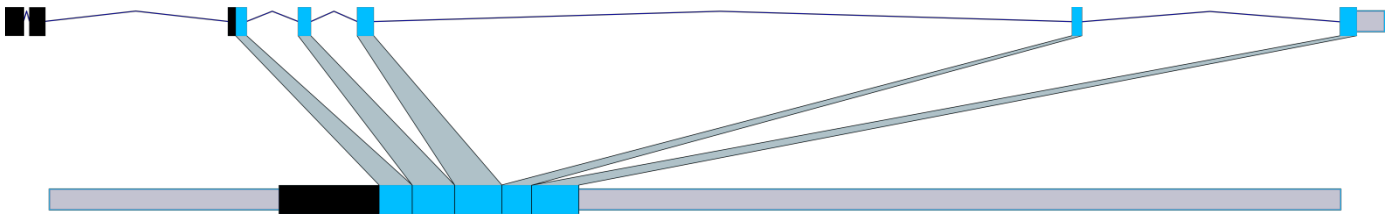
Drosophila melanogaster 98%



Homo sapiens 88%

SFT2D3

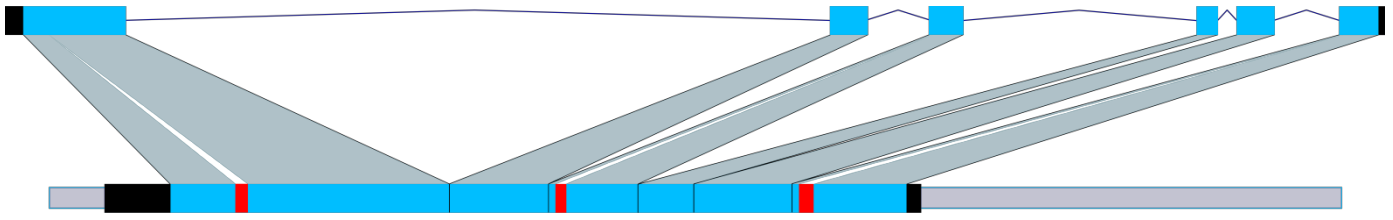
Caenorhabditis elegans 62%



Homo sapiens 67%

TBCC

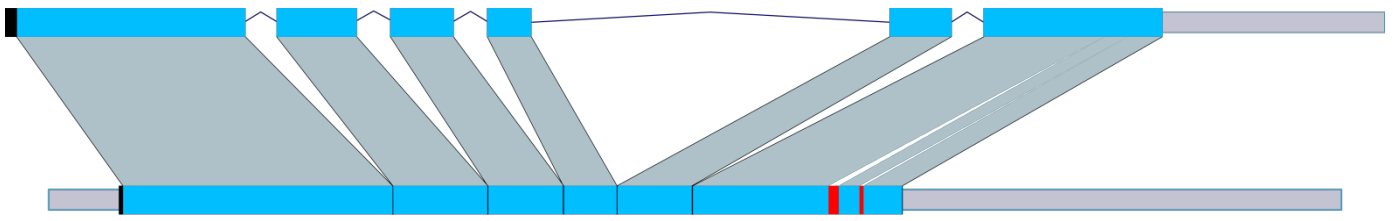
Caenorhabditis elegans 92%



Homo sapiens 90%

TRMT12

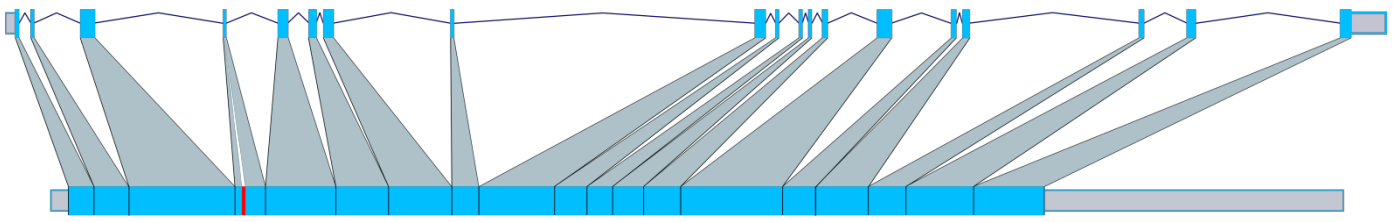
Ornithorhynchus anatinus 93%



Homo sapiens 97%

TTC30B

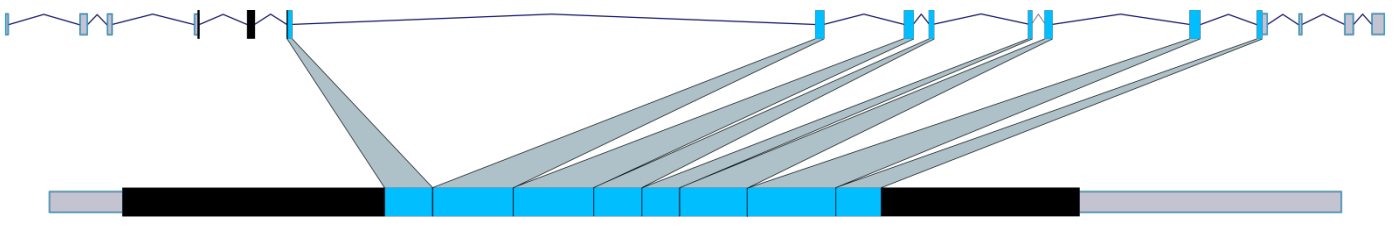
Xenopus laevis 99%



Homo sapiens 99%

UTP3

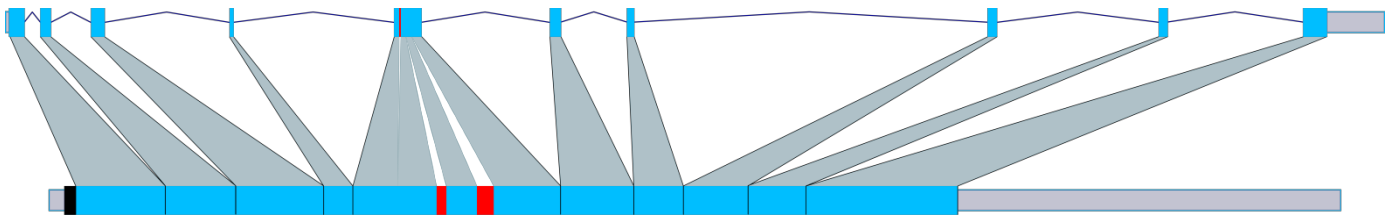
Xenopus laevis 86%



Homo sapiens 53%

ZNF830

Gallus gallus 100%



Homo sapiens 99%