Rozprawa doktorska pt.

Identyfikacja i wielkoskalowa analiza retrogenów w genomach zwierzęcych

(ang. Identification and large-scale analysis of retrogenes in animal genomes)

Michał Kabza

Promotor: prof. dr hab. Izabela Makałowska

Zakład Bioinformatyki

Instytut Biologii Molekularnej i Biotechnologii

Wydział Biologii

Uniwersytet im. Adama Mickiewicza w Poznaniu

Poznań 2016

Składam serdeczne podziękowania:

Pani prof. dr hab. Izabeli Makałowskiej

za przekazaną wiedzę, życzliwość, cierpliwość oraz wspaniałą atmosferę panującą w Zakładzie Bioinformatyki

Panu prof. dr hab. Pawłowi Golikowi

oraz Panu prof. dr hab. Wiesławowi Babikowi

za wysiłek włożony w recenzowanie niniejszej rozprawy

Panu prof. dr hab. Bogdanowi Jackowiakowi, Dziekanowi Wydziału Biologii

Dyrekcji, Pracownikom i Doktorantom Instytutu Biologii Molekularnej i Biotechnologii

Koleżankom i Kolegom z Zakładu Bioinformatyki

w szczególności dr Joannie Ciomborowskiej, dr Michałowi Szcześniakowi, mgr Elżbiecie Kaja, mgr Wojciechowi Rosikiewiczowi oraz mgr Magdalenie Kubiak

Szczególne wyrazy wdzięczności kieruję w stronę mojej Rodziny

Publikacje stanowiące podstawę rozprawy doktorskiej

1. **Kabza M**, Ciomborowska J, Makałowska I.
   RetrogeneDB--a database of animal retrogenes.
   Mol Biol Evol. 2014 Jul;31(7):1646-8.

2. **Kabza M**, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polański A, Makałowska I.
   Inter-population Differences in Retrogene Loss and Expression in Humans.
   PLoS Genet. 2015 Oct 16;11(10):e1005579.

Finansowanie

Niniejsza praca powstała przy finansowym udziale:

1. Narodowego Centrum Nauki (grant 2013/09/N/NZ2/01221 dla M.K.)
2. KNOW Poznańskie Konsorcjum RNA

Spis treści

I. STRESZCZENIE

Streszczenie

W mojej pracy doktorskiej skupiłem się głównie na wielkoskalowej identyfikacji retrokopii w genomach zwierzęcych i analizie ich potencjalnej funkcjonalności, jak również na zdarzeniach ewolucyjnych wpływających na repertuar retrogenów obecnych w ludzkich genomach. Retrokopie to kopie istniejących genów powstałe na drodze retropozycji, procesu w którym obecne w komórce cząsteczki RNA ulegają odwrotnej transkrypcji i są następnie wklejane do genomu. W przeciwieństwie do kopii genów powstałych na drodze duplikacji DNA, retrokopie przez długi czas były zwykle klasyfikowane jako pseudogeny, głównie ze względu na utratę elementów regulatorowych swych genów rodzicielskich (Kaessmann et al. 2009). Dopiero liczne prace w okresie ostatnich 20 lat podważyły ten stan rzeczy i dowiodły, że retrokopie mogą stać się funcjonalnymi genami (retrogenami) i wykazywać szeroki zakres funkcjonalności (Kaessmann et al. 2009). Niektóre retrogeny zachowują otwartą ramkę odczytu swojego genu rodzicielskiego i funkcjonują jako geny kodujące białko (Brosius 1999). Retrogeny mogą również stracić swój potencjał kodujący i działać jako długie niekodujące RNA (Dai et al. 2008), gąbki miRNA (Poliseno et al. 2010) lub źródło krótkich regulatorowych RNA (Kaessmann et al. 2009). Poza swoją rolą w formowaniu nowych genów, retrokopie okazały się doskonałymi markerami ewolucyjnymi, dającymi nam wgląd w procesy takie jak powstawanie chromosomów płci u ssaków (Potrzebowski et al. 2008) czy zmiany aktywności retrotranspozonów typu LINE w trakcie ewolucji (Kaessmann et al. 2009).

W pierwszej z przedstawionych publikacji opisuję RetrogeneDB, nową bazę danych zawierającą adnotacje retrokopii. W przeciwieństwie do poprzednich tego typu baz danych, takich jak HOPPSIGEN (Khelifi et al. 2005) czy RCPedia (Navarro and Galante 2013), RetrogeneDB nie ogranicza się jedynie do wybranych organizmów modelowych i zawiera przewidywania retrokopii dla 62 genomów zwierzęcych pobranych z bazy danych Ensembl (wydanie 73). Stworzony przeze mnie potok analityczny do wykrywania retrokopii użyty w bazie RetrogeneDB oparty jest o przyrównanie całego proteomu danego organizmu do jego genomu i dalsze poszukiwanie oznak retropozycji (Zhang et al. 2011, Navarro and Galante 2013) wśród znalezionych rejonów podobieństwa. Podejście oparte o przyrównanie proteomu zostało wybrane ponieważ generuje ono bardziej konserwatywne wyniki niż w przypadku wykorzystania przyrównań sekwencji mRNA, co czyni je bardziej odpowiednim do analizy relatywnie słabo zaadnotowanych genomów zsekwencjonowanych z niskim stopniem pokrycia. Dane dotyczące wykrytych retrokopii uzupełnione są o dodatkowe informacje na temat ich potencjalnej funkcjonalności, wliczając w to

zachowanie otwartej ramki odczytu czy oszacowania poziomu ekspresji retrokopii w oparciu o dane RNA–Seq. Baza RetrogeneDB jest obecnie dostępna pod adresem http://retrogenedb.amu.edu.pl i zawiera łącznie 84 808 przewidzianych retrokopii, z czego 64 225 nie jest obecnych w bazie Ensembl. Interfejs WWW bazy danych pozwala na przeglądanie jej zasobów, przeszukiwanie ich za pomocą licznych kryteriów, pobrane zawartości bazy w postaci plików tekstowych i poszukiwanie podobieństwa sekwencji za pomocą programu BLAST.

Druga z publikacji stanowiących podstawę mojej pracy doktorskiej opisuje z kolei analizę międzypopulacyjnych różnic dotyczących repertuaru retrokopii w genomie człowieka i ich ekspresji. W odróżnieniu od wcześniejszych analiz (Abyzov et al. 2013, Ewing et al. 2013, Schrider et al. 2013), skoncentrowanych głównie na identyfikacji nowych zjawisk retropozycji, moim głównym celem było wykrycie utraty ancestralnych, funkcjonalnych retrokopii (retrogenów) w różnych ludzkich populacjach. Aby osiągnąć ten cel, wykorzystałem genomowe warianty strukturalne wykryte w ramach Fazy 1 Projektu Sekwencjonowania 1000 Genomów (1000 Genomes Project Consortium et al. 2012) i porównałem ich pozycje z adnotacjami ludzkich retrokopii z bazy RetrogeneDB. Ekspresja retrokopii, służąca jako wyznacznik ich funkcjonalności, została oszacowana w oparciu o dane RNA–Seq z zestawów danych Illumina Bodymap 2 i Geuvadis Consortium (Lappalainen et al. 2013). Obecność retrokopii w genomach innych naczelnych została zbadana z wykorzystanem nowego podejścia bazującego na przyrównaniu całych genomów PECAN (Paten et al. 2009) dostępnym w bazie danych Ensembl. Łącznie byłem w stanie zidentyfikować 214 długich delecji pokrywających się z 190 retrokopiami, z których 19 wykazywało potencjalną funkcjonalność (tzn. aktywność transkrypcyjną). Jedenaście spośród tych retrogenów okazało się być ancestralne, a ich utratę wytłumaczyć można jedynie delecją genomową, nie zaś nową retropozycją. Ekspresja każdego z tych retrogenów została potwierdzona przez moją koleżankę, Magdalenę Kubiak, za pomocą metody RT–PCR. Przeprowadziłem również analizę ekspresji różnicowej genów między pięcioma populacjami (CEU, GBR, FIN, TSI, YRI), wykrywając 9 retrokopii o statystycznie istotnych różnicach w poziomie ekspresji. Dodatkowo, we współpracy z Politechniką Śląską stworzyłem nową metodę identyfikacji retrokopii nieobecnych w genomie referencyjnym, opartą o składanie de novo krótkich odczytów, które nie zostały zmapowane do genomu. Zastosowanie tej metody do analizy genomów 500 osobników zsekwencjonowanych w ramach Fazy 1 Projektu Sekwencjonowania 1000 Genomów pozwoliło na odkrycie 5 nowych, dotychczas nieznanych retrokopii, z których trzy okazały się ancestralne i polimorficzne (tzn. nie są obecne w genomach wszystkich badanych osobników).

Referencje

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.
- Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L; 1000 Genomes Project Consortium, Lee C, Gerstein M. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res. 2013 Dec;23(12):2042-52.
- Brosius J. Many G-protein-coupled receptors are encoded by retrogenes. Trends Genet. 1999 Aug;15(8):304-5.
- Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, Eyre-Walker A, Du W, Long M. The evolution of courtship behaviors through the origination of a new gene in Drosophila. Proc Natl Acad Sci U S A. 2008 May 27;105(21):7478-83.
- Ewing AD, Ballinger TJ, Earl D; Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L, Wilson RK, Haussler D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 2013 Mar 13;14(3):R22.
- Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009 Jan;10(1):19-31.
- Khelifi A, Duret L, Mouchiroud D. HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D59-66.
- Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM; Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 26;501(7468):506-11.
- Navarro FC, Galante PA. RCPedia: a database of retrocopied genes. Bioinformatics. 2013 May 1;29(9):1235-7.

- Navarro FC, Galante PA. A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biol Evol. 2015 Jul 29;7(8):2265-75.
- Paten B, Herrero J, Beal K, Birney E. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics. 2009 Feb 1;25(3):295-301.
- Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010 Jun 24;465(7301):1033-8.
- Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. 2008 Apr 1;6(4):e80.
- Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 2013;9(1):e1003242.
- Zhang YE, Vibranovski MD, Krinsky BH, Long M. A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. Bioinformatics. 2011 Jul 1;27(13):1749-53.

II. SUMMARY

Summary

In my PhD thesis I focused on the large–scale identification of retrocopies in animal genomes and the analysis of their potential functionality, as well as on the evolutionary events affecting retrocopy repertoire in human genomes. Retrocopies are copies of existing genes created by retroposition, a process of reverse transcription of cellular RNAs and their subsequent insertion into the genome. Unlike gene copies created by DNA–mediated duplication, retroposed genes used to be routinely classified as pseudogenes, mostly because of the loss of regulatory elements of their parental genes (Kaessmann et al. 2009). Numerous studies in the past 20 years, however, challenged that view and proved that retrocopies can act as functional genes (retrogenes) and exhibit a very broad range of functionalities (Kaessmann et al. 2009). Some retrogenes may retain the open reading frames of their parental genes and function as protein coding genes (Brosius 1999). Retrogenes can also lose their coding potential and act as noncoding RNAs (Dai et al. 2008), miRNA sponges (Poliseno et al. 2010) or the source of short interfering RNAs (Kaessmann et al. 2009). Besides their role in formation of new genes, retrocopies are extremely useful markers of evolutionary history, providing insight into such phenomena as origination of mammalian sex chromosomes (Potrzebowski et al. 2008) or changes in the activity of LINE retrotransposons during evolution (Kaessmann et al. 2009).

In the first publication I present RetrogeneDB, a new database containing retrocopy annotations. Unlike previous similar databases, such as HOPPSIGEN (Khelifi et al. 2005) or RCPedia (Navarro and Galante 2013), RetrogeneDB is not limited to model organisms and contains retrocopy predictions for 62 animal genomes downloaded from Ensembl 73 databases. I created a retrocopy identification pipeline that is based on the alignment of the whole proteome to the genome and further screening of identified matches for the signs of retroposition (Zhang et al. 2011, Navarro and Galante 2013). Proteome–based approach was chosen because such method produces more conservative results than retrocopy detection utilizing mRNA sequence alignments (Navarro and Galante 2015), which makes it suitable for relatively poorly annotated, low–coverage genomes. Detected retrocopies are supplemented by additional data regarding their potential functionality, including information on open reading frame conservation and expression estimations based on RNA–Seq data. RetrogeneDB is currently located at http://retrogenedb.amu.edu.pl and overall contains 84 808 retrocopies, 64 225 of which are not annotated in the Ensembl databases. The web interface allows users to browse or searched database resources using numerous criteria, download all data as text files, and search for sequence similarity using BLAST.

The second publication forming the basis of my PhD thesis describes the analysis of inter–population differences in retrocopy repertoire and expression. In contrast to previous studies (Abyzov et al. 2013, Ewing et al. 2013, Schrider et al. 2013), mostly focused on detection of novel retroposition events, my primal goal was to detect the loss of ancestral, functional retrocopies (retrogenes) in different human populations. In order to achieve that, I used the structural variants detected during Phase I of the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2012) and compared their location to human retrocopies annotated in RetrogeneDB. Retrocopy expression, and therefore potential functionality, was assessed using Illumina Bodymap 2 and Geuvadis Consortium RNA–Seq datasets (Lappalainen et al. 2013). The evolutionary conservation of retrocopies in different primate species was established using a novel approach based on PECAN whole genome alignment (Paten et al. 2009) available in Ensembl databases. Overall, I was able to detect 214 long deletions that affected 190 retrocopies, 19 of which are potentially functional (i.e. show signs of transcriptional activity). Eleven of these retrogenes turned out to be ancestral and therefore their loss could only be attributed to genomic deletion, not novel retroposition. The expression of each of these 11 retrocopies was experimentally confirmed by my colleague, Magdalena Kubiak, using RT–PCR. I also performed differential gene expression analysis between five populations (CEU, GBR, FIN, TSI, YRI), detecting 9 retrocopies that show statistically significant differences in expression level. Finally, in cooperation with Silesian University of Technology, I created the new method of detecting novel retrocopies absent from human reference genome, based on de novo assembly of short reads that failed to map to the genome. Application of this method to the genomes of 500 individuals sequenced during Phase I of the the 1000 Genomes Project yielded 5 novel, previously unknown retroposed gene copies, three of which turned out to be ancient and polymorphic (i.e. absent in the genomes of certain individuals).

References

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.
- Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L; 1000 Genomes Project Consortium, Lee C, Gerstein M. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res. 2013 Dec;23(12):2042-52.
- Brosius J. Many G-protein-coupled receptors are encoded by retrogenes. Trends Genet. 1999

Aug;15(8):304-5.

- Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, Eyre-Walker A, Du W, Long M. The evolution of courtship behaviors through the origination of a new gene in Drosophila. Proc Natl Acad Sci U S A. 2008 May 27;105(21):7478-83.
- Ewing AD, Ballinger TJ, Earl D; Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L, Wilson RK, Haussler D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 2013 Mar 13;14(3):R22.
- Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009 Jan;10(1):19-31.
- Khelifi A, Duret L, Mouchiroud D. HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D59-66.
- Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM; Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 26;501(7468):506-11.
- Navarro FC, Galante PA. RCPedia: a database of retrocopied genes. Bioinformatics. 2013 May 1;29(9):1235-7.
- Navarro FC, Galante PA. A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biol Evol. 2015 Jul 29;7(8):2265-75.
- Paten B, Herrero J, Beal K, Birney E. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics. 2009 Feb 1;25(3):295-301.
- Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010 Jun 24;465(7301):1033-8.
- Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H. Chromosomal gene movements reflect the recent origin and biology of therian sex

chromosomes. PLoS Biol. 2008 Apr 1;6(4):e80.

- Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 2013;9(1):e1003242.
- Zhang YE, Vibranovski MD, Krinsky BH, Long M. A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. Bioinformatics. 2011 Jul 1;27(13):1749-53.

III. OŚWIADCZENIE DOKTORANTA

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

Wydział Biologii

OŚWIADCZENIE DOKTORANTA DOTYZĄCE JEGO UDZIAŁU W POWSTAWANIU PRAC NAUKOWYCH STANOWIĄCYCH ROZPRAWĘ DOKTORSKĄ

1. RetrogeneDB—A Database of Animal Retrogenes
(**Michał Kabza**, Joanna Ciomborowska, Izabela Makałowska*)

Rola doktoranta: Pierwszy autor

Udział doktoranta w badaniach opisywanych w tej publikacji obejmował:

- udział w opracowaniu koncepcji badań
- opracowanie oraz implementację potoku analitycznego służacego do identyfikacji retrokopii
- analizę ekspresji wykrytych retrokopii
- stworzenie bazy danych i interfejsu webowego umożliwiającego dostęp do jej zasobów
- współudział w przygotowaniu manuskryptu

Udział procentowy: 80%

2. Inter-population Differences in Retrogene Loss and Expression in Humans
(**Michał Kabza***, Magdalena Regina Kubiak, Agnieszka Danek, Wojciech Rosikiewicz, Sebastian Deorowicz, Andrzej Polański, Izabela Makałowska*)

Rola doktoranta: Pierwszy autor

Udział doktoranta w badaniach opisywanych w tej publikacji obejmował:

- udział w opracowaniu koncepcji badań
- przeprowadzenie analizy ekspresji i konserwacji ewolucyjnej retrokopii
- przeprowadzenie analizy ancestralnych, funkcjonalnych retrokopii ulegających delecji w różnych ludzkich populacjach
- wykrywane nowych retrokopii nieobecnych w genomie człowieka w oparciu o odczyty niemapujące się do genomu referencyjnego
- współudział w przygotowaniu manuskryptu

Udział procentowy: 45%

Poznań, dnia 11.01.2016

Michał Kabza

mgr Michał Kabza
Zakład Bioinformatyki
Wydział Biologii UAM w Poznaniu

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

Wydział Biologii

W związku z wykorzystaniem przez mgr Michała Kabzę publikacji pt. „RetrogeneDB – a database of animal retrogenes" jako elementu rozprawy doktorskiej, a opublikowanej przed otwarciem przewodu doktorskiego, oświadczamy, że praca ta wykonana została całkowicie pod opieką promotora, prof. dr hab. Makałowskiej. Prof. dr hab. Makałowska pełniła obowiązki opiekuna naukowego od początku studiów doktoranckich, a wyżej wymieniona publikacja stanowiła element realizowanego projektu i wykonana została na poczet rozprawy doktorskiej. Jednocześnie oświadczamy, że publikacja ta nie stanowiła podstawy otwarcia przewodu doktorskiego mgr Michała Kabzy, jako że w chwili otwarcia jego dorobek naukowy obejmował dwie inne wymienione poniżej publikacje naukowe:

1. **ERISdb: a database of plant splice sites and splicing signals**
   (Szcześniak MW, **Kabza M**, Pokrzywa R, Gudyś A, Makałowska I)
   Plant Cell Physiol. 2013 Feb;54(2):e10.

2. **Ewolucja struktury genów**
   (Makałowska I, **Kabza M**, Ciomborowska J)
   Kosmos 2009, 1-2(282-283):5-16

Poznań, dnia 15.03.2016

Michał Kabza

mgr Michał Kabza
Zakład Bioinformatyki
Wydział Biologii UAM w Poznaniu

Poznań, dnia 18.03.2016

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

IV. OŚWIADCZENIA WSPÓŁAUTORÓW

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

Wydział Biologii

W związku z wykorzystaniem przez mgr Michała Kabzę poniżej wymienionych publikacji jako rozprawy doktorskiej oświadczam, iż udział mój, jako promotora, polegał przede wszystkim na wspólnym opracowaniu koncepcji badań i przygotowaniu manuskryptów. Jednocześnie stwierdzam, iż we wszystkich wymienionych publikacjach wkład pracy mgr Michała Kabzy był niezwykle duży. Przeprowadzone przez mgr Kabzę analizy i opracowania wyników były fundamentalne dla powstania poniżej wymienionych prac.

1. **RetrogeneDB—A Database of Animal Retrogenes**
   (Michał Kabza, Joanna Ciomborowska, **Izabela Makałowska***)
   Udział procentowy: 10%

2. **Inter-population Differences in Retrogene Loss and Expression in Humans**
   (Michał Kabza*, Magdalena Regina Kubiak, Agnieszka Danek, Wojciech Rosikiewicz, Sebastian Deorowicz, Andrzej Polański, **Izabela Makałowska***)
   Udział procentowy: 10%

Poznań, dnia 11.01.2016

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

UAM

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

Wydział Biologii

RetrogeneDB—A Database of Animal Retrogenes

(Michał Kabza, **Joanna Ciomborowska**, Izabela Makałowska*)

Rola: współautor

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

- testowanie bazy danych RetrogeneDB
- współudział w przygotowaniu manuskryptu

Udział procentowy: 10%

Poznań, dnia 11.01.2016

JCiomborowska

dr Joanna Ciomborowska
Zakład Bioinformatyki
Wydział Biologii UAM w Poznaniu

UAM

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

Wydział Biologii

Inter-population Differences in Retrogene Loss and Expression in Humans

(Michał Kabza*, **Magdalena Regina Kubiak**, Agnieszka Danek, Wojciech Rosikiewicz, Sebastian Deorowicz, Andrzej Polański, Izabela Makałowska*)

Rola: współautor

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

- wykonanie eksperymentów (PCR, real–time PCR)
- współudział w interpretacji danych otrzymanych w wyniku analiz i eksperymentów
- współudział w przygotowaniu manuskryptu

Udział procentowy: 20%

Poznań, dnia 11.01.2016

Magdalena Kubiak

mgr Magdalena Regina Kubiak
Zakład Bioinformatyki
Wydział Biologii UAM w Poznaniu

Inter-population Differences in Retrogene Loss and Expression in Humans

(Michał Kabza*, Magdalena Regina Kubiak, **Agnieszka Danek**, Wojciech Rosikiewicz, Sebastian Deorowicz, Andrzej Polański, Izabela Makałowska*)

Rola: współautor

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

- pomoc podczas przeprowadzenia części analiz komputerowych (wykrycie długich delecji nakładających się z retrokopiami, składanie de novo krótkich odczytów niezmapowanych do genomu)

Udział procentowy: 10%

Gliwice, dnia 12.01.2016

Agnieszka Danek

dr inż. Agnieszka Danek
Instytut Informatyki
Politechnika Śląska

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

Inter-population Differences in Retrogene Loss and Expression in Humans

(Michał Kabza*, Magdalena Regina Kubiak, Agnieszka Danek, **Wojciech Rosikiewicz**, Sebastian Deorowicz, Andrzej Polański, Izabela Makałowska*)

Rola: współautor

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

- współudział w interpretacji danych otrzymanych w wyniku analiz i eksperymentów
- współudział w przygotowaniu manuskryptu

Udział procentowy: 5%

Poznań, dnia 11.01.2016

W. Rosikiewicz

mgr Wojciech Rosikiewicz
Zakład Bioinformatyki
Wydział Biologii UAM w Poznaniu

Inter-population Differences in Retrogene Loss and Expression in Humans

(Michał Kabza*, Magdalena Regina Kubiak, Agnieszka Danek, Wojciech Rosikiewicz, **Sebastian Deorowicz**, Andrzej Polański, Izabela Makałowska*)

Rola: współautor

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

- dostarczenie narzędzi komputerowych i serwera do przeprowadzenia części analiz komputerowych
- koordynowanie badań związanych z identyfikacją długich delecji i składaniem de novo krótkich odczytów niezmapowanych do genomu

Udział procentowy: 5%

Gliwice, dnia 12.01.2016

……………………………………

dr hab. inż. Sebastian Deorowicz
Instytut Informatyki
Politechnika Śląska

Inter-population Differences in Retrogene Loss and Expression in Humans

(Michał Kabza*, Magdalena Regina Kubiak, Agnieszka Danek, Wojciech Rosikiewicz, Sebastian Deorowicz, **Andrzej Polański**, Izabela Makałowska*)

Rola: współautor

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

- dostarczenie narzędzi komputerowych i serwera do przeprowadzenia części analiz komputerowych
- koordynowanie badań związanych z identyfikacją długich delecji i składaniem de novo krótkich odczytów niezmapowanych do genomu

Udział procentowy: 5%

Gliwice, dnia ....30 XII 2015

....

prof. dr hab. inż. Andrzej Polański
Instytut Informatyki
Politechnika Śląska w Gliwicach

V. PUBLIKACJE STANOWIĄCE PODSTAWĘ ROZPRAWY DOKTORSKIEJ

Kabza M, Ciomborowska J, Makałowska I.

RetrogeneDB--a database of animal retrogenes.

Molecular Biology and Evolution 2014 Jul;31(7):1646-8.

RetrogeneDB—A Database of Animal Retrogenes

Michał Kabza,[1] Joanna Ciomborowska,[1] and Izabela Makałowska*,[1]

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

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

Associate editor: Naruya Saitou

Abstract

Retrocopies of protein-coding genes, reverse transcribed and inserted into the genome copies of mature RNA, have commonly been categorized as pseudogenes with no biological importance. However, recent studies showed that they play important role in the genomes evolution and shaping interspecies differences. Here, we present RetrogeneDB, a database of retrocopies in 62 animal genomes. RetrogeneDB contains information about retrocopies, their genomic localization, parental genes, ORF conservation, and expression. To our best knowledge, this is the most complete retrocopies database providing information for dozens of species previously never analyzed in the context of protein-coding genes retroposition. The database is available at http://retrogenedb.amu.edu.pl.

Key words: **retroposition, gene duplication, retrogene, database.**

Retrogenes, for a long time considered to be not important copies of parental genes are nowadays called “seeds of the evolution,” because they made a significant contribution to genomes evolution (Brosius 1991). It has been shown that they play very important role in the diversification of transcriptomes and proteomes and may be responsible for the wealth of species-specific features (Betrán et al. 2002; Balasubramanian et al. 2009; Szcześniak et al. 2011). As duplicates of their parental genes, they evolve relatively fast, so these genes may acquire novel functions. Retrocopies of protein-coding genes are also known to be involved in many diseases (Prendergast 2001; Ciomborowska et al. 2013).

Analyses of retroduplications have been mostly limited to the few mammalian model species (mainly human and mouse) and fruit fly (Kaessmann et al. 2009). Nonmammalian vertebrates have been largely overlooked in retrocopies studies, and our knowledge of their evolution in other animals is even more limited. Although retrocopies are annotated in major genomic databases (Ensembl [Flicek et al. 2014], UCSC Genome Browser [Meyer et al. 2013], National Center for Biotechnology Information Gene [Maglott et al. 2011]), they are often annotated just as “pseudogenes,” the same way as duplicates originated via DNA-based mechanisms. The same problem refers to more specialized database Pseudogene.org (www.pseudogene.org, last accessed January 2014). The most complete retrocopies’ annotations are in Ensembl database; although they are very good for human and mouse, the quality is very poor for remaining genomes. There are only two databases fully dedicated to retrocopies: RCPedia (Navarro and Galante 2013) and HOPPSIGEN (Khelifi et al. 2005). However, the first one contains data only for a few primate species, and the latter is limited to human and mouse.

We have analyzed genomes of 62 animal species to identify retrocopies. The search was done based on the similarities between reference genomic sequence and proteins coded by multiexon genes in a given species. To increase accuracy, we applied several criteria to call a genomic region a retrocopy: Length of the alignment at least 150 bp, minimum of 50% coverage of parental gene, minimum of 50% identity, and loss of at least two introns among others (for details see supplementary file S1, Supplementary Material online). Resulting data set was additionally manually inspected to exclude potential false positives, especially copies of transposons annotated as protein-coding genes, which in some genomes totaled for as many as few thousands. Our strategy led to identification of 84,808 retrocopies, including 6,277 protein-coding genes not recognized previously as retrogenes. A total of 64,225 retrocopies identified by us are not present in the Ensembl database, this includes 139 retrocopies in the human and as many as 2,205 in the mouse genome, which belong to the best annotated. Because of our stringent requirements, applied in the order to generate a high-quality data set, the number of identified retrocopies in a given species is considerably lower than in most other databases. However, this method gave consistently good results in both, well and poorly annotated, low-coverage genomes, for example, alpaca or dolphin.

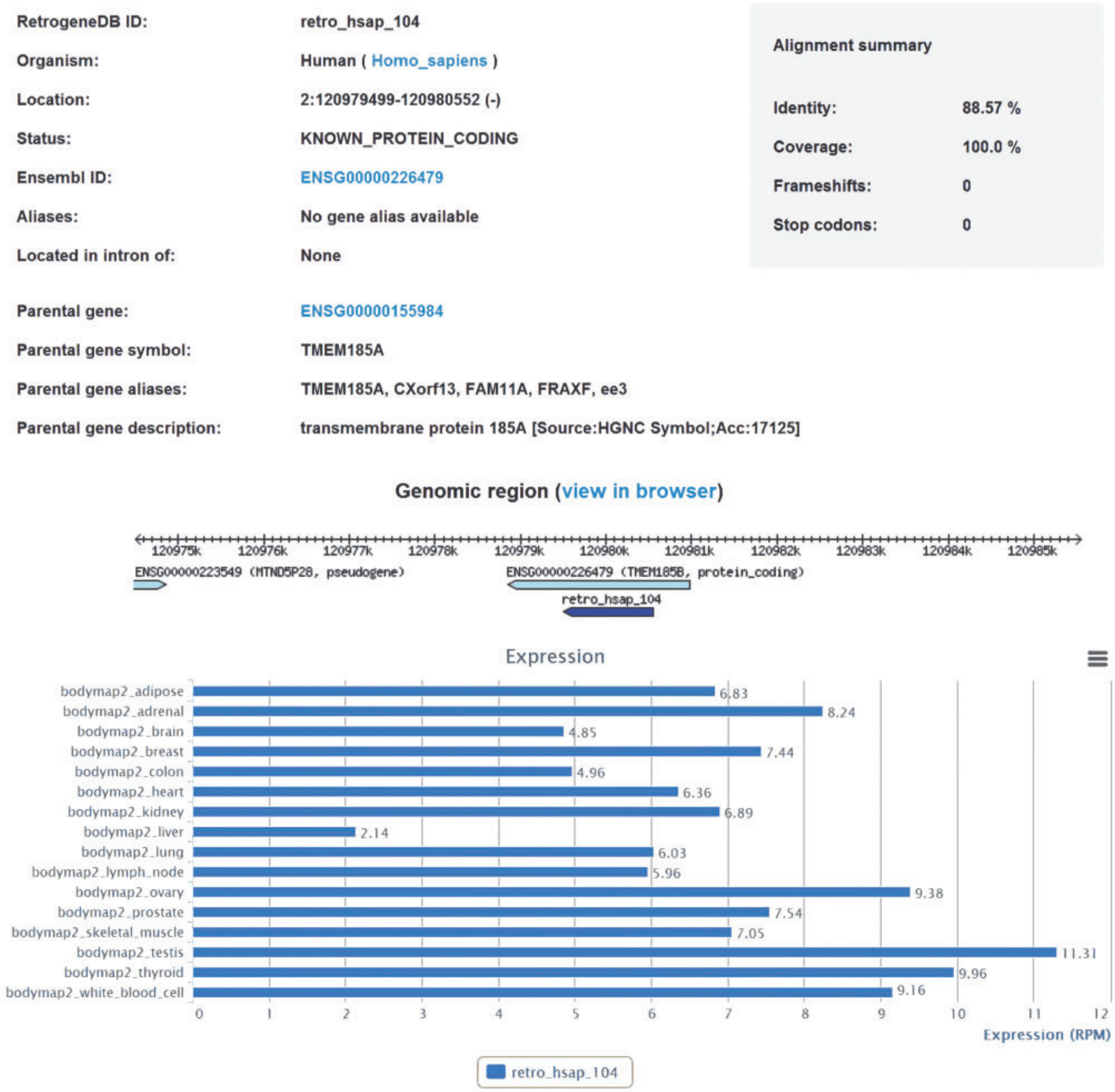
The number of retrocopies differs significantly even between closely related species, for example, 4,927 in human vs. 3,285 in chimpanzee. This may be resulting from differences in annotations and from species-specific retroposition events. In addition, retrocopies are polymorphic and higher number of retrocopies in human (vs. chimpanzee) may reflect a large amount of human population data (Abyzov et al. 2013).

Retrocopies, as a second copy of the existing gene, evolve relatively quickly and accumulate mutations. However, many of them gain functionality and become subjected to purifying selection (Vinckenbosch et al. 2006; Yu et al. 2007). We compared retrocopies with their progenitors to single out those with conserved ORF, that is, without internal stop codons or frameshifts over the entire alignment. Conserved ORFs in

Downloaded from http://mbe.oxfordjournals.org/ by guest on October 8, 2015

© The Author 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access



RetrogeneDB ID:	retro_hsap_104
Organism:	Human (Homo_sapiens)
Location:	2:120979499-120980552 (-)
Status:	KNOWN_PROTEIN_CODING
Ensembl ID:	ENSG00000226479
Aliases:	No gene alias available
Located in intron of:	None
Parental gene:	ENSG00000155984
Parental gene symbol:	TMEM185A
Parental gene aliases:	TMEM185A, CXorf13, FAM11A, FRAXF, ee3
Parental gene description:	transmembrane protein 185A [Source:HGNC Symbol;Acc:17125]

Alignment summary

Identity:	88.57 %
Coverage:	100.0 %
Frameshifts:	0
Stop codons:	0

Genomic region (view in browser)

Fig. 1. Example of RetrogeneDB record with selected data.

mammals account for 10–25% of retrocopies. In nonmammalian animals, the fraction is much higher, considerably over 50% and in some species close to 100. However, the conservation of the ORF over the length of alignment does not automatically imply that a retrocopy is efficiently translated, even if it is expressed. In selected species, we also identified expressed retrocopies based on the RNA-seq data. Because of the high similarity to parental genes, in the process of reads mapping, we made sure they uniquely and perfectly map to retrocopies (supplementary file S1, Supplementary Material online). This led to the underestimation of retrocopies expression level but prevented false-positive predictions of expressed retrocopies. Approximately 10–20% of mammalian retrocopies are expressed in at least one library at minimal level of 1 RPM (reads per million mapped). In lizard, this number is higher with almost 40% of expressed retrocopies. Majority of expressed retrocopies in marsupials, egg-laying mammals, and nonmammalian species have conserved ORFs. However, in placental mammals, the fraction of expressed retrocopies with conserved ORF is lower, from only 30% in human up to 65% in horse.

All the data are stored in MySQL database (www.mysql.com, last accessed September 2013), and the web interface was developed using Django framework (www.djangoproject.com, last accessed January 2014). The database is available at http://retrogenedb.amu.edu.pl (last accessed April 26, 2014) and can be searched either from the retrocopy or the parental gene perspective. The retrocopy search can be done based on

Downloaded from http://mbe.oxfordjournals.org/ by guest on October 8, 2015

the genomic localization, key words, parental gene name, and retrocopy ID, and results can be filtered based on the retrocopy type, ORF conservation, or expression. In addition, a JBrowse genome browser was implemented allowing retrocopy inspection in the genomic context (fig. 1). The search from parental gene perspective enables to identify all retrocopies of a given gene or all orthologs, which were retroposed in any other species. Users can also perform sequence-based search using BLAST tool.

Supplementary Material

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

Acknowledgments

This work was supported by the National Science Centre of Poland grant number 2013/09/N/NZ2/01221 to M.K. and grant number 2011/01/N/NZ2/01701 to J.C. and European Union grant PIRSES-GA-2009-247633 to I.M.

References

Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L; 1000 Genomes Project Consortium, Lee C, Gerstein M. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res.* 23(12):2042–2052.

Balasubramanian S, Zheng D, Liu Y-J, Fang G, Frankish A, Carriero N, Robilotto R, Cayting P, Gerstein M. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol.* 10:R2.

Betrán 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.

Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makałowski W, Makałowska I. 2013. "Orphan" retrogenes in the human genome. *Mol Biol Evol.* 30:384–396.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749–D755.

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

Khelifi A, Duret L, Mouchiroud D. 2005. HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic Acids Res.* 33:D59–D66.

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

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* 41:D64–D69.

Navarro FC, Galante PA. 2013. RCPedia: a database of retrocopied genes. *Bioinformatics* 29:1235–1237.

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

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

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.

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

Downloaded from http://mbe.oxfordjournals.org/ by guest on October 8, 2015

Kabza M, Ciomborowska J, Makałowska I.

RetrogeneDB--a database of animal retrogenes.

Molecular Biology and Evolution 2014 Jul;31(7):1646-8.

Supplementary materials

Materials and Methods

Data

All the genomic data (including genomic sequences, protein sequences, annotations and homology relationships) used in this study were downloaded from Ensembl release 73 databases (Flicek et al. 2013). To estimate expression of both retrogenes and parental genes we utilized following short read libraries from the NCBI SRA database (Leinonen et al. 2011): ERP000546 (human), SRP007412 (mouse, chimpanzee, gorilla, orangutan, macaque, opossum), SRP021940 (horse), SRP009831 (anole lizard) and DRP000627 (coelacanth).

Retrocopy identification

In order to identify retrocopies we aligned the proteomic sequences, obtained from Ensemble database, to the hard masked genome sequence of the given organism using LAST. We used following parameters: BLOSUM62 substitution matrix, gap existence/extension penalty : 11/2, frameshift penalty : 15, drop-off value : 20. In the analysis we considered only proteins originating from protein-coding genes and we excluded all genes annotated as containing reverse transcriptase. As multiple proteins can give hits to the same genomic locus, alignments were subsequently clustered using BEDTools (Quinlan and Hall 2010) with the requirement of at least 1 bp overlap on the same strand. Clusters overlapping known protein-coding genes were removed from the analysis. For each remaining cluster (and therefore potential retrogene locus) we chose the optimal alignments (i.e. having the highest score within the cluster) and suboptimal alignment (with score of at least 98% of the optimal alignment). Optimal alignments are most likely to represent alignments of parental genes to their retrogenes. However, if parental gene has very close paralogs they may occasionally give slightly better alignments to retrogene locus than the actual parental gene. For this reason, we screened both optimal and suboptimal alignments for signs of retroposition. We considered the locus a retrogene if every optimal or suboptimal alignment within the cluster showed following characteristics: the length of at least 150 bp on the genomic sequence, minimum 50% of sequence identity at protein level, alignment coverage of parental gene protein greater than 50% and the loss of at least 2 introns inferred from the alignment. In this case, to reduce the number of false positives, we decided to use minimal length of 75 bp for lost introns, as described previously (Marques et al. 2005). If the locus was classified as a retrogene, the best alignment was used to determine such parameters as precise retrogene location on the genome, parental gene and percent identity (in case of multiple optimal alignments, this alignment was chosen randomly) Identified retrogenes were subsequently screened for the overlap with pseudogenes annotated in Ensembl. Retrogenes showing at least 50% of the overlap with annotated pseudogenes were classified as

'KNOWN_PSEUDOGENE', while the rest were assigned to the 'NOVEL' category.

Detecting retrogenes amongst known, protein-coding genes

To identify annotated, protein-coding genes that originated via retroposition, we first aligned all protein sequences originating from protein-coding genes of a given organism to itself using LAST (Kiełbasa et al. 2011). Proteome sequences were limited to products of genes annotated in Ensembl as ‘protein-coding’. Alignments were subsequently filtered and those between isoforms of the same gene were removed. Next, we selected genes whose coding sequence consisted of one exon in every transcript (if the gene encoded more than one protein, the longest isoform was chosen). For each gene, we searched for the best scored alignment and examined for signs of retroposition. To call the gene a retrogene, the best alignment must have met following criteria: neither tested gene nor its potential parental gene should show reverse transcriptase activity (detected from Ensembl protein families descriptions), coding sequence of the tested gene is at least 150 bp long, alignment identity and coverage of both proteins is greater than 50% and the aligned fragment of potential parental gene (excluding first and last 10 amino acids) must span at least two introns. All found retrogenes were classified as 'KNOWN_PROTEIN_CODING'.

Manual curation

In order to ensure maximum quality of the data, we manually searched for retrogenes originating from parental genes that are likely annotation errors. Parental genes with large numbers of retrogenes (more than 50 for mammals, 5 for other organisms) were screened, and genes originating from transposons and those showing poor or strange profiles of evolutionary conservation were removed.

Expression estimation

In order to estimate expression of both retrogenes and their parental genes, we utilized a two-step approach, similar to the one used in RCPedia (Navarro and Galante 2013). We downloaded short read libraries for various tissues from selected records of NCBI SRA database and filtered them using Fastx Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). We only kept the reads for which at least 95% of bases had the quality score of 20 or more. First, we mapped the reads to the known transcripts from Ensembl 73 databases using Tophat2 (Kim et al. 2013). Unmapped reads were then remapped to the genome using Bowtie2 (Langmead and Salzberg 2012). To ensure unique mapping, we only kept reads which mapped with quality of at least 20 (corresponding to mapping error probability <= 0.01). Expression values were calculated using HTSeq (http://www-huber.embl.de/users/anders/HTSeq) and custom Python scripts. Expression estimates were normalized using the overall number of mapped reads, resulting in RPM (reads per million mapped reads) values.

References

- Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. Ensembl 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D48-55.
- Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011 Mar;21(3):487-93.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013 Apr 25;14(4):R36.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9.
- Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21.
- Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 2005 Nov;3(11):e357.
- Navarro FC, Galante PA. RCPedia: a database of retrocopied genes. Bioinformatics. 2013 May 1;29(9):1235-7.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2.

Kabza M, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polański A, Makałowska I.

Inter-population Differences in Retrogene Loss and Expression in Humans.

PLOS Genetics 2015 Oct 16;11(10):e1005579.

RESEARCH ARTICLE

Inter-population Differences in Retrogene Loss and Expression in Humans

Michał Kabza[1]*, Magdalena Regina Kubiak[1], Agnieszka Danek[2], Wojciech Rosikiewicz[1], Sebastian Deorowicz[2], Andrzej Polański[2], Izabela Makałowska[1]*

[1] Department of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland, [2] Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland

* mkabza@amu.edu.pl (MK); izabel@amu.edu.pl (IM)

OPEN ACCESS

Citation: Kabza M, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polański A, et al. (2015) Inter-population Differences in Retrogene Loss and Expression in Humans. PLoS Genet 11(10): e1005579. doi:10.1371/journal.pgen.1005579

Editor: Takashi Gojobori, National Institute of Genetics, JAPAN

Received: June 3, 2015

Accepted: September 15, 2015

Published: October 16, 2015

Copyright: © 2015 Kabza et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Science Centre (PL), https://www.ncn.gov.pl: grant 2011/01/B/ST6/06868 to AP, grant 2013/11/B/NZ2/02598 to IM, grant 2013/09/N/NZ2/01221 to MK; KNOW RNA Research Centre in Poznań, 01/KNOW2/2014. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Gene retroposition leads to considerable genetic variation between individuals. Recent studies revealed the presence of at least 208 retroduplication variations (RDVs), a class of polymorphisms, in which a retrocopy is present or absent from individual genomes. Most of these RDVs resulted from recent retroduplications. In this study, we used the results of Phase 1 from the 1000 Genomes Project to investigate the variation in loss of ancestral (i.e. shared with other primates) retrocopies among different human populations. In addition, we examined retrocopy expression levels using RNA-Seq data derived from the Ilumina Body-Map project, as well as data from lymphoblastoid cell lines provided by the Geuvadis Consortium. We also developed a new approach to detect novel retrocopies absent from the reference human genome. We experimentally confirmed the existence of the detected retrocopies and determined their presence or absence in the human genomes of 17 different populations. Altogether, we were able to detect 193 RDVs; the majority resulted from retrocopy deletion. Most of these RDVs had not been previously reported. We experimentally confirmed the expression of 11 ancestral retrogenes that underwent deletion in certain individuals. The frequency of their deletion, with the exception of one retrogene, is very low. The expression, conservation and low rate of deletion of the remaining 10 retrocopies may suggest some functionality. Aside from the presence or absence of expressed retrocopies, we also searched for differences in retrocopy expression levels between populations, finding 9 retrogenes that undergo statistically significant differential expression.

Author Summary

Many retrogenes, long considered to be genomic "junk", were recently revealed as vitally important. Retroposition plays an important role in shaping differences between species, populations and individuals. Variation may result either from new retroposition events and/or retrocopy loss, as well as from differences in retrocopy expression. Genome analysis of 1092 individuals from various populations and transcriptomes from 50 individuals, revealed differences between populations in the frequency of transcriptionally active ancient retrocopy loss, as well as differences in retrocopy expression levels. Overall, these results provide new insights into the genomic events of inter-population variation, which has not been evaluated much with respect to gene loss.

Introduction

Retro(pseudo)genes are gene copies that originate through the process of reverse transcription of mRNAs and their subsequent insertion into the genome. This process, called retroposition is usually mediated by enzymes provided by retrotransposable elements and creates intronless copies of existing genes [1]. Due to the lack of promoter sequence of their parental genes, retrocopies were long assumed to be nonfunctional and simply "junk DNA". However, multiple findings challenged that view and a few mechanisms were proposed for retrogene transcription [2]. Functional retrocopies (retrogenes) may have an intact open reading frame and thus be protein-coding, but can also act as long noncoding RNAs [3], sources of short interfering RNAs [4] and microRNA sponges [5]. The loss of parental regulatory sequences and potential replacement with those recruited from the new locus of integration is believed to be the most possible reason that retrogenes are able to undergo neofunctionalization more often than other types of gene copies [6] and take part in the shaping of lineage- and species-specific features [7]. As recently shown, retrogenes can also replace their parental genes [8]. Apart from their functionality, retrocopies are also useful phylogenetic markers, providing insight into such evolutionary processes as sex chromosome origination [9] or changes in retrotransposable element activity and germ line gene expression [2].

In recent years, numerous studies have shown that many retrogenes are vitally important and a number of them are involved in diseases. A good example is the *RHOB* gene (ras homolog gene family member B [MIM: 165370]), a tumor suppressor that belongs to the Rho GTPases family [10]. A mutation in another retrogene, *RNF113A* (ring finger protein 113A [MIM: 300951]), results in trichothiodystrophy [11]. On the other hand, insertion of the *PPIA* cDNA (peptidyl-prolyl isomerase A [MIM: 123840]) into the *TRIM5* gene (tripartite motif-containing protein 5 [MIM: 608487]) confers resistance to HIV in owl monkeys [12].

Retroposition gives rise to considerable genetic variation between individuals. Recent developments in sequencing technology allow researchers to move beyond the analysis of individual genomes from model organisms to the study of retrocopies within a population. Large-scale sequencing projects, such as the 1000 Genomes Project [13] enable the exploration of differences in copy-number variation within human populations. Three recent studies [14–16] on the retrocopy repertoire in human populations revealed a total of 208 polymorphic retrocopies [17] called retroduplication variations (RDVs). In addition, two of them [14, 15] aided in reconstructing the phylogenetic tree of human populations. Thus, proving the value of RDV polymorphisms as genomic markers for population history.

Despite these advances, many questions remained unanswered. Current methods of discovering novel retropositions, (i.e. not annotated on the reference genome), utilize paired-end reads and require at least one read from the pair to map to the parental gene of the retrocopy. As a result, this approach allows only for the detection of retrocopies that originated relatively recently in evolutionary history and show little sequence divergence compared to their parental genes. Evolutionary events that happened earlier in the human lineage and resulted in currently observed RDVs have been less explored. Moreover, we know surprisingly little about the functional aspects of different kinds of retrogene polymorphisms. Single nucleotide polymorphisms located in retrogene promoters and regulatory sequences might result in different levels of retrogene expression between individuals and populations. Presence or absence of a functional retrocopy might have far reaching biological implications. Due to lack of reliable experimental data, these issues have thus far been ignored; yet, this will most certainly change in the near future.

In this study, we focused on aspects not covered by aforementioned publications. We used the results of Phase 1 from the 1000 Genomes Project to investigate variation in loss of ancient (i.e. shared with other primates) retrocopies among human populations. In addition, we utilized RNA-Seq data from the Ilumina BodyMap project as well as data from the lymphoblastoid cell lines of 50 individuals provided by the Geuvadis Consortium [18] to examine variation in the expression level of retrocopies.

Results

Identification of RDVs from known retrocopies in human populations

To analyze retrocopy number variation resulting from retrocopy loss in human populations we used 4,927 human retrocopies from RetrogeneDB [19] and the genomic data from Phase 1 of the 1000 Genomes Project [13]. The genomic data consisted of whole genome sequencing of 1,092 individuals from 14 populations. By mapping human retrocopies to genomic variants detected during Phase 1 of the 1000 Genomes Project, we identified 214 indels that affected 190 retrocopies. To be more explicit, 190 retrocopies annotated in the human reference genome were at least partially missing (i.e. at least 100 bp of the retrocopy sequence was deleted) in some of analyzed genomes. Next, we calculated allele frequencies of detected indels in 14 available populations (S1 Table and RVD Maps online material). Out of all identified indels involving retrocopies, 67 were population specific and 11 were observed in all investigated populations. Forty-eight indels were relatively widespread and affected 6–8 populations, and 88 occurred in only 2–5 populations.

Indels that were detected only rarely in populations have, with some exceptions, low frequencies and are observed in 0.5%–2% of alleles. This may suggest that these indels represent either relatively new deletions or deleterious deletions, and thus, were subjected to negative selective pressure. The highest rate of absence was observed in the case of retrogene retro_-hsap_1441 (Fig 1). This retrogene is present in about 20% of alleles in Asian populations, in about 50% of the alleles in European and American populations, and in about 70% in populations with African ancestry. The most likely explanation for this phenomenon is the emergence of a new retroposition in Africa, which spread to other continents, or alternatively, a deletion that originated in Asia. However, based on available data we cannot distinguish between these two scenarios.

Detection of the loss of ancestral, expressed retrocopies

Due to the methodology applied, previous studies [14–16], indicated that the majority of retrocopy polymorphisms resulted from novel retropositions. In this study, we focused on the identification of polymorphisms that resulted from a retrocopy deletion. We can assume that the lack of a retrogene observed at low frequency and only in selected populations represents a deletion. However, when absence of the retrogene is common across populations and detected at relatively high frequency, the polymorphism may be a result of either deletion or new retroposition. Therefore, to identify retrocopy deletion events, we performed comparative analysis across fourteen Eutherian species in order to identify retrocopy orthologs that originated prior to human speciation.

In the case of retroposition, a reciprocal sequence similarity search is not sufficient to establish orthology, since many genes undergo independent retroduplication in different species

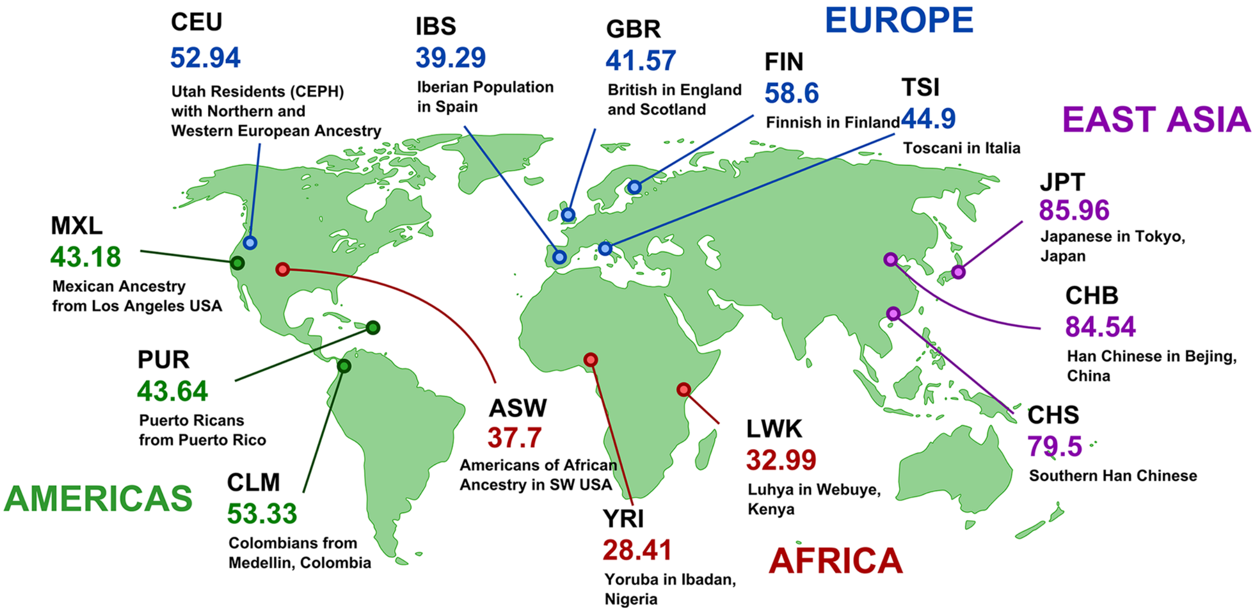

EUROPE
CEU
52.94
Utah Residents (CEPH) with Northern and Western European Ancestry
IBS
39.29
Iberian Population in Spain
GBR
41.57
British in England and Scotland
FIN
58.6
Finnish in Finland
TSI
44.9
Toscani in Italia
EAST ASIA
JPT
85.96
Japanese in Tokyo, Japan
CHB
84.54
Han Chinese in Bejing, China
CHS
79.5
Southern Han Chinese
MXL
43.18
Mexican Ancestry from Los Angeles USA
PUR
43.64
Puerto Ricans from Puerto Rico
AMERICAS
CLM
53.33
Colombians from Medellin, Colombia
ASW
37.7
Americans of African Ancestry in SW USA
YRI
28.41
Yoruba in Ibadan, Nigeria
LWK
32.99
Luhya in Webuye, Kenya
AFRICA

Fig 1. Frequencies of retro_hsap_1441 absence in different human populations. Maps for all retrocopies are available at http://rhesus.amu.edu.pl/RetrogeneMaps/.

doi:10.1371/journal.pgen.1005579.g001

[20]. Hence, in order to pinpoint orthologous retrocopies we used information concerning the location of retrocopies identified by us in mammalian genomes [19] and mapped them on the genomic sequences alignments using Ensembl release 73. First, we checked for orthologous genes in major mammalian lineages. We classified a retrocopy as common for a lineage if it was observed in any two species from this lineage. The requirement of retrocopy presence in a minimum of two but not necessarily all genomes is based on the assumption that the probability of independent retroposition of the same gene at the same location is close to zero. Using this approach, we identified as many as 1,282 retroposition events that took place in the Hominidae ancestor genome. As many as 510 retropositions are common to Catarrhini, and 360 to primates. In comparison, we did not detect any retrocopies common to Glires and only 84 shared by mouse and rat (Fig 2). This analysis confirms a burst of retroposition in Primates and shows that it was especially intensive in Hominidae [7].

While considering only human retrocopies, we found that 1,954 were present in at least one other genome. From these, 1,174 originated in the common ancestor of Hominidae and 57 are very ancient, as they are present in at least one genome of analyzed Eutheria other than Eurchontoglires (Fig 2). From the set of polymorphic retrocopies, 68 were among those we identified as ancestral. Therefore, with high confidence we can say that the polymorphism of these 68 retrocopies resulted from deletion and not from a new retroposition event.

Most retrocopies are non functional and therefore, their loss is in most cases neutral. However, many retrocopies are expressed and this, together with conservation may indicate some functionality. To assess retrocopy expression, we used RNA-Seq samples from ten individuals selected from five different populations (CEU, GBR, FIN, TSI, YRI) derived from the Geuvadis RNA sequencing project [18], as well as data for 16 human tissues from the Illumina Bodymap 2.0 project. To distinguish between parental gene and the retrocopy, only uniquely mapped reads were considered. As expressed we considered a retrocopy with a normalized expression value of at least 1 RPM (reads per million mapped reads) in at least one sample, either one library from the Body Map project or one individual from the Geuvadis RNA sequencing

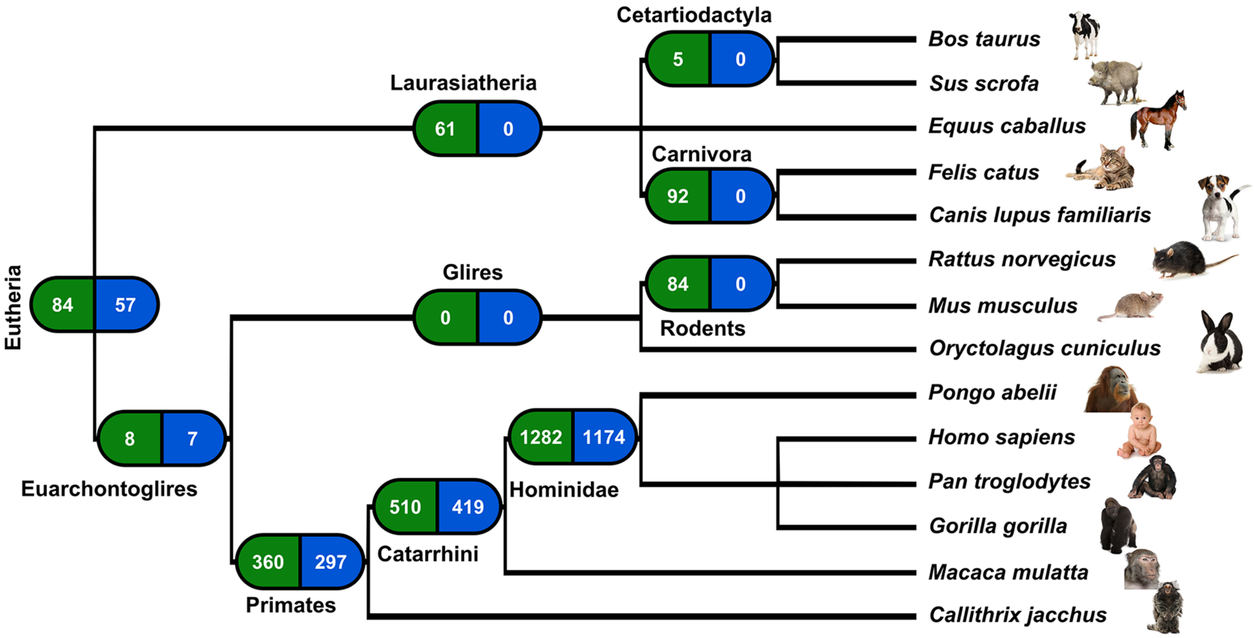


Fig 2. Retroposition events in Eutheria. Left boxes show number of retrocopies originated in a given lineage ancestors' genome; right boxes show number of retrocopies detected in the human genome.

doi:10.1371/journal.pgen.1005579.g002

project. These criteria were met by 588 retrocopies, and 11 of them were also ancestral and absent in some individuals (Fig 3).

Characterisation of ancestral, expressed retrocopies undergoing deletion

We calculated Fisher's exact test to assess the relationship between expression, ancestrality and retrogene loss. The obtained p-values were all higher than 0.05 and therefore, neither conserved nor expressed retrocopies are less likely to be polymorphic. Most ancestral and expressed retrocopies are deleted in only a small fraction (0.5–2%) of alleles from certain populations, suggesting that their deletions are either relatively new or slightly deleterious; thus, were subjected to negative selective pressure (Table 1).

To confirm and evaluate the expression of 11 transcriptionally active ancestral retrocopies undergoing deletion, we decided to use regular PCR followed by quantitative PCR (real-time PCR). In the case of two retrogenes (retro_hsap_4873 and retro_hsap_2011) we were not able to identify primers suitable for qPCR; therefore, we selected primers for standard PCR only. We ran standard PCR in 16 cDNA libraries from various organs, and sequenced amplification products to confirm primer specificity. Sequencing was necessary, as primers specific for a given retrogene could still amplify RNA expressed from a parental gene. Indeed, in the case of one retrocopy (retro_hsap_88) products were not always specific (S1 Fig). In addition, primers designed for two other retrocopies (retro_hsap_2905 and retro_hsap_4123) formed dimers. Finally, primers for retro_hsap_477 functioned only at a much lower temperature than is required for qPCR. For these four retrocopies, as in the case of retro_hsap_4873 and retro_hsap_2011, we were not able to identify alternative pairs of primers meeting qPCR requirements. Nevertheless, we managed to confirm the expression of all 11 retrocopies by regular PCR (we found working primers for longer products). Sequences of all primers are provided in S2 Table.

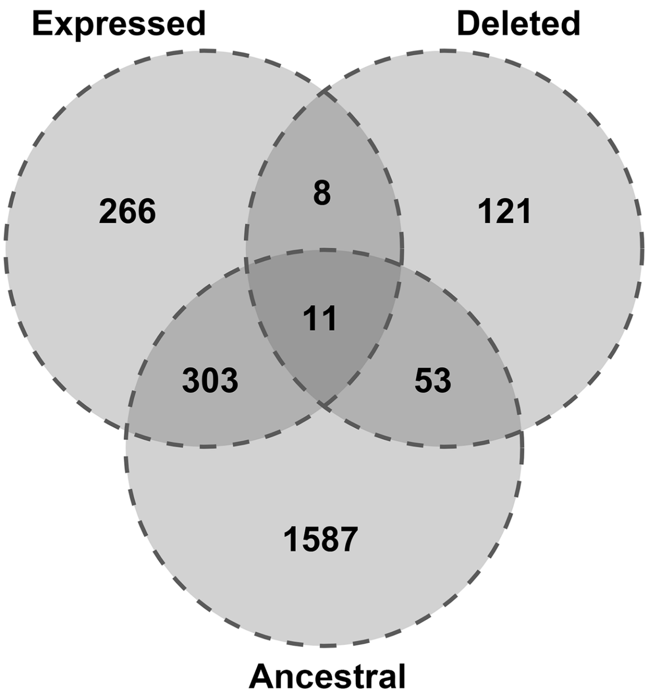


Fig 3. Expression and conservation of retrocopies deleted relative to the reference genome.

doi:10.1371/journal.pgen.1005579.g003

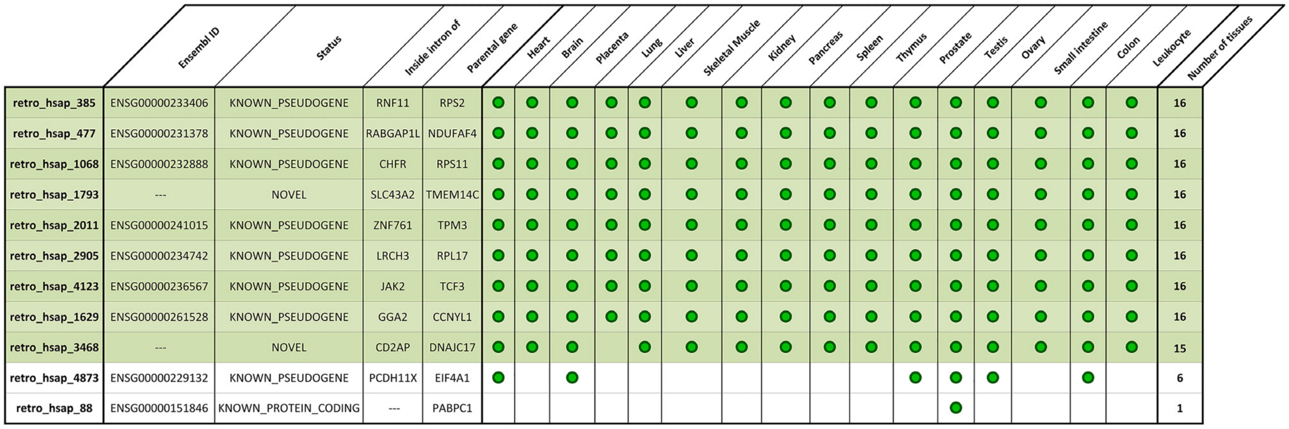
The analysis showed that eight retrogenes are expressed in all sixteen analyzed libraries and one (retro_hsap_3468) is expressed in all but lung. In contrast, retro_hsap_88 is expressed only in testis. In the case of retro_hsap_4873, we could with high confidence, confirm expression in six organs (Fig 4 and S1 Fig).

Due to the above-mentioned reasons, qPCR could be performed on only five retrocopies. It revealed low-level expression in nearly all organs. High expression is observed only for retro_hsap_1793 in placenta and some retrocopies show moderate expression in pancreas, spleen,

Table 1. Frequencies of ancestral and expressed retrogene loss in human populations.

Retrocopy	Indel Length	Population													
		African			Ad Mixed American			European					East Asian		
		ASW	YRI	LWK	MXL	PUR	CLM	CEU	IBS	GBR	FIN	TSI	JPT	CHB	CHS
retro_hsap_88	1272	0	0	3.09	0.76	0	1.67	0.59	0	0	1.08	0.51	1.69	1.03	0
retro_hsap_385	6257	0	0	0	0	1.82	0.83	0	0	0	0	0	0	0	0
retro_hsap_477	16238	0	0	0	0	0	0	0	0	0	0	0	0	1.03	0.5
retro_hsap_1068	46133	0.82	0	0	0	0	0	0	0	0	0	0.51	0	0	0
retro_hsap_1629	5062	0	0	0	0	0	0	0	0	0.56	0	0	0	0	0
retro_hsap_1793	4082	0.82	0.57	1.03	0	1.82	0	0	0	0	0	0	0	0	0
retro_hsap_2011	36761	0	0	0	0	0	0	0	0	0	1.08	0	0	0	0
retro_hsap_2905	6836	0	0	0	0	0	0	0	0	0	0	0.51	0	0	0
retro_hsap_3468	2181	0.82	0	0	3.03	3.64	5	5.29	3.57	11.24	3.76	5.61	0	0	0
retro_hsap_4123	9011	0.82	0.57	1.03	0	0	0	0	0	0	0	0	0	0	0
retro_hsap_4873	1260	15.31	11.28	15.07	0.99	1.22	0	0	0	0	0	0	0	0	0

doi:10.1371/journal.pgen.1005579.t001



	Ensembl ID	Status	Inside intron of	Parental gene	Heart	Brain	Placenta	Lung	Liver	Skeletal Muscle	Kidney	Pancreas	Spleen	Thymus	Prostate	Testis	Ovary	Small intestine	Colon	Leukocyte	Number of tissues
retro_hsap_385	ENSG00000233406	KNOWN_PSEUDOGENE	RNF11	RPS2	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_477	ENSG00000231378	KNOWN_PSEUDOGENE	RABGAP1L	NDUFAF4	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_1068	ENSG00000232888	KNOWN_PSEUDOGENE	CHFR	RPS11	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_1793	---	NOVEL	SLC43A2	TMEM14C	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_2011	ENSG00000241015	KNOWN_PSEUDOGENE	ZNF761	TPM3	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_2905	ENSG00000234742	KNOWN_PSEUDOGENE	LRCH3	RPL17	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_4123	ENSG00000236567	KNOWN_PSEUDOGENE	JAK2	TCF3	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_1629	ENSG00000261528	KNOWN_PSEUDOGENE	GGA2	CCNYL1	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	16
retro_hsap_3468	---	NOVEL	CD2AP	DNAJC17	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	15
retro_hsap_4873	ENSG00000229132	KNOWN_PSEUDOGENE	PCDH11X	EIF4A1	●		●								●	●	●		●		6
retro_hsap_88	ENSG00000151846	KNOWN_PROTEIN_CODING	---	PABPC1												●					1

Fig 4. Expression pattern of deleted expressed retrocopies based on standard PCR.

doi:10.1371/journal.pgen.1005579.g004

testis, ovary, and leukocyte (Fig 5). In addition, although retro_hsap_3468 exhibits expression in all but one organ when amplified by standard PCR, qPCR did not yield significant products in five libraries (Fig 5).

Based on our earlier retrocopy analysis [19], we determined that only one out of 11 lost expressed retrocopies (retro_hsap_88) is annotated as a single exon gene (Ensembl: ENSG00000151846) coding for the cytoplasmic 3 poly(A) binding protein (*PABPC3* [MIM: 604680]). Eight retrogenes are annotated in the Ensembl database as processed pseudogenes, and two (retro_hsap_3468 and retro_hsap_1793) are novel, i.e. identified by us [19] and are not annotated in the Ensembl or other databases (Fig 4). Three retrocopies originated from

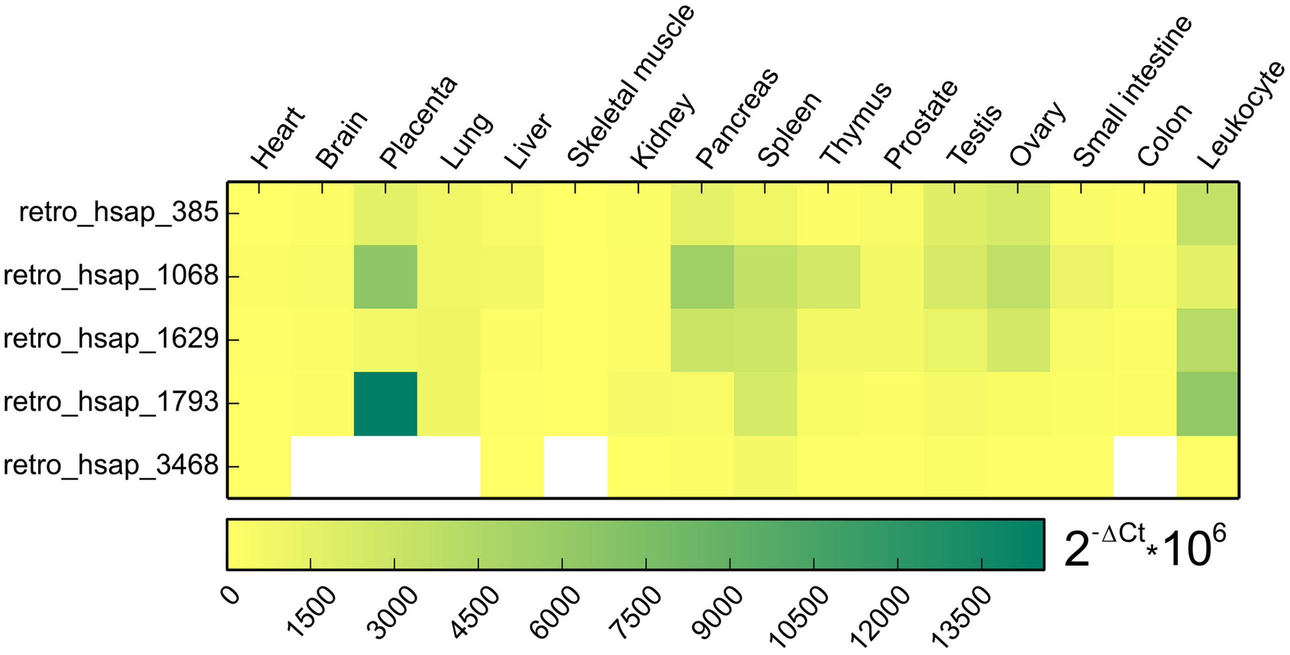


Fig 5. Heat map representing expression pattern of five retrogenes based on qPCR. White color indicates tissues with no significant expression level (Ct > 32).

doi:10.1371/journal.pgen.1005579.g005

ribosomal protein coding genes, which is not surprising as these genes yielded an exceptionally high number of retrocopies [21].

The function of these 11 retrogenes is not known, with the exception of retro_hsap_88, which codes for a known protein. Only three of the remaining 10 retrocopies have conserved open reading frames (ORF), i.e. there are no stop codons or frame shifts over the entire retrocopy-parental gene alignment. Interestingly, these three retrocopies are those originated from genes encoding for ribosomal proteins.

With one exception (retro_hsap_88), all analyzed retrocopies are located in the introns of other genes; which is quite typical for retrocopies. Interestingly, retrogene located in the first intron of the zinc finger protein 761 gene (*ZNF761*), retro_hsap_2011 (annotated in the Gene database as *TPM3P9*), exapted the first exon of its host gene, according to annotations. This modification allowed the retrocopy to acquire a regulatory machinery, and consequently, the ability to be expressed, which we confirmed by standard PCR. In addition, part of the *ZNF761* intron was incorporated as part of a retrogene exon (Fig 6). Another possible scenario is that a retrogene was incorporated into an existing two-exon splice variant. In any event, this retrocopy represents yet another example of the formation of new genes via retroposition followed by structural evolution [6, 22].

Differences between populations in retrogenes expression

Evolutionary changes involve more than just SNPs or mutations in the coding sequence. The differences between species, populations or individuals may also result from mutations in the regulatory regions, and can result in varying expression levels of an affected gene. Therefore, in addition to the analysis of retrocopy absence, we also analyzed retrocopy expression level differences between populations. To perform this analysis we utilized the DESeq2 differential expression package [23] and retrocopy expression estimations for lymphoblastoid cell lines from 50 individuals within five populations (CEU, GBR, FIN, TSI, YRI); 10 individuals per population. The analysis was performed on all 4,927 retrocopies downloaded from RetrogeneDB. In general, the differences detected were not significant. However, in nine retrogenes, we observe notable fold changes (Table 2). As expected, most differences occurred between European and African populations, although the British population was revealed to be somehow distinct from populations originating from continental Europe. The most frequently observed difference is between British and Yoruba populations. Three retrocopies have

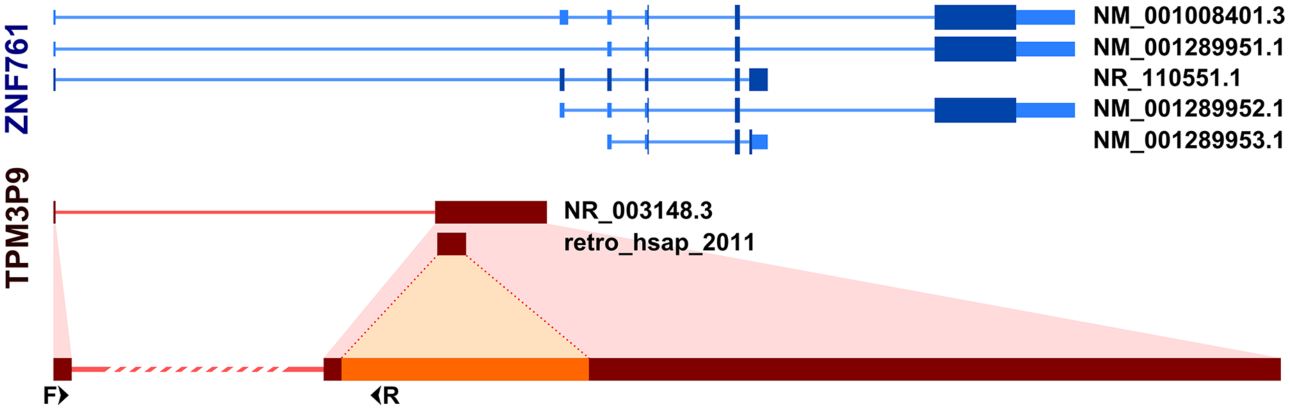


Fig 6. Incorporation of host gene exon by retrogene retro_hsap_2011. Splice variants of *ZNF761* gene (upper part) and *TPM3P9* retrogene (lower part). F and R refer to forward and reverse primer binding sites, respectively.

doi:10.1371/journal.pgen.1005579.g006

Table 2. Retrocopies undergoing differential expression between analyzed populations.

Retrocopy	Parental gene	Conserved ORF	Other tissues	Reference population	Tested population	Fold change
retro_hsap_108	RHOA	Yes	All 16 tissues	GBR	FIN	0.43
retro_hsap_1259	RCN1	No	Heart, lung, ovary, prostate	GBR	YRI	0.47
retro_hsap_1692	TMEM254	Yes	-	GBR	YRI	2.10
retro_hsap_1750	TAF5L	No	Lymph node	CEU	GBR	2.20
retro_hsap_2310	RPL23A	No	-	FIN	YRI	0.49
retro_hsap_2684	PDCL3	No	-	GBR	YRI	0.41
retro_hsap_3129	MYL12B	Yes	-	GBR	YRI	0.44
retro_hsap_3129	MYL12B	Yes	-	FIN	YRI	0.48
retro_hsap_3265	RPL10	Yes	Lung, skeletal muscle	TSI	YRI	2.29
retro_hsap_4127	RPL23A	No	-	GBR	YRI	2.58

doi:10.1371/journal.pgen.1005579.t002

significantly higher expression levels in individuals from the British population. One of them, retro_hsap_3129, is also exhibits a higher expression in the Finnish population when compared with Yoruba. Two retrogenes demonstrate a higher level of expression in Youruba than in the British population. The British population has one retrocopy expressed at a higher level than the Finnish population, and one expressed at lower levels compared to Utah residents of European ancestry. The remaining differences between populations include an additional retrogene in which the expression is significantly higher in Finnish compared to Yoruba, and one retrocopy with lower expression in individuals from Toscani when compared to individuals from Yoruba.

The detection of retrocopies absent from the reference genome

We developed a new methodology in order to identify novel (i.e. not annotated in the reference genome) retrocopies deleted in some individuals. We performed de novo assembly of unmapped reads from 500 individual genomes included in Phase 1 of The 1000 Genomes Project. This yielded 6,911 contigs, 2,901 of which were singletons detected in only one individual. We excluded sequences that showed similarity to human genome patches, alternative assemblies, as well as genomic contaminations. 1,233 (17.8%) sequences were filtered out as bacterial and viral contaminations. Strong similarity to GRCh37.p10 and HuRef human genome assemblies was observed in 861 (12.4%) and 1,767 (25.6%) respectively. Finally, we applied the retrocopy identification pipeline used previously in RetrogeneDB [19] to detect novel retrocopies. As a result, five novel retrocopies were identified. Four of the identified retrocopies were found in only a few individuals (1 to 10). However, retrocopy rdn1 was relatively frequent and was identified in 40 individuals (Table 3).

All contigs containing retrocopies show strong (at least 99% identity) similarity to some human clone sequences in the BLAST nr/nt database, suggesting that they were sequenced, but were not included in any major human genome assembly. Comparison of detected retrocopies with the results of previous studies [14–16] reveals that these retrocopies are in fact, novel.

Detected retrocopies show significant divergence between their sequences and sequences of their parental genes, indicating old retroposition events. We were able to find the orthologs of discovered retrocopies in the chimpanzee and/or gorilla genomes, confirming that they are indeed, ancestral retroduplications that were subsequently lost during human evolution. This analysis also allowed us to determine exact locations and sizes of deletions for retrocopies rdn1, rdn2 and rdn3 (Table 3 and Fig 7).

Table 3. Retrocopies absent in the reference genome.

Retrocopy	rdn1	rdn2	rdn3	rdn4	rdn5
Length (bp)	373	459	331	338	460
Parental gene	C9orf85	RPL21	PTGES3	RARRES2	RPL6
Deletion site in reference genome	chr3: 66287175–66287285	chr16: 90107060–90107061	chr9: 104121215–104121216	—	—
Identity (protein level)	66.67%	72.02%	52.21%	61.19%	82.47%
Number of individuals	40	9	10	1	2

doi:10.1371/journal.pgen.1005579.t003

To confirm the identified retrocopies and their deletions in some individuals, we performed PCR on 17 human genomes supplied by the 1000 Genome Project (Coriell Cell Repositories). Individuals were selected based on previous bioinformatic analyses, and each originated from a different population. Using primers from identified retrocopy sequences, we confirmed the existence of four retrocopies (rdn1–rdn4). Retrocopy rdn5 contains a large number of repetitive sequences. Therefore, we could not obtain a product specific enough to confirm the presence of this particular retrocopy: it was excluded from further studies. However, we were able to prove the deletions of retrocopies rdn1 to 3 in some individuals. Apparently, retrocopy rdn4 was present in all investigated genomes. In addition, utilizing primers designed from the flanking regions of the deletion, we confirmed the size of the deleted regions (Table 3), as well as the homo- and heterozygosity of the studied individuals (Fig 8).

None of discovered retrocopies has a conserved open reading frame, which may suggest that they are non-functional. Using standard PCR in a pooled cDNA library, we searched for the expression of these retrocopies. We obtained an amplification product in only one retrocopy, rdn2. Its expression is also confirmed by three EST sequences deposited in the dbEST (GenBank: DA381977, AV702346, AV629627).

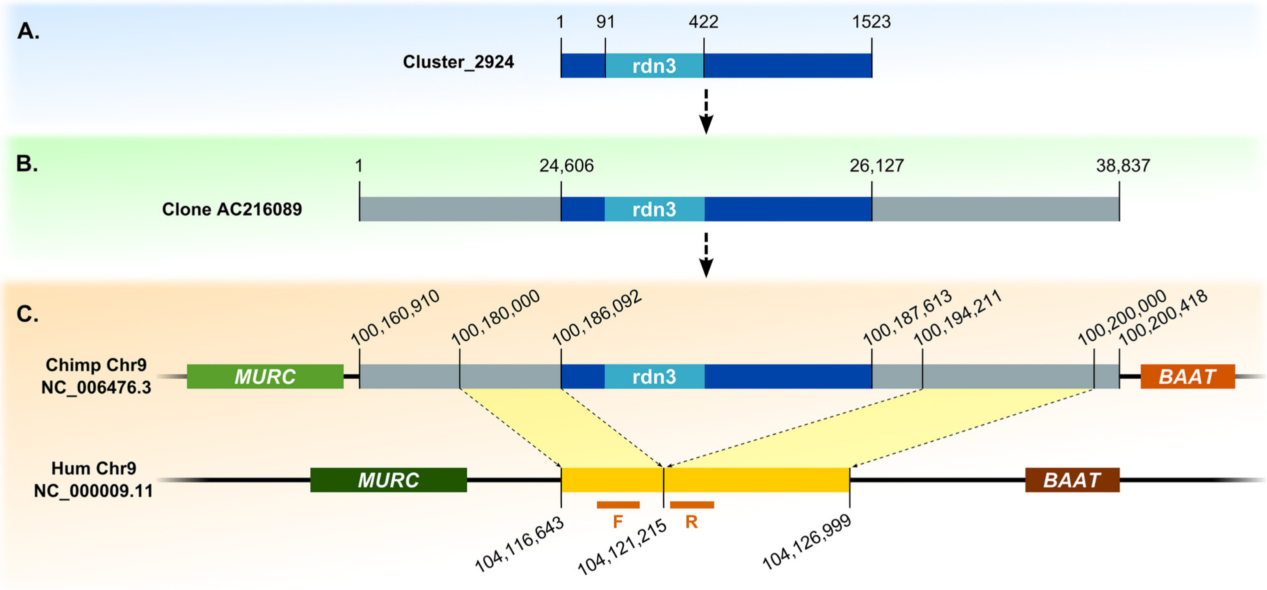


Fig 7. Detection of novel retrocopy deletion sites, example of retrocopy rdn3. (A) Contig and BAC containing its sequence. (B) Alignment with chimpanzee genome. (C) Identification of indel site. F and R refer to forward and reverse primers designed for examined indel site in human genome, respectively.

doi:10.1371/journal.pgen.1005579.g007

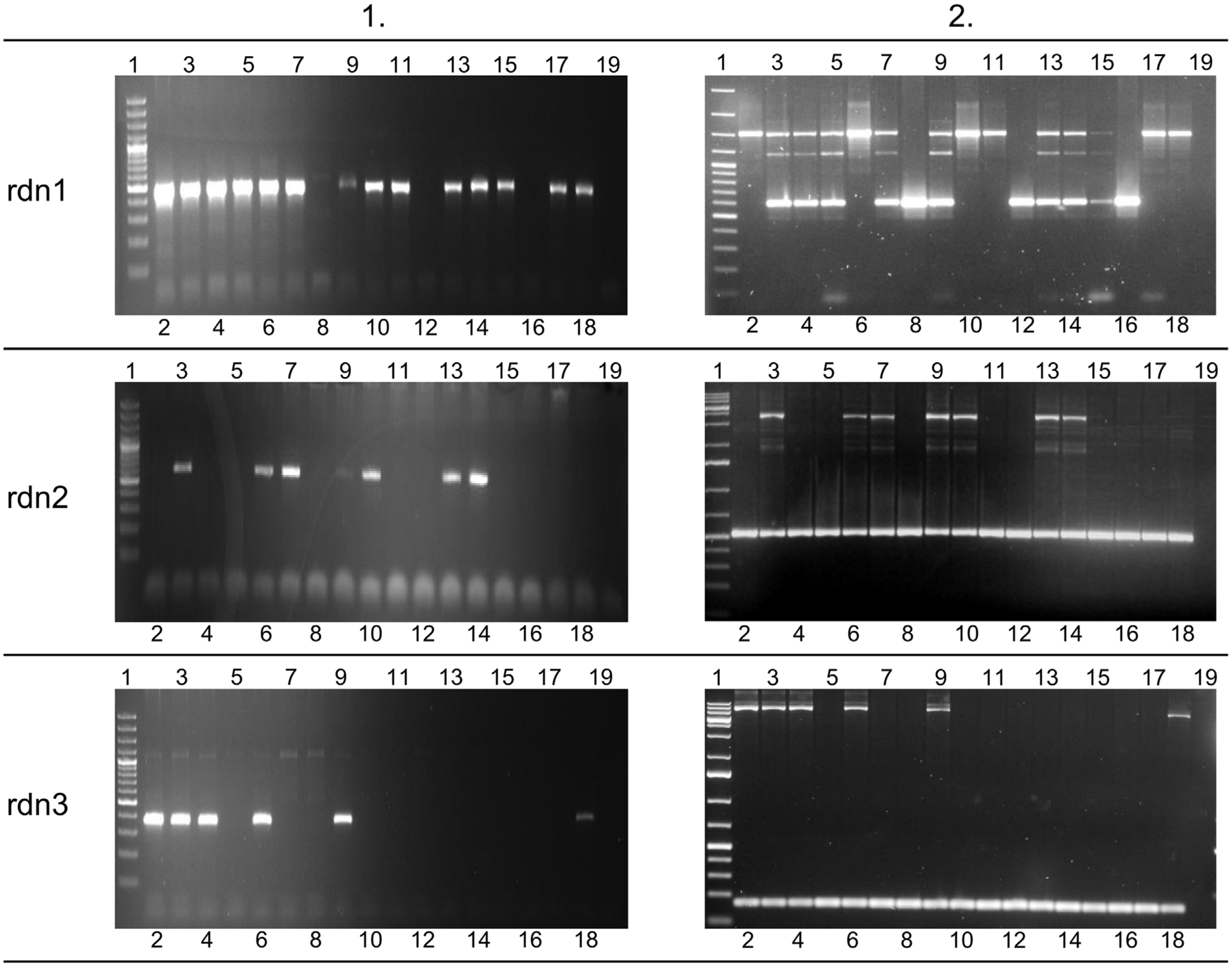


Fig 8. Agarose gel results for novel retrogenes. Left side of figure (column 1) represents presence or absence of novel retrogenes in 17 examined genomes. Agarose gels on the right side of figure (column 2) show PCR products corresponding to the region of deletion. This indicates homo- or heterozygotic character of 17 studied individuals. Lane 1 –GeneRuler 100 bp Plus DNA Ladder or GeneRuler 1 kb DNA Ladder (Thermo Scientific); lanes 2–18 –genomic DNA templates (see S1 Table for details), lane 19 –negative control (water instead of DNA).

doi:10.1371/journal.pgen.1005579.g008

Discussion

Retroduplication variations (RDVs) are a class of genomic sequence polymorphisms associated with the presence or absence of retrocopies in individual genomes. Until recently, our knowledge of the prevalence and significance of this phenomenon was limited. The advent of large-scale sequencing projects, such as The 1000 Genomes Project, allowed researchers to detect such polymorphisms on a massive scale, including both known and novel (not present in the reference genome) retrocopies. Particularly the latter class has drawn much attention, as their discovery is much more challenging than determining the presence or absence of an annotated retrocopy. Previous studies [14–16] focused on relatively recent retroduplication events, which is hardly surprising given that they are very informative on the history of human populations. Two of these studies were indeed able to reconstruct population phylogenetic trees from detected retroduplication variations. Identifying these novel retrocopies required mapping short reads to the genomic sequences spanning splice sites of their parental genes. However, not every novel human retroduplication polymorphism can be discovered in that way. For example, retrocopies that originated very early in the human lineage might have diverged too much from their parental genes to be mapped to their sequences. Similarly, some ancestral retrocopies may have been lost from some of the ancient genomes and are thus, very infrequent in contemporary human populations. As a result, they were not included in the human reference genome. Finally, some retrocopies may not be included in the reference genome, simply due to genome misassembly.

Using our two approaches, we were able to detect 193 RDVs (190 RDVs affecting annotated retrogenes and 3 RDVs of retrocopies absent from the reference genome). This is slightly less than the sum of all previous studies, but significantly more than in any of these studies individually (176 [15], 58 [14], and 91 [16]). However, due to a different approach, our results complement rather than overlap previous studies. While previous studies mainly focused on RDVs from recent retroposition events, our dataset revealed RDVs resulting from retrocopy deletion. We found orthologs for only 56 retrocopies missing in some human genomes. Nevertheless, we are confident that indels resulting from a deletion dominate in this set, since frequency of retrocopy absence is low for the majority of them; absence in more than 50% of individuals was observed in only a few of the retrogenes. For new retropositions, we would expect absence to be close to 100% of alleles in at least some populations.

In this study we developed an approach to detect and characterize hitherto undiscovered retroduplication variations in the human genome. To accomplish this, we de novo assembled the unmapped reads from 500 individuals whose genomes were sequenced in the 1000 Genomes Project. Contigs from different individuals were subsequently combined and assembled into novel genomic regions. Finally, genomic regions were filtered to exclude cross-species contaminations and sequences from alternative human genome assemblies (HuRef) and reference genome patches (GRCh37.p10). These filtered sequences were searched for retrocopies using the RetrogeneDB pipeline [19]. Using this strategy, we were able to discover 5 novel retrocopies, none of which were previously detected [14–16]. We experimentally confirmed the existence of the detected retrocopies and checked for their presence and absence in the human genomes from 17 different populations. The profile of retrocopy presence varied greatly between retrocopies, with some of them being ubiquitous and some detected in only few populations. Detected retrocopies showed significant divergence between their sequences and sequences of their parental genes, indicating an old retroposition event. For all of these retrocopies, we were able to find the orthologs in the chimpanzee or gorilla genome, confirming that they are, in fact ancestral retroduplications that were subsequently lost during human evolution.

The number of discovered novel retrocopies (i.e. not annotated on the reference genome) is much lower than in previous studies [14–16]. This is mostly due to the fact that the current human reference genome is of very high quality. As a result, most polymorphic retroduplications with sequences clearly distinguishable from their parental genes are simply present in the reference genome. Our results suggest that we have reached the limit where the reference genome contains almost the entire human retrogene repertoire, with the exception of recent retroduplications, for which obtaining the exact sequence from the genome sequencing data is not trivial, as they are extremely similar to their parental genes. This is certainly not the case for most sequenced genomes, and our approach can be used to complement the retrogene set in many low-coverage, draft euakryotic genomes, for which resequencing data is available. Our pipeline can also be potentially applied to detect retrocopies specific to genomes of ancient, extinct organisms, such as neanderthal [24] or woolly mammoth [25].

The role of retrocopy polymorphisms as markers for human population history is clearly established, but our findings suggest that they can also provide great insight into ongoing evolutionary processes. The availability of population RNA-Seq data from the Geuvadis

Consortium [18] allowed us not only to study changes in retrocopy sequences and repertoire, but also consider their potential functional implications. We particularly focused on expressed retrogenes that are absent in the genomes of some individuals. RDVs seem to affect both expressed and non-expressed retrocopies indiscriminately, but their allele frequency for most deletions of expressed retrocopies is relatively low ($< 2\%$), suggesting that they are new or slightly deleterious.

We experimentally confirmed the expression of 11 ancestral transcriptionally active retrogenes undergoing deletions. The frequency of their deletion is very low with the exception of retrogene retro_hsap_4873, which is absent in over 10% of alleles in African populations. The expression, ancient origin and low rate of deletion of the remaining 10 retrocopies may suggest some functionality, but further studies would be required to establish their function and the possible consequences of their deletion.

Numerous studies revealed a tendency of retrogenes to be expressed in the testis [9, 26]. It has been hypothesized that this specific transcription may be the result of the hypertranscription state observed in meiotic and postmeiotic spermatogenic cells [27]. Alternatively, retrocopies could be preferentially inserted into actively transcribed, and therefore open chromatin [28]. Since the retroposition occurs in the germ line, retrocopies may be primarily inserted into, or near to genes expressed in the germ line, which may enable or enhance their expression in testis. Another hypothesis links this testis-specific expression with an escape from the male meiotic sex chromosome inactivation [29]. To verify the tendency of retrocopies to be expressed in testis, we performed expression analysis on 16 cDNA libraries from various human organs. Our findings confirmed testis-specific expression of only one retrogene. This supports results from our previous studies, showing that many retrogenes have broad expression patterns [8]. However, the set of analyzed retrogenes is relatively small and may not be fully representative.

Apart from the presence or absence of expressed retrocopies, we also searched for more subtle differences in retrocopy expression levels between populations. Overall, we detected 9 retrogenes that undergo statistically significant differential expression. Similar to inter-population differences in RDVs frequencies, retrogene expression differences tend to coincide with geographical locations, with most differences occurring between African and European populations. Observed differences most likely arise from different frequencies of short sequence variants (SNPs and indels) in the regulatory sequences of the retrogenes studied, but we cannot exclude more complicated scenarios, such as variation in DNA methylation. At this point, we still know very little concerning the functional implications of discovered expression differences, and this will require further study.

Materials and Methods

Detecting RDVs of known retrocopies

The set of human retrocopies generated by us from the RetrogeneDB database [19] was used for the identification of RDVs from known retrocopies. Sequence variants detected for 1,092 individuals during Phase 1 of the 1000 Genomes Project [13] were downloaded and searched for long deletions (structural variants) that overlapped retrocopy loci using BEDtools [30]. Only deletions that resulted in the loss of at least 100 bp of the retrocopy sequence were reserved for further analysis. For each deletion its frequency was calculated in each of the 14 analyzed populations (ASW, CEU, CHB, CHS, CLM, FIN, GBR, IBS, JPT, LWK, MXL, PUR, TSI and YRI).

Retrocopy conservation analysis

Orthologs of human retrocopies in 13 other eutherian species were identified using Amniota vertebrates PECAN [31] whole genome alignment from Ensembl release 73 [32], requiring a reciprocal overlap of at least 50% between retrocopy sequences from different organisms in the alignment. To be more specific, we used bx-python package to extract genome alignment blocks corresponding to retrocopies annotated in RetrogeneDB. BEDtools [30] software was then used to find sets of overlapping blocks that constituted orthologous groups of retrocopies. We considered a human retrocopy to be ancestral if it had an ortholog in at least one other primate species (chimpanzee, gorilla, orangutan, macaque or marmoset).

De novo assembly of novel retrocopy sequences and indel site identification

To search for novel retrocopies in the human genome, short reads that failed to align to the human reference genome for 500 individuals representing all populations included in Phase 1 of The 1000 Genomes Project were downloaded. Short reads from each individual were separately assembled using SOAPdenovo [33] and only contigs with a minimal length of 500 bp were retained. Then, clustering on all the assembled contigs using CD-Hit-EST [34] was performed, requiring at least 95% identity over at least 70% of the sequence. For each cluster, consensus sequence using CAP3 [35] was calculated, obtaining the set of novel genomic regions. Sequences that showed similarity to human genome patches and alternative assemblies (GRCh37.p10, HuRef), as well as genomic contaminations from viruses, bacteria and parasitic organisms were excluded from further analysis. Finally, we applied the retrocopy identification pipeline from RetrogeneDB [19] to detect and characterize retrocopies in assembled genomic regions. The pipeline takes a very stringent approach to identifying gene copies that originated via retroposition. The retrocopy identification process is based on the alignment of the whole proteome to the genome, using LAST software [36]. Observed alignments must meet numerous criteria to be considered retrocopies, including a length of at least 150 bp on the genomic sequence, alignment identity and protein coverage greater than 50%, and the loss of at least 2 introns inferred from the alignment.

To extend contig sequences, unassembled BAC clones and fosmids were searched using BLAST [37]. Identified human BACs and fosmids were then aligned to chimpanzee and/or gorilla genomes to identify homologous sequences. Deletion sites in the reference genome were detected based on comparison of sequence and gene locations in human, chimpanzee and gorilla genomes (Fig 7).

Novel retrogenes and indels confirmation

Novel retrogenes and indels were confirmed experimentally by PCR in the genomes of 17 individuals from various origins (S3 Table) from Coriell Cell Repositories. PCR was carried out using primers designed from the retrocopy sequence and sequences flanking the RDV.

Expression pattern analysis

Expression of retrogenes was analyzed in Multiple Tissue cDNA Panels from Clontech—Human MTC Panel I and II (catalog no. 636742 and 636743 respectively). This set of tissues contained the following: heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, small intestine w/o mucosal lining, colon and peripheral leukocyte cDNA.

Standard PCR

All PCR primers used in this study were designed or verified using Primer-BLAST with the following parameters: primer melting temperature (T_m) 55–60°C and GC content between 40% and 60% (S2 Table).

Standard PCR amplification was done using EconoTaq PLUS 2X Master Mix (Lucigen). The reaction was carried out in a total volume of 10 μl, containing 1 μl of DNA template (genomic DNA or cDNA), 1X EconoTaq PLUS Master Mix, and 1 μM of each primer (S2A and S2B Table). Thermal cycling conditions were as follows: 2 min at 94°C, followed by 40 cycles of 94°C for 30 sec, 55–60°C for 15 sec, 72°C for 1 min, and a final 5 min at 72°C and 4°C hold. PlatinumTaq DNA Polymerase (Life Technologies) was used for long-size products amplification. PCR mixes in 25 μl volumes contained: 1 μl of DNA template, 1X High Fidelity PCR Buffer, 0.2 mM of each dNTP, 2 mM MgSO4, 1 μM of each primer (S2C Table), and 1 unit of Platinum Taq High Fidelity Polymerase. PCR reactions were performed as follows: 2 min at 94°C, followed by 30 cycles of 94°C for 30 sec, 55–57°C for 30 sec, 72°C for 12 min, and a final 12 min at 72°C and 4°C hold.

Electrophoreses were done in 1.5% agarose gels containing GelRed (Biotium) in 1x TAE buffer. Half of the PCR reaction volume was used for gel electrophoresis. When additional or nonspecific products were observed, the expected size product was excised from the gel and purified using Gel Extraction Kit (Bio Basic INC). Re-PCR reaction was performed using the extracted DNA as the template and EconoTaq PLUS 2X Master Mix mentioned before. When sufficient product was obtained, the remaining half of the total reaction volume was purified with 5 μl mix of Exonuclease I and FastAP (Thermo Scientific). The samples were incubated at 37°C for 30 minutes, followed by enzyme deactivation at 80°C for 15 minutes.

Following enzymatic purification, the PCR products expected to confirm retrogene expression or presence of novel retrogenes in 17 human genomes were sequenced. Sequencing was performed with the Big Dye V3.1 Terminator Kit (Applied Biosystems) using forward and reverse primers designed for each retrogene (S2A and S2B Table) on an ABI Prism 3130xl machine (Applied Biosystems). Additionally, short PCR products corresponding to genomic loci with deletion (S2C Table) were purified and sequenced.

The BioEdit Sequence Alignment Editor [38] was used for parental gene, retrogene and sequencing result global alignment. Sequencing result specificity was verified by megaBLAST analysis.

Real-time PCR

Real-time polymerase chain reaction was performed in Applied Biosystems QuantStudio 6 & 7 Flex Real-Time PCR System using Power SYBR Green PCR Master Mix (Applied Biosystems) in 40 cycles and with Tm = 60°C. GAPDH gene was used as an endogenous control. Results were analyzed using QuantStudio Real-Time PCR Software v1.0 (Applied Biosystems). The cut-off value for Ct (cycle threshold) was established as 32, based on previously published studies [8]. Additionally, results were exported into a RDML data format [39] and used for PCR efficiency estimation using LinRegPCR software [40]. An efficiency value over 1.8 was obtained for all retrogenes and considered as acceptable. The relative amount of each retrogenes against GAPDH was calculated using the equation $2^{-\Delta Ct} \times 10^6$, where $\Delta Ct = (Ct_{retrogene}\text{-}Ct_{GAPDH})$.

Differential expression analysis

In order to analyze retrocopy expression, RNA-Seq samples from 10 individuals for each of 5 populations (CEU, GBR, FIN, TSI, YRI) from the Geuvadis RNA sequencing project [18] were used, as well as data for 16 human tissues from the Illumina Bodymap 2.0 project. Tophat [41], provided by Ensembl release 73 annotations, was used to align short reads from each sample to the reference sequence consisting of GRCh37.p10 genome assembly and de novo assembled novel retrocopy sequences. Retrocopy expression estimation was based on the number of read pairs that mapped concordantly and uniquely (mapping quality equal or greater to 50) to each locus. Only retrocopies with a normalized expression value of at least 1 RPM (reads per million mapped reads), were considered to be expressed. Furthermore, the DESeq2 package [23] was used to search for significant differences in retrocopy expression between analyzed populations, requiring at least a 2 fold expression change and adjusted p-value less than 0.05. Loci with low coverage (mean number of reads per biological replicate less than 50) were excluded from the analysis.

Supporting Information

S1 Fig. Agarose gel results showing PCR products obtained from 16 cDNA templates. Black arrows indicate PCR products of retrogenes undergoing deletions in various human populations (A-J). Lane 1 –GeneRuler Low Range DNA Ladder (Thermo Scientific); 2 –Heart; 3 –Brain; 4 –Placenta; 5 –Lung; 6 –Liver; 7 –Skeletal Muscle; 8 –Kidney; 9 –Pancreas; 10 –Spleen; 11 –Thymus; 12 –Prostate; 13 –Testis; 14 –Ovary, 15 –Small intestine 16 –Colon; 17 –Leukocyte; NTC—No template control (water instead of cDNA).
(TIF)

S1 Table. Frequencies of retrogenes loss in human populations.
(XLSX)

S2 Table. PCR primers used for experimental validation. PCR primers used in: A. retrogenes expression analysis, B. detection of novel retrogenes in 17 human genomes and expression analysis, C. detection of novel retrogene insertion sites in the reference genome. * Set of primers used also in real-time PCR. **Second pair of primers for retrogene retro_hsap_2011 was used to confirm a longer splice variant (TPM3P9-002, ENST00000424846). ***Product size with/without deletion.
(XLSX)

S3 Table. Human genomes from Coriell Cell Repositories examined in order to confirm novel retrogenes and identify indels.
(XLSX)

Author Contributions

Conceived and designed the experiments: MK IM. Performed the experiments: MK MRK AD. Analyzed the data: MK MRK WR. Contributed reagents/materials/analysis tools: MK SD AP IM. Wrote the paper: MK MRK WR IM.

References

1. Maestre J, Tchénio T, Dhellin O, Heidmann T. mRNA retroposition in human cells: processed pseudogene formation. The EMBO journal. 1995; 14(24):6333–8. PubMed PMID: 8557053
2. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nature reviews Genetics. 2009; 10(1):19–31. doi: 10.1038/nrg2487 PMID: 19030023
3. Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, et al. The evolution of courtship behaviors through the origination of a new gene in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(21):7478–83. doi: 10.1073/pnas.0800693105 PMID: 18508971

4. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008; 453(7194):539–43. doi: 10.1038/nature06908 PMID: 18404146

5. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010; 465(7301):1033–8. doi: 10.1038/nature09144 PMID: 20577206

6. Brosius J. The contribution of RNAs and retroposition to evolutionary novelties. Genetica. 2003; 118(2–3):99–116. PMID: 12868601.

7. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. PLoS biology. 2005; 3(11):e357. PMID: 16201836

8. Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makalowski W, Makalowska I. "Orphan" retrogenes in the human genome. Mol Biol Evol. 2013; 30(2):384–96. doi: 10.1093/molbev/mss235 PMID: 23066043

9. Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS biology. 2008; 6(4):e80. doi: 10.1371/journal.pbio.0060080 PMID: 18384235

10. Prendergast GC. Actin' up: RhoB in cancer and apoptosis. Nature Reviews Cancer. 2001; 1(2):162–8. PMID: 11905808

11. Corbett MA, Dudding-Byth T, Crock PA, Botta E, Christie LM, Nardo T, et al. A novel X-linked trichothiodystrophy associated with a nonsense mutation in RNF113A. J Med Genet. 2015; 52(4):269–74. doi: 10.1136/jmedgenet-2014-102418 PMID: 25612912

12. Sayah DM, Sokolskaja E, Berthoux L, Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV–1. Nature. 2004; 430(6999):569–73. PMID: 15243629

13. Consortium GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. doi: 10.1038/nature11632 PMID: 23128226

14. Ewing AD, Ballinger TJ, Earl D, Platform BIGSaAPa, Harris CC, Ding L, et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome biology. 2013; 14(3):R22. doi: 10.1186/gb-2013-14-3-r22 PMID: 23497673

15. Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, et al. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome research. 2013; 23(12):2042–52. doi: 10.1101/gr.154625.113 PMID: 24026178

16. Schrider DR, Navarro FCP, Galante PAF, Parmigiani RB, Camargo AA, Hahn MW, et al. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS genetics. 2013; 9(1):e1003242. doi: 10.1371/journal.pgen.1003242 PMID: 23359205

17. Richardson SR, Salvador-Palomeque C, Faulkner GJ. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. BioEssays: news and reviews in molecular, cellular and developmental biology. 2014; 36(5):475–81.

18. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501(7468):506–11. doi: 10.1038/nature12531 PMID: 24037378

19. Kabza M, Ciomborowska J, Makalowska I. RetrogeneDB–-a database of animal retrogenes. Mol Biol Evol. 2014; 31(7):1646–8. doi: 10.1093/molbev/msu139 PMID: 24739306

20. Pan D, Zhang L. Burst of young retrogenes and independent retrogene formation in mammals. PloS one. 2009; 4(3):e5040. doi: 10.1371/journal.pone.0005040 PMID: 19325906

21. Jun J, Ryvkin P, Hemphill E, Mandoiu I, Nelson C. The Birth of New Genes by RNA- and DNA-Mediated Duplication during Mammalian Evolution. dxdoiorg. 2009; 16(10):1429–44.

22. Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol Biol Evol. 2011; 28(1):33–7. doi: 10.1093/molbev/msq260 PMID: 20889727

23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. PMID: 25516281

24. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science (New York, NY). 2010; 328(5979):710–22.

25. Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 2008; 456(7220):387–90. doi: 10.1038/nature07446 PMID: 19020620

26. Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(9):3220–5. PMID: 16492757

27. Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mechanisms of development. 2001; 106(1–2):3–23. PMID: 11472831

28. Fontanillas P, Hartl DL, Reuter M. Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. PLoS genetics. 2007; 3(11):e210. PMID: 18081425

29. Emerson JJ, Kaessmann H, Betrán E, Long M. Extensive gene traffic on the mammalian X chromosome. Science (New York, NY). 2004; 303(5657):537–40.

30. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010; 26(6):841–2.

31. Paten B, Herrero J, Beal K, Birney E. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics. 2009; 25(3):295–301. doi: 10.1093/bioinformatics/btn630 PMID: 19056777

32. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic acids research. 2014; 42(Database issue):D749–55. doi: 10.1093/nar/gkt1196 PMID: 24316576

33. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012; 1(1):18. doi: 10.1186/2047-217X-1-18 PMID: 23587118

34. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England). 2012; 28(23):3150–2.

35. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome research. 1999; 9(9):868–77. PMID: 10508846

36. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011; 21(3):487–93. doi: 10.1101/gr.113985.110 PMID: 21209072

37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. PMID: 2231712

38. Hall T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series. 1999; 41:95–8.

39. Lefever S, Hellemans J, Pattyn F, Przybylski DR, Taylor C, Geurts R, et al. RDML: structured language and reporting guidelines for real-time quantitative PCR data. Nucleic Acids Res. 2009; 37(7):2065–9. doi: 10.1093/nar/gkp056 PMID: 19223324

40. Ramakers C, Ruijter JM, Deprez RH, Moorman AF. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett. 2003; 339(1):62–6. PMID: 12618301

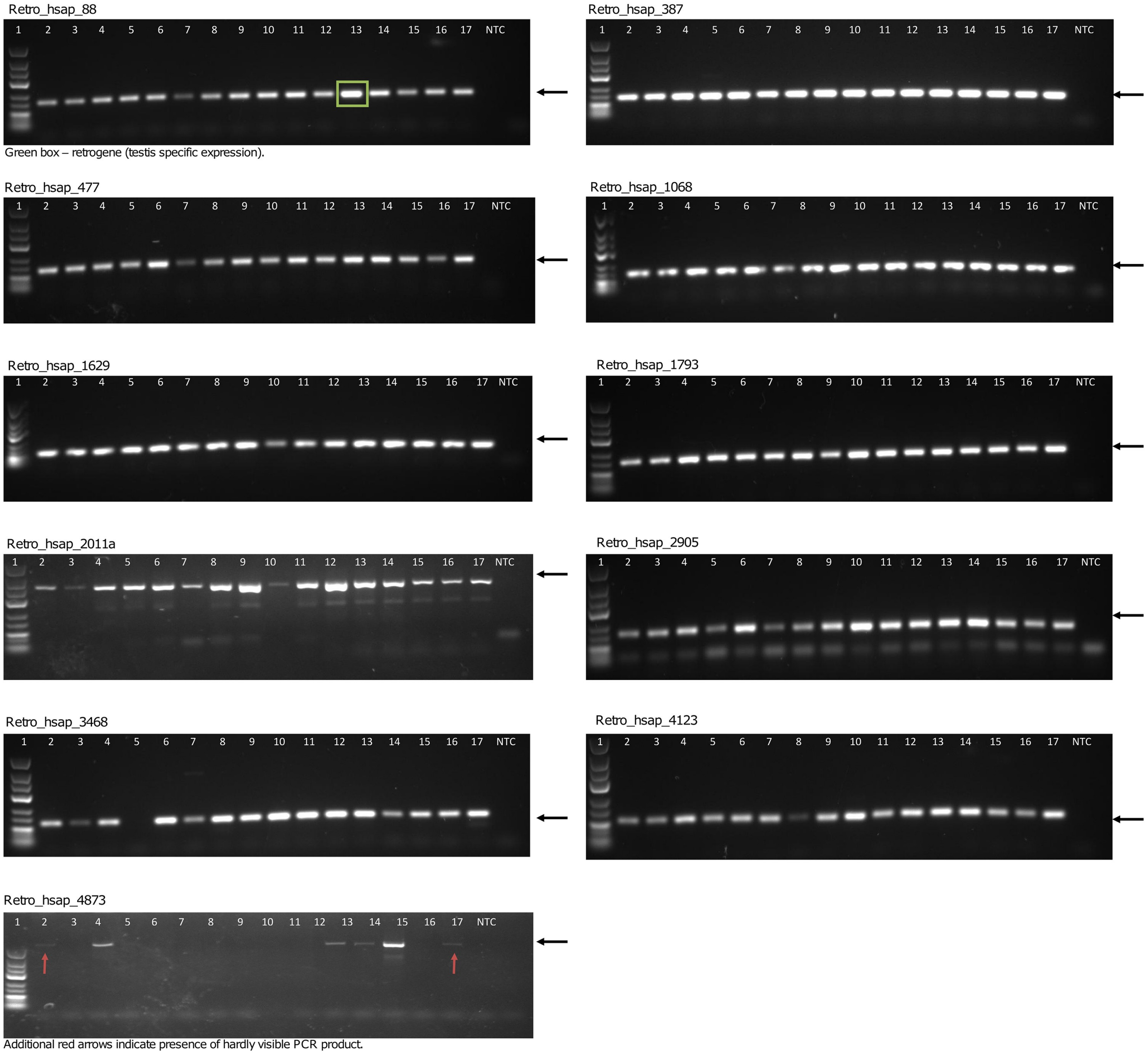
41. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013; 14(4):R36. doi: 10.1186/gb-2013-14-4-r36 PMID: 23618408

Kabza M, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polański A, Makałowska I.

Inter-population Differences in Retrogene Loss and Expression in Humans.

PLOS Genetics 2015 Oct 16;11(10):e1005579.

Supplementary materials



Green box – retrogene (testis specific expression).

Additional red arrows indicate presence of hardly visible PCR product.

S1 Fig. Agarose gel results showing PCR products obtained from 16 cDNA templates. *Black arrows indicate PCR products of retrogenes undergoing deletions in various human populations (A-J). Lane 1 – GeneRuler Low Range DNA Ladder (Thermo Scientific); 2 – Heart; 3 – Brain; 4 – Placenta; 5 – Lung; 6 – Liver; 7 – Skeletal Muscle; 8 – Kidney; 9 – Pancreas; 10 – Spleen; 11 – Thymus; 12 – Prostate; 13 – Testis; 14 – Ovary, 15 – Small intestine 16 – Colon; 17 – Leukocyte; NTC – No template control (water instead of cDNA).*

S1 Table. Frequencies of retrogenes loss in human populations

No.	Retrogene ID	Indel			Status		Population													
							AFR, African			AMR, Ad Mixed American			EUR, European					EAS, East Asian		
		Start	End	Length	Conserved	Expressed	ASW	YRI	LWK	MXL	PUR	CLM	CEU	IBS	GBR	FIN	TSI	JPT	CHB	CHS
1	retro_hsap_37	12853864	12919352	65488	no	no	0	0	1,55	1,52	1,82	4,17	2,94	3,57	2,25	1,08	2,04	6,18	2,58	3,5
1	retro_hsap_37	12898652	13013315	114663	no	no	0	1,14	0,52	2,27	0,91	4,17	1,18	0	1,12	0	0,51	2,25	0,52	1,5
1	retro_hsap_37	12907341	13183071	275730	no	no	0	0,57	1,03	0	0	0	0	0	0	0	0	1,12	1,55	0,5
2	retro_hsap_88	25671963	25673235	1272	conserved	expressed	0	0	3,09	0,76	0	1,67	0,59	0	0	1,08	0,51	1,69	1,03	0
3	retro_hsap_111	1587090	1653303	66213	no	expressed	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	retro_hsap_116	15920717	15937082	16365	no	no	0,82	0	0	0	0	0	0	0	0	0	0	0	0	0
5	retro_hsap_118	16150596	16156384	5788	no	no	1,64	0	0	5,3	5,45	2,5	5,29	0	3,93	6,45	7,65	0	0	0
5	retro_hsap_118	16151934	16155438	3504	no	no	11,48	5,68	10,31	32,58	28,18	31,67	34,12	25	39,89	31,72	37,24	0	4,64	12
6	retro_hsap_128	27502051	27515654	13603	conserved	no	0	0	0	0,76	0	0	0	0	0	0	0	0	0	0
7	retro_hsap_146	46243510	46251900	8390	conserved	no	0	2,27	0,52	0	0	0	0	0	0	0	0	0	0	0
8	retro_hsap_163	62654000	62657149	3149	conserved	no	1,64	0	2,06	18,18	4,55	10	3,53	0	2,81	3,76	3,06	38,2	34,54	34
8	retro_hsap_163	62654661	62657843	3182	conserved	no	0	0	2,06	13,64	3,64	9,17	3,53	0	2,25	3,76	3,06	28,09	23,71	25
8	retro_hsap_163	62654736	62657154	2418	conserved	no	1,64	0	2,06	18,18	4,55	10	3,53	0	2,81	3,76	3,06	39,89	34,54	34,5
9	retro_hsap_164	62654000	62657149	3149	no	no	1,64	0	2,06	18,18	4,55	10	3,53	0	2,81	3,76	3,06	38,2	34,54	34
9	retro_hsap_164	62654661	62657843	3182	no	no	0	0	2,06	13,64	3,64	9,17	3,53	0	2,25	3,76	3,06	28,09	23,71	25
9	retro_hsap_164	62654736	62657154	2418	no	no	1,64	0	2,06	18,18	4,55	10	3,53	0	2,81	3,76	3,06	39,89	34,54	34,5
10	retro_hsap_169	73539134	73729779	190645	no	no	0	0	0	0	1,82	0	0	0	0	0	0	0	0	0
11	retro_hsap_222	148002073	148249381	247308	no	no	2,46	1,14	1,03	0	0	0	0	0	0	0	0	0	0	0
12	retro_hsap_226	149294682	149680227	385545	no	no	0	0,57	0	0,76	0	0	0,59	0	0,56	0,54	0,51	0,56	0	0
13	retro_hsap_237	158455169	158513838	58669	no	no	0	0	0	0	0	0	0	0	0	0,54	0	0	0	0
13	retro_hsap_237	158492323	158497838	5515	no	no	0,82	0	0	0	0	0	0	0	0,56	0,54	0,51	0	0	0
14	retro_hsap_258	168178264	168180240	1976	conserved	no	0	0	0	0,76	0	2,5	0	0	0	0	0	0	0	0
15	retro_hsap_355	25591100	25662599	71499	no	no	20,49	12,5	7,22	15,15	34,55	28,33	42,35	57,14	38,76	34,41	37,24	6,74	7,73	2,5
15	retro_hsap_355	25614200	25661699	47499	no	no	25,41	16,48	11,34	18,18	35,45	30	41,76	67,86	39,89	36,56	37,24	6,18	10,82	3
16	retro_hsap_361	28445917	28448337	2420	conserved	no	0	0	1,55	0	0	0	0	0	0	0,54	0	0	0	0
17	retro_hsap_382	50861088	50877151	16063	no	no	0	0	0	0	0,91	0	0	0	0	0	0	0	0	0
18	retro_hsap_383	50861088	50877151	16063	no	no	0	0	0	0	0,91	0	0	0	0	0	0	0	0	0
19	retro_hsap_385	51714044	51720301	6257	conserved	expressed	0	0	0	0	1,82	0,83	0	0	0	0	0	0	0	0
20	retro_hsap_401	65443323	65871480	428157	no	no	0	0	0	0	0	0	0,59	0	0	0	0	0	0	0
21	retro_hsap_402	66040938	66043901	2963	conserved	no	3,28	5,68	8,76	0	3,64	0	0	0	0	0	0	0	0	0
22	retro_hsap_421	92579197	92585020	5823	conserved	no	1,64	2,27	3,09	0	0,91	0	0	0	0	0	0	0	0	0
23	retro_hsap_441	119998638	120034562	35924	no	no	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
24	retro_hsap_445	120109975	120147301	37326	no	no	0,82	0,57	0	0	0,91	0,83	0	0	0	0	0	0	0	0
25	retro_hsap_452	149294682	149680227	385545	no	no	0	0,57	0	0,76	0	0	0,59	0	0,56	0,54	0,51	0,56	0	0
26	retro_hsap_467	168546430	168548007	1577	no	expressed	0	1,14	1,55	0	0	2,5	0	0	0,56	2,15	1,02	1,69	2,06	10
27	retro_hsap_477	174802989	174819227	16238	conserved	expressed	0	0	0	0	0	0	0	0	0	0	0	0	1,03	0,5
28	retro_hsap_496	204315397	204316837	1440	no	no	0	0	0	0	0	0	0	0	0	0	0	0,56	0	1
29	retro_hsap_505	213027399	213029207	1808	conserved	no	1,64	0,57	4,64	0	0	0	0	0	0	0	0	0	0	0
30	retro_hsap_510	220436372	220440884	4512	no	expressed	0	0	0	0,76	0	0	0	0	0	0	0	4,49	7,22	6,5
31	retro_hsap_524	229816935	229829794	12859	conserved	no	0	0	0	0	0	0	0,59	0	0	0	0	0	0	0
32	retro_hsap_539	247339436	247348942	9506	no	no	0,82	0	0	0	0	0	0	0	0	0	0	0	0	0
33	retro_hsap_546	19001659	19061470	59811	conserved	no	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
34	retro_hsap_572	48913600	48984899	71299	no	no	9,02	19,32	14,43	3,79	2,73	0,83	0	0	0,56	3,76	0	0	0	0
35	retro_hsap_582	69544695	69555321	10626	no	no	0	0	0	0	0,91	0	0	0	0	0	0	0	0	0
36	retro_hsap_586	70182372	70187739	5367	no	no	0	0	0,52	0	0	0	0	0	0	0	0	0	0	0
37	retro_hsap_611	93974119	93987739	13620	no	no	0	0	0,52	0	0	0,83	0	0	0	0	0	0	0	0

38	retro_hsap_643	5903829	5906719	2890	no	no	0,82	1,14	0	0	0	0	0	0	0	0	0	0	0	0
39	retro_hsap_644	8550461	8560143	9682	no	no	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
40	retro_hsap_653	27603746	27706369	102623	no	no	0,82	0	0	0	0	0	0,59	0	0,56	0,54	2,55	0	0	0
40	retro_hsap_653	27638253	27642095	3842	no	no	10,66	9,09	8,76	14,39	12,73	6,67	5,29	3,57	5,62	6,45	7,65	37,08	36,08	41
41	retro_hsap_685	71057372	71060540	3168	no	no	3,28	5,68	2,58	0	0,91	0,83	0	0	0	0	0	0	0	0
42	retro_hsap_687	71281002	71291091	10089	conserved	no	6,56	1,7	6,7	36,36	10,91	13,33	6,47	7,14	7,3	5,91	10,71	53,93	38,14	43
43	retro_hsap_700	81785115	81792121	7006	no	no	0	0	0	0,76	0,91	1,67	1,76	0	1,12	0,54	0	0	0	0
44	retro_hsap_714	95041735	95045717	3982	no	no	0	0	0	6,06	0	2,5	0,59	7,14	0,56	0,54	0	7,87	12,37	9
45	retro_hsap_748	4272152	4327519	55367	no	no	2,46	5,11	1,03	7,58	2,73	3,33	1,18	0	2,25	6,99	2,04	9,55	13,92	14,5
46	retro_hsap_749	4323937	4375127	51190	no	no	0,82	0,57	0	0,76	2,73	0	0	10,71	0,56	2,15	0,51	1,69	2,58	1
47	retro_hsap_753	14451066	14464379	13313	no	no	0,82	0,57	0	0	0	0	0	0	0	0	0	0	0	0,5
48	retro_hsap_838	3416397	3615503	199106	no	no	0,82	0	4,12	0	0	0	0	0	0	0	0	0	0	0
49	retro_hsap_842	4272152	4327519	55367	no	no	2,46	5,11	1,03	7,58	2,73	3,33	1,18	0	2,25	6,99	2,04	9,55	13,92	14,5
50	retro_hsap_843	4323937	4375127	51190	no	no	0,82	0,57	0	0,76	2,73	0	0	10,71	0,56	2,15	0,51	1,69	2,58	1
51	retro_hsap_856	18606213	18621784	15571	no	no	0	0	0	0	0	0,83	0	0	0,56	0	0	0	0	0
52	retro_hsap_858	25606577	25629107	22530	no	no	0	2,27	6,19	0	0,91	0	0	0	0	0	0,51	0	0	0
53	retro_hsap_888	67474655	67754573	279918	no	no	0	0	0	0	0	0,83	0	0	0	0	1,02	2,25	5,15	3,5
54	retro_hsap_906	86531790	86534997	3207	conserved	no	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
55	retro_hsap_954	7716352	7721237	4885	no	no	0	0,57	2,06	0	0,91	0	0	0	0	0	0	0	0	0
56	retro_hsap_956	7951241	8058198	106957	no	no	0	0,57	0,52	0	1,82	0,83	0	0	0	1,08	0	0	0	0,5
57	retro_hsap_966	13000468	13015289	14821	no	no	0	0,57	0	0	0	0	0	0	0	1,08	0	0	0	0
58	retro_hsap_1033	77115334	77118231	2897	conserved	no	0,82	4,55	0,52	0	0	0	0	0	0	0	0	0	0	0
59	retro_hsap_1065	126997189	127019806	22617	conserved	no	0	0	0	0	0	0	0	0	0,56	0	0,51	0	0	0
59	retro_hsap_1065	126997902	127009734	11832	conserved	no	0	0	0	0,76	0	1,67	2,94	0	2,81	2,69	1,53	0	0	0
59	retro_hsap_1065	126998016	127013350	15334	conserved	no	0	0	0	0,76	0	0,83	0	0	0,56	0	1,02	0	0	0
60	retro_hsap_1068	133378446	133424579	46133	conserved	expressed	0,82	0	0	0	0	0	0	0	0	0	0,51	0	0	0
61	retro_hsap_1095	25088940	25131895	42955	conserved	no	0	0	0	0	0	0	1,18	0	0	0	0	0	0	0
62	retro_hsap_1153	105369051	105373544	4493	no	no	13,93	15,91	19,07	5,3	4,55	10	1,76	7,14	4,49	6,99	4,59	20,22	20,62	22,5
63	retro_hsap_1207	57865912	57892003	26091	no	no	4,92	5,68	9,28	0,76	0	0	0	0	0	0	0,51	0	0	0
64	retro_hsap_1246	25928917	25942046	13129	no	no	0,82	0	1,03	0	0	0	0	0	0	0	0	0,56	0	0
65	retro_hsap_1256	41862020	41873571	11551	conserved	no	0	0	0,52	0	0	0	0	0	0	0	0	0	0	0
66	retro_hsap_1257	42642373	42647693	5320	no	no	0	0,57	0	0	0	0	0	0	0	0	0	0	0	0
67	retro_hsap_1290	99293017	99294862	1845	conserved	no	0,82	0,57	1,03	0	0	0	0	0	0	0	0	0	0	1
68	retro_hsap_1297	107510029	107518103	8074	conserved	no	0	0	0	0	1,82	0	0	0	0	0	0	0	0	0
69	retro_hsap_1305	22026268	22036271	10003	no	no	4,1	7,95	10,82	0	0	0,83	0	0	0	0	0	0	0	0
70	retro_hsap_1308	28712325	28733928	21603	no	no	0	0	0	0	0	0	0	0	0	0,54	0	0	0	0
71	retro_hsap_1366	77434252	77435332	1080	no	no	0	1,14	4,12	0	0	0	0	0	0	0	0	0	0	0
72	retro_hsap_1370	88391508	88423179	31671	conserved	no	0	0	0	0	0	0	0,59	0	0,56	0,54	0	0	0	0
73	retro_hsap_1380	103648834	103651641	2807	conserved	no	0	1,7	0,52	0	0	0	0	0	0	0	0	0	0	0
74	retro_hsap_1408	44588871	44591714	2843	no	no	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
75	retro_hsap_1413	51038426	51045098	6672	no	no	0	0	0	0	0	0,83	0	0	0	0	0	0	0	0
76	retro_hsap_1438	70351203	70356293	5090	no	expressed	0	0,57	0,52	0	0	0	0	0	0	0	0	0	0	0
77	retro_hsap_1441	73997050	74024449	27399	no	no	37,7	28,41	32,99	43,18	43,64	53,33	52,94	39,29	41,57	58,6	44,9	85,96	84,54	79,5
78	retro_hsap_1468	106689962	107181698	491736	no	no	1,64	1,7	0,52	0	0,91	0	6,47	0	0	0	0	1,12	1,03	0
78	retro_hsap_1468	106800263	107275814	475551	no	no	1,64	0	0,52	0	0	0	2,94	0	0	0	0	1,12	0	0
78	retro_hsap_1468	106855178	106936548	81370	no	no	4,92	1,14	1,55	1,52	0	0	4,71	0	0,56	2,15	0,51	1,69	0,52	0
78	retro_hsap_1468	106879059	107096555	217496	no	no	1,64	0,57	1,55	0,76	0,91	0,83	5,88	0	0,56	0,54	2,55	1,12	0	0
78	retro_hsap_1468	106885900	106918199	32299	no	no	12,3	15,91	18,56	7,58	7,27	10,83	22,94	14,29	17,42	24,19	11,22	3,37	2,06	0
79	retro_hsap_1481	35683966	35721560	37594	no	expressed	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
80	retro_hsap_1519	80190539	80252602	62063	conserved	no	0	0	0	0	0	0	0	0	0	0	0	0	0,52	0
81	retro_hsap_1578	59842499	59933821	91322	no	no	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
82	retro_hsap_1629	23509597	23514659	5062	conserved	expressed	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
83	retro_hsap_1678	34455602	34489219	33617	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0

83	retro_hsap_1678	34455678	34457659	1981	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0,52	0
84	retro_hsap_1697	68684243	68687175	2932	conserved	no	0,82	0	0	0	0	0	0	0	0	0	0	0	0	0
85	retro_hsap_1707	76887741	76945274	57533	no	no	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
86	retro_hsap_1708	78079278	78095651	16373	no	no	0	0	0	0	0	0	0,59	0	0,56	0,54	0	0	0	0
87	retro_hsap_1724	18351500	18465799	114299	no	no	24,59	11,36	16,49	35,61	30	35	44,12	32,14	47,75	46,24	40,82	7,3	12,89	5,5
88	retro_hsap_1759	41552813	41557324	4511	no	no	0,82	0,57	0	0	0	0	0	0	0	0	0	0	0	0
89	retro_hsap_1793	1502809	1506891	4082	conserved	expressed	0,82	0,57	1,03	0	1,82	0	0	0	0	0	0	0	0	0
90	retro_hsap_1803	10613995	10651447	37452	no	no	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
91	retro_hsap_1804	10761660	10766613	4953	conserved	no	1,64	3,41	3,09	0	0	0	0	0	0	0	0	0	0	0
92	retro_hsap_1837	44281451	45168500	887049	conserved	expressed	0	0	0	0	0	0	0	0	0	0	0	0	0	0
93	retro_hsap_1901	1903090	1984846	81756	no	no	0	0	0	0,76	0	0	0	0	1,12	0	0	0	0	0
94	retro_hsap_1924	21289408	21289913	505	no	no	0	2,84	0	0	0	0	0,59	3,57	0	4,3	1,02	0,56	0	0,5
95	retro_hsap_1927	23747804	23751321	3517	no	expressed	58,2	70,45	74,23	29,55	22,73	17,5	7,65	21,43	10,11	14,52	7,14	26,97	23,71	27
96	retro_hsap_1968	14599517	14601715	2198	no	no	0	0	1,03	0	0,91	0	0	0	0	0	0	0	0	0
97	retro_hsap_1980	20595809	20711949	116140	no	no	4,92	10,23	5,67	3,79	8,18	4,17	4,12	10,71	4,49	2,69	7,65	9,55	11,34	8,5
98	retro_hsap_2011	53936258	53973019	36761	conserved	expressed	0	0	0	0	0	0	0	0	0	1,08	0	0	0	0
99	retro_hsap_2030	12753336	12754493	1157	no	expressed	2,46	0	1,55	16,67	11,82	12,5	9,41	7,14	8,99	10,75	5,61	4,49	7,22	9
100	retro_hsap_2041	22532162	22571556	39394	no	no	0	0	0	0	0	0	0	0	0	0	0	1,12	0	0,5
101	retro_hsap_2075	11394742	11400295	5553	conserved	no	2,46	2,84	6,7	0	0	1,67	0	0	0	0	0	0	0	0
102	retro_hsap_2083	24554849	24558847	3998	no	no	0	0	1,03	0	0	0	0	0	0	0	0	0	0	0
103	retro_hsap_2085	26217266	26221401	4135	conserved	no	0	0	0	0	0	0	0	0	0	0	0	0	0,52	0
104	retro_hsap_2102	52793394	52798859	5465	no	no	0	0	1,03	0	0	0	0	0	0	0	0	0	0	0
105	retro_hsap_2109	61073859	61082549	8690	no	no	0	0	0,52	0	0	0	0	0	0	0	0	0	0	0
106	retro_hsap_2121	84215894	84261674	45780	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0,52	0
107	retro_hsap_2154	130249195	130254979	5784	no	no	0	0	1,55	0	0	0	0	0	0	0	0	0	0	0
108	retro_hsap_2212	222370001	222402067	32066	conserved	no	0	0,57	0	0	0	0	0	0	0	0	0	0	0	0
109	retro_hsap_2225	3732804	3738027	5223	conserved	no	0,82	0	0	5,3	2,73	5	8,24	10,71	2,25	5,38	4,59	7,3	12,89	13
110	retro_hsap_2262	58688881	58691262	2381	conserved	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
111	retro_hsap_2266	62735864	62761450	25586	no	no	0	0	4,12	0	0	0	0	0	0	0	0	0	0	0
112	retro_hsap_2298	102124389	102129179	4790	conserved	no	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
112	retro_hsap_2298	102126059	102128599	2540	conserved	no	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0,5
113	retro_hsap_2305	109310323	109312258	1935	no	no	23,77	18,18	13,4	20,45	36,36	28,33	39,41	42,86	34,27	29,03	46,43	6,18	1,55	1,5
114	retro_hsap_2318	132087258	132096606	9348	conserved	no	2,46	6,25	9,28	0	0	0	0	0	0,56	0	0	0	0,52	0
115	retro_hsap_2342	174155880	174209070	53190	conserved	no	0	0	0	0	0	0	0	0	0	0,54	0	0	0	0
116	retro_hsap_2358	187025287	187029178	3891	no	no	0	0	0	0	0	0	0	0	0	0	0	0,56	0	0
117	retro_hsap_2399	234648330	234659987	11657	no	no	0,82	3,41	3,61	0	0	0,83	0	0	0	0	0	0	0	0
117	retro_hsap_2399	234650078	234661758	11680	no	no	0,82	2,84	3,61	0	0	0,83	0	0	0	0	0	0	0	0
118	retro_hsap_2475	34663722	34665222	1500	conserved	no	0	0,57	3,61	0	0	0	1,18	0	0	0	0	0	0	0
119	retro_hsap_2523	32432179	32434435	2256	no	no	5,74	2,84	6,19	1,52	2,73	0	0,59	0	0	0	0,51	0	0	0
120	retro_hsap_2549	25065839	25110120	44281	no	no	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
121	retro_hsap_2550	25065839	25110120	44281	no	no	0	0	0	0	0	0	0	0	0,56	0	0	0	0	0
122	retro_hsap_2602	1770991	1780588	9597	no	no	0,82	1,14	0,52	0	0	0	0	0	0	0	0	0	0	0
123	retro_hsap_2620	26036805	26170483	133678	no	no	0	0	0	0	0,91	0	0,59	0	0	0	0	0	0	0
124	retro_hsap_2745	195669672	195729104	59432	no	no	0	0	0	0	0	0	0	0	0	1,08	0	0	0,52	0,5
125	retro_hsap_2905	197576281	197583117	6836	conserved	expressed	0	0	0	0	0	0	0	0	0	0	0,51	0	0	0
126	retro_hsap_2909	9151311	9690368	539057	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0
127	retro_hsap_2914	15760384	15768829	8445	conserved	no	5,74	2,27	2,06	0	0	0	0	0	0	0	0	0	0	0
128	retro_hsap_2926	43898070	43902532	4462	no	no	0	0	0	0	0,91	0	0	0	0	0	0	0	0	0
129	retro_hsap_2938	63665192	63698156	32964	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
130	retro_hsap_2948	79298986	79300656	1670	conserved	no	0,82	3,41	1,03	0	0	0	0	0	0	0	0	0	0	0
131	retro_hsap_3007	165966966	165967740	774	no	no	12,3	13,07	9,28	0	0	0,83	0	0	0	0	0	0	0	0
132	retro_hsap_3027	17543641	17569643	26002	no	no	0	0	0	0,76	0	0	0	0	0,56	0	0	1,69	2,06	1
133	retro_hsap_3028	22700766	22777132	76366	no	no	0	0	0	0	0	0	0	0	0	0	0	0,56	0	0

134	retro_hsap_3144	17527012	17606097	79085	no	no	0	0	6,19	0,76	0,91	0,83	1,76	0	0,56	2,69	1,53	1,12	0	0
135	retro_hsap_3145	17527012	17606097	79085	no	no	0	0	6,19	0,76	0,91	0,83	1,76	0	0,56	2,69	1,53	1,12	0	0
136	retro_hsap_3146	17609510	17619280	9770	no	no	4,1	4,55	1,03	0,76	1,82	0,83	2,94	0	1,69	4,3	3,06	2,25	0,52	2
137	retro_hsap_3147	17618868	17645544	26676	no	no	0,82	2,84	0	0	0,91	0	4,12	3,57	2,81	3,23	2,04	1,12	0	1,5
138	retro_hsap_3148	17618868	17645544	26676	no	no	0,82	2,84	0	0	0,91	0	4,12	3,57	2,81	3,23	2,04	1,12	0	1,5
139	retro_hsap_3149	17636107	17649848	13741	no	no	1,64	2,27	0,52	0	0	0	2,94	0	1,69	0,54	2,04	0,56	5,15	10,5
139	retro_hsap_3149	17639240	17651718	12478	no	no	1,64	2,27	0,52	0	0	0	1,18	0	0	0	2,04	0,56	5,67	11
139	retro_hsap_3149	17646433	17694447	48014	no	no	0	0,57	0	0	0	0	0,59	0	0	0	1,02	0	0	2
140	retro_hsap_3172	51577104	51580306	3202	no	no	10,66	9,66	3,09	0	0,91	0,83	0	0	0	0	0	0	0	0
141	retro_hsap_3177	55485600	55493454	7854	conserved	no	0	0	0	0	0	0,83	0	0	0	0	0	0	0	0
142	retro_hsap_3220	96704395	96705197	802	conserved	no	0	0	1,03	0	0	0	0	0	0	0	0	0	0	0
143	retro_hsap_3221	97047854	97096157	48303	conserved	no	0	0	0	0,76	4,55	0,83	2,94	0	2,25	6,45	3,06	0	0	0
144	retro_hsap_3244	137807484	137816900	9416	no	no	2,46	0	0	0,76	0	0,83	2,35	3,57	3,37	2,69	2,55	0	0	0
145	retro_hsap_3260	156848606	156850880	2274	conserved	no	0	0	0	0	0	0	0	0	0	0	0	7,3	4,12	3,5
146	retro_hsap_3277	178348361	178353196	4835	conserved	no	2,46	2,84	2,06	2,27	13,64	10	10	10,71	10,67	13,98	13,78	0	0	0
147	retro_hsap_3280	180375026	180430527	55501	conserved	no	13,93	7,95	10,82	33,33	31,82	32,5	34,71	32,14	29,78	36,02	34,18	28,09	32,99	20
148	retro_hsap_3286	5375709	5378380	2671	conserved	no	0	0	0	0	0	0	0,59	0	0	0	0,51	0	0	0
149	retro_hsap_3293	17342773	17357155	14382	no	no	0,82	0	0	0	0	0,83	0	0	2,81	0	0	0	0	0
150	retro_hsap_3295	17527012	17606097	79085	no	no	0	0	6,19	0,76	0,91	0,83	1,76	0	0,56	2,69	1,53	1,12	0	0
151	retro_hsap_3296	17527012	17606097	79085	no	no	0	0	6,19	0,76	0,91	0,83	1,76	0	0,56	2,69	1,53	1,12	0	0
152	retro_hsap_3297	17527012	17606097	79085	no	no	0	0	6,19	0,76	0,91	0,83	1,76	0	0,56	2,69	1,53	1,12	0	0
153	retro_hsap_3298	17527012	17606097	79085	no	no	0	0	6,19	0,76	0,91	0,83	1,76	0	0,56	2,69	1,53	1,12	0	0
154	retro_hsap_3299	17646433	17694447	48014	no	no	0	0,57	0	0	0	0	0,59	0	0	0	1,02	0	0	2
155	retro_hsap_3302	23303355	23304789	1434	no	no	0	1,14	0	0	0	0	0	0	0	0	0	0	0	0
156	retro_hsap_3343	79652780	79659332	6552	no	no	1,64	0,57	2,58	1,52	0	0	0	0	0	0	0,51	0	0	0
157	retro_hsap_3420	7984273	7987007	2734	conserved	no	0	0	0	0	0	0	0	0	0	0,54	0	0	0	0
158	retro_hsap_3448	31220481	31312867	92386	no	no	0,82	0	0	0	0	0	0	0	0,56	0	0	0	0	0
159	retro_hsap_3468	47564728	47566909	2181	conserved	expressed	0,82	0	0	3,03	3,64	5	5,29	3,57	11,24	3,76	5,61	0	0	0
160	retro_hsap_3497	90506539	90509898	3359	conserved	no	0	1,14	1,03	2,27	0	0	0	0	0	0	0	0	0	0
161	retro_hsap_3572	29740815	29886060	145245	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0
161	retro_hsap_3572	29766542	29922999	156457	no	no	0	0	0	0	0	0	0	0	0	0	0	0,56	0	0
162	retro_hsap_3574	31220481	31312867	92386	no	no	0,82	0	0	0	0	0	0	0	0,56	0	0	0	0	0
163	retro_hsap_3600	57709732	58316811	607079	no	no	0	0	0	0,76	0	0	0	0	0	0	0	0	0	0
164	retro_hsap_3633	93596328	93599179	2851	no	no	0,82	1,14	1,03	0	0	0	0	0	0	0	0	0	0	0
165	retro_hsap_3694	3374358	3417555	43197	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
166	retro_hsap_3702	16247618	16330928	83310	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0,52	0
167	retro_hsap_3744	64544521	65238651	694130	conserved	no	0	0	0	0	0	0	0,59	0	0,56	0	0,51	0	0	0
168	retro_hsap_3780	124982594	124987784	5190	no	no	0	0	0	0	0	1,67	0	0	0	0	0	0	0	0
169	retro_hsap_3781	124982594	124987784	5190	no	no	0	0	0	0	0	1,67	0	0	0	0	0	0	0	0
170	retro_hsap_3798	138362584	138376351	13767	conserved	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
171	retro_hsap_3852	56742257	57178108	435851	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0
172	retro_hsap_3853	56742257	57178108	435851	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0
172	retro_hsap_3853	56877587	56879089	1502	no	no	3,28	10,8	5,67	0	0	0	0	0	0	0	0	0	0	0
173	retro_hsap_3861	64544521	65238651	694130	conserved	no	0	0	0	0	0	0	0,59	0	0,56	0	0,51	0	0	0
174	retro_hsap_3872	76128156	76626961	498805	no	no	1,64	0	0,52	0,76	0	0	0,59	0	0	0	0,51	0	0	0
175	retro_hsap_3884	102726666	102740091	13425	no	expressed	0	0	0	0	0	0,83	0	0	0	0	0	0	0	0
176	retro_hsap_3894	118590976	118599042	8066	no	no	0	1,14	0	0	0,91	0	0	0	0	0	0	0	0	0
177	retro_hsap_3895	121059274	121084704	25430	conserved	no	0	0	0	0	0	0	0	0	0	0	0	0	0,52	0
178	retro_hsap_3937	18850880	18865271	14391	no	no	0	0	0	0,76	0,91	0,83	0	0	1,12	0	2,04	0	0	0
179	retro_hsap_3940	28091134	28105267	14133	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
180	retro_hsap_3949	48066058	48072448	6390	no	no	4,92	13,07	7,22	0	0,91	0,83	0,59	0	0	2,15	0	1,12	0	0,5
181	retro_hsap_3952	51402350	51405313	2963	conserved	no	0,82	0	0	0	0	0	0	0	1,12	0	0	0	0,52	0,5
182	retro_hsap_4020	11870990	12515284	644294	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0

183	retro_hsap_4021	11870990	12515284	644294	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0
184	retro_hsap_4051	48458704	48460226	1522	conserved	no	0,82	0	0,52	0	0	0	0	0	0	0	0	0	0	0
185	retro_hsap_4123	5106619	5115630	9011	conserved	expressed	0,82	0,57	1,03	0	0	0	0	0	0	0	0	0	0	0
186	retro_hsap_4170	43642661	44253809	611148	no	no	0,82	0	0	0	0	1,67	0,59	0	0	1,08	0	0	0	0
187	retro_hsap_4212	117606275	117609289	3014	conserved	no	0	0	0	0	0,91	0	0	0	0	0	0,51	0	0	0
188	retro_hsap_4218	128358593	128366122	7529	no	expressed	0	0	0	0	0	0	0	0	1,12	0	0	0	0	0
189	retro_hsap_4226	6648894	6701884	52990	no	no	0	0	0	0	0	0	0	0	0	0	0	0	0	0,5
189	retro_hsap_4226	6663084	6671193	8109	no	no	0	0	0	10,61	2,73	9,17	0	0	0	0	0	0	0	0,5
190	retro_hsap_4244	33015516	33023806	8290	conserved	no	0,82	0,57	0,52	0	0	0	0	0	0	0,54	0	0	0	0
191	retro_hsap_4282	91020166	91029459	9293	no	no	0	0	1,03	0	0	0	0	0	0	0	0	0	0	0
192	retro_hsap_4290	102878567	102880666	2099	no	no	0	0	0	0	0	0	0	0	0	0,54	0	0	0	0
193	retro_hsap_4296	109739492	109742051	2559	conserved	no	3,28	5,68	12,89	0	0	0	0	0	0	0	0	0	0	0
194	retro_hsap_4714	75025962	75142729	116767	no	no	0	0	0,68	0	0	0	0	0	0	0	2,05	0	0	0
195	retro_hsap_4727	91639200	91754099	114899	no	no	0	0,75	0,68	0,99	0	0	0	0	0	0	2,05	0	0	0
196	retro_hsap_4837	53478158	53491615	13457	no	no	3,06	3,01	0,68	1,98	1,22	1,1	0	0	0	0	2,05	0	0	0
197	retro_hsap_4845	65020882	65044067	23185	conserved	no	0	0	0,68	0	0	1,1	0	0	0	3,31	2,05	0	0	0
198	retro_hsap_4873	91367534	91368794	1260	conserved	expressed	15,31	11,28	15,07	0,99	1,22	0	0	0	0	0	0	0	0	0
199	retro_hsap_4911	142602572	142801112	198540	no	no	0	0	0,68	0	0	1,1	0	0	0	0,66	2,05	0	0	0

S2 Table. PCR primers used for experimental validation. PCR primers used in:
A. retrogenes expression analysis, B. detection of novel retrogenes in 17 human genomes and expression analysis, C. detection of novel retrogene insertion sites in the reference genome.

	Retrocopy	Primer name	Primer Sequence	Primer length (bp)	Annealing temperature (°C)	Product size (bp)
A.	retro_hsap_88	retro_hs_88F	GCGCTCTACTCCTGTAACGG	20	60	147
		retro_hs_88R	TCGTAGAGCATCGCCTCAGT	20		
	retro_hsap_385	retro_hs_385F*	AGGCCACCTTTGATGCCATT	20	60	126
		retro_hs_385R*	CTCTGGCGTGGGCCTTGAT	19		
	retro_hsap_477	retro_hs_477F	AGGTTTCCTGCACTGGTGTT	20	55	122
		retro_hs_477R	CCATTCCTCTCAGTTCTCTAGGTT	24		
	retro_hsap_1068	retro_hs_1068F*	TCTGTGAGCTTCAGTTCCAATGAG	24	60	130
		retro_hs_1068R*	TGAAAGATGGTCGGCTGCTTT	21		
	retro_hsap_1629	retro_hs_1629F*	CCTGCCCAGTCGGATTTAGG	20	60	106
		retro_hs_1629R*	GCGTACTTTCCCTTGGATGGT	21		
	retro_hsap_1793	retro_hs_1793F*	TTGTCGGCTGTGTGAAGCA	19	60	141
		retro_hs_1793R*	CCGCCAAGTTTCCAGGTGTA	20		
	retro_hsap_2011	retro_hs_2011aF	ACCAGGCAGATGATGCAGAGGAGT	24	60	381
		retro_hs_2011aR	TCGTTCCTCTGTGCATCCCAT	21		
		retro_hs_2011bF**	AGATTCGTGCAGACCAGGA	19	55	416
		retro_hs_2011bR**	GGCCCAGATTTCAATAACCT	20		
	retro_hsap_2905	retro_hs_2905F	CTCTCTGAGCGTGTTTCCTTGTTC	24	60	150
		retro_hs_2905R	AACACGAAGATTGGAACCTCTTGA	24		
	retro_hsap_3468	retro_hs_3468F	ACTGGTGTCGTGGAGTTTGG	20	60	115
		retro_hs_3468R	CACCCTGGAGTCCTCCTTTG	20		
	retro_hsap_4123	retro_hs_4123F*	CTCTCCGGATCACTCAAGCA	20	60	103
		retro_hs_4123R*	GACTCCTGCTCCAGGCTAGT	20		
	retro_hsap_4873	retro_hs_4873F	CAAGGAAACCACAGGTATGT	20	60	778
		retro_hs_4873R	TAGGTATCTCCGGTTAAGCA	20		
B.	rdn1	rdn1F	TTGGAAGTCAAAAGTTGGTCAGA	23	55	445
		rdn1R	CCAAGTTTGGTAAGAGTGTTCAGA	24		
	rdn2	rdn2F	AGAAACATTCTTGTAGTCCAGAGGT	25	55	593
		rdn2R	CTGAACTGCATGTGAACCAATCA	23		
	rdn3	rdn3F	CTTCTTCTGAATCATCTTCCCAGC	24	58	377
		rdn3R	ATATGGATAGATGCCCAACTGC	22		
	rdn4	rdn4F	CAGAGAAGTGGCTGGAAGGACTG	23	58	362
		rdn4R	TGGGGATACAATGGGGTCCTCT	22		
	rdn5	rdn5F	AGTCAGCTATTTAATGAGGTTCTTA	25	56	519
		rdn5F	ATCCTACTGAAGATGTACCAGTTA	24		
C.	rdn1	long1F	GCGAAGTAAGGGGATTTTTGCT	22	55	602 / ~1,500***
		long1R	GGAGTAGTTTGTTGGTCTTCCCT	23		
	rdn2	long2F	CATTTCGGTGGCTTAGCAGC	20	55	529 / ~4,000***
		long2R	CTGAGGAGGGGCATTCATGG	20		
	rdn3	long3F	GTTGACAGCAGTAACAGTTCTCATT	25	57	154 / ~10,000***
		long3R	ATATCTTACGTGGACGGCAGTAGTC	25		

* Set of primers used also in real-time PCR.

** Second pair of primers for retrogene retro_hsap_2011 was used to confirm a longer splice variant (TPM3P9-002, ENST00000424846).

*** Product size with/without deletion.

S3 Table. Human genomes from Coriell Cell Repositories examined in order to confirm novel retrogenes and identify indels.

No.	Id	Description	Gender
1	NA18968	Japanese in Tokyo, Japan	Female
2	HG01384	Colombian in Medellin, Colombia	Female
3	NA20532	Toscani, Italia	Male
4	NA21091	Gujarati Indians in Houston, Texas, USA	Male
5	HG00271	Finnish, Finland	Male
6	NA21297	Maasai in Kinyawa, Kenya	Female
7	NA12878	Ceph/Utah Pedigree 1463	Female
8	HG01097	Puerto Rican, Puerto Rico	Male
9	HG00864	Chinese Dai in Xishuangbanna, China	Female
10	HG01571	Peruvian in Lima, Peru	Male
11	HG01860	Kinh in Ho Chi Minh City, Vietnam	Male
12	HG00102	British From England And Scotland, UK	Female
13	HG02489	African Ancestry From Barbados, The Caribbean	Male
14	NA19240	Yoruba in Ibadan, Nigeria	Female
15	NA20322	African Ancestry, Southwest USA	Female
16	NA19729	Mexican Ancestry in Los Angeles, California, USA	Male
17	HG01702	Iberian Populations, Spain	Female