A Dual Origin of the \textit{Xist} Gene from a Protein-Coding Gene and a Set of Transposable Elements

Eugeny A. Elisaphenko\textsuperscript{1}, Nikolay N. Kolesnikov\textsuperscript{1}, Alexander I. Shevchenko\textsuperscript{1}, Igor B. Rogozin\textsuperscript{1,2}, Tatyana B. Nesterova\textsuperscript{3}, Neil Brockdorff\textsuperscript{4}, Suren M. Zakian\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}Institute of Cytology and Genetics, Russian Academy of Sciences, Siberian Department, Novosibirsk, Russia, \textsuperscript{2}National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, \textsuperscript{3}Medical Research Council, Clinical Sciences Centre, Imperial College Faculty of Medicine, London, United Kingdom, \textsuperscript{4}Department of Biochemistry, University of Oxford, Oxford, United Kingdom

Abstract

X-chromosome inactivation, which occurs in female eutherian mammals is controlled by a complex X-linked locus termed the X-inactivation center (XIC). Previously it was proposed that genes of the XIC evolved, at least in part, as a result of pseudogenization of protein-coding genes. In this study we show that the key XIC gene \textit{Xist}, which displays fragmentary homology to a protein-coding gene \textit{Lnx3}, emerged \textit{de novo} in early eutherians by integration of mobile elements which gave rise to simple tandem repeats. The \textit{Xist} gene promoter region and four out of ten exons found in eutherians retain homology to exons of the \textit{Lnx3} gene. The remaining six \textit{Xist} exons including those with simple tandem repeats detectable in their structure have similarity to different transposable elements. Integration of mobile elements into \textit{Xist} accompanies the overall evolution of the gene and presumably continues in contemporary eutherian species. Additionally we showed that the combination of remnants of protein-coding sequences and mobile elements is not unique to the \textit{Xist} gene and is found in other XIC genes producing non-coding nuclear RNA.

Introduction

The majority of genes on one of the two X chromosomes are inactivated in female mammals during early development. It is postulated that this phenomenon is required to equalize X chromosome gene expression in males (XY) and females (XX) [1]. In eutherians X inactivation is controlled by the X-linked inactivation center (XIC). The XIC is a complex genetic locus comprising of a number of genes producing non-coding nuclear RNAs [2]. Key amongst these is the X inactive specific transcript (Xist), that plays a pivotal role in X inactivation. Xist RNA spreads in cis along the X chromosome to be inactivated, initiates transcriptional silencing, and triggers chromatin modifications that maintain an inactive state of the genes [3].

Duret et al. [4] proposed that the XIC genes \textit{Tsx} (Testis specific X-linked gene), \textit{Xist}, \textit{Enox(Jps)} (Expressed neighbour of \textit{Xist}), \textit{Fix} (Mus musculus \textit{FTX} non-coding RNA), and \textit{Cahp2} (Cellular nucleic acid binding protein 2) originated through pseudogenization of the protein-coding genes \textit{Fip1l2} (Polyadenylation factor I complex (Saccharomyces cerevisiae), like 2), \textit{Lnx3} (Ligand of numb-protein X\textsubscript{3}), \textit{Rasl11c} (RAS-like, family 11, member C\textsubscript{3}, \textit{Uspl} (Ubiquitin specific peptidase), and \textit{Wave4(Wasf3)} (Wiskott-Aldrich syndrome protein family member 2\textsubscript{3}, respectively. This assumption is based on the fact that the genes in question are located within the regions of synteny in chicken (\textit{Gallus gallus}) and human (\textit{Homo sapiens}) and contain homologous regions. Xist may have evolved in part from the \textit{Lnx3} gene, which in chicken produces an mRNA encoding a protein similar to PDZ domain containing ring finger 1 (LOC422320, ACC XM_420296.1). \textit{Lnx3} exons 4 and 11 show similarity to exons h4/m4 and h5/m6 of human and mouse \textit{XIST/Xist}, respectively. The function of \textit{Lnx3} in chicken is unknown, and this gene ceased to exist in eutherians after contributing to \textit{Xist} formation. Six exons of \textit{Fip1l2} show similarity to the orthologous region of human and mouse. Three of them correspond to mouse \textit{Tsx} or human pseudo-\textit{TSX}. Four \textit{Rasl11c} exons are homologous to sequences in the canine and bovine XIC, and one \textit{Rasl11c} exon displays similarity to a sequence in the human XIC [4]. The contribution of \textit{Uspl} and \textit{Wave4(Wasf3)} to the formation of XIC genes remains to be clarified.

In this study we have conducted an independent analysis of the origin and evolution of the X inactivation center and the \textit{Xist} gene in eutherians using a set of bioinformatic approaches. We have compiled a consensus \textit{Xist} gene reconstructed from sequences that had undergone considerable divergence in each species. As a result, additional data have been obtained demonstrating how the XIC originated from a region containing protein-coding genes. In particular, we have demonstrated that the genes \textit{Enox(Jps)} and \textit{Fix} of the XIC contain exons homologous to those in cognate protein-coding genes \textit{Uspl} and \textit{Wave4(Wasf3)}, and we have discovered at least three additional exons of \textit{Lnx3} in the \textit{Xist} gene sequence. Moreover we have found that many \textit{Xist} exons originated from mobile elements of various classes, which gave rise to simple
tandem repeats detectable in the structure of the gene. Based on the analysis of the Xist gene consensus, we have reconstructed an ancestral structure of this gene, which may have existed 100 million years ago (Mya) in the eutherian ancestor, and traced its evolution to contemporary eutherian species.

Results

Xist consensus

To examine further the origin of the Xist gene we have generated an Xist gene consensus from genomic sequences of rat, bovine, canine, and human Xist/XIST, belonging to four eutherian orders (Rodentia, Cetartiodactyla, Carnivora, and Primates, respectively). Based on PipMaker [5] comparison we removed all species-specific repeats (e.g. species-specific SINEs, tandem repeats, pseudogenes etc) from the annotated Xist sequences and then aligned the sequences by ClustalX [6]. The resulting alignment was then manually corrected. The Xist consensus sequences were generated by the AnnHyb program (http://www.bioinformatics.org/annhyb). The resulting Xist gene consensus contains 10 exons and 500-bp flanking regions at the 5’ and 3’ ends. The size of the genomic locus is about 30 kb (Table 1, Fig. 1).

The correspondence between exons in consensus, human and mouse Xist is shown in table 2. The consensus includes all known types of tandem repeats (A-F) (Table 3), which are essential components of the gene. Additionally we detected a previously unrecognized repeat H (Table 3) and mobile elements (Fig. S1) in the Xist consensus sequence, that were reconstructed from fragments found in the nucleotide sequences of the gene in different species. The Xist gene consensus sequence was used in further analyses described below.

Homology of Xist and Lnx3 exons

It has been shown previously that chicken Lnx3 exons 4 and 11 have similarity to human XIST exons 4 and 6 [4]. Using FASTA, SSEARCH and WUBLAST programs we performed a more comprehensive comparison between the chicken Lnx3 gene and human XIST to obtain additional evidence that Lnx3 is the precursor of Xist. In addition to the previously found similarity we discovered that Lnx3 exons 3 and 5 are similar to human XIST introns 3 and 4 (59 and 60%, respectively). Interestingly, human XIST intron 4 corresponds to rodent Lnx3 exon 5, which displays 65% homology to Lnx3 exon 5 (Fig. 2, Fig. S2). We also discovered additional similarity between the XIST promoter region and Lnx3 exons 1 and 2 (Fig. 2, Fig. S2). Moreover, the same fragment of ancient LINE, belonging to the L3 family (widespread in the genomes of the majority of vertebrate species), is located at the orthologous position in Lnx3 and the Xist consensus 5’ region. The similarity between the chicken mobile element and mammalian L3 consensus is 64.7%. Comparison between the Xist consensus described above and chicken Lnx3 revealed the same homologous regions as in human XIST (Fig. 2, Fig. S2). The P-value indicates that the similarities are statistically significant (Table 4). Therefore, our data substantiates and adds to previous observations that eutherian Xist consensus has extended homology to Lnx3 in the promoter region, exons 1, 5–7, and intron 4.

Contribution of transposons to the formation of the Xist gene

We compared consensus exon sequences with the database of chordate mobile elements using the program WU-BLAST 2.0 and elevated search sensitivity by setting the word size W = 3 instead of the default value W = 10. We identified a short fragment of MIR3 in Xist exon 1. The consensus exon 2 (m2/pha2) is similar to the ancient L3CR element (Fig. S3, Table 2). Unexpectedly, exon 3 of the consensus, which corresponds to human exon 2, has similarity to an L1 fragment and is present in the spliced transcript (ACC M97168). The boundaries of this human exon are inside the L1 sequence, i.e., human exon 2 is composed of a 64 nt fragment of a mobile element which is included in mature XIST RNA (Fig. 1). Presumably, a similar phenomenon takes place in the splicing of canine Xist RNA too (Fig. 1). Consensus exon 4 (m3/h3) is also similar to L1 sequence (Fig. S3). A considerable part of exon 8 (h6/m7) is similar to mobile elements of different classes (SINEs, LINEs, and DNA transposons) (Fig. 1). Exon 9 (pm7/h7) shows 64% similarity to endogenous retroviruses. Exon 10 (m8/h8) displays the highest similarity to DNA transposons. Almost the entire exon displays a 61% similarity to the hAT-10_XT element and the central part, 66% similarity to the ZOMBI transposon (Fig. S3).

The degree of sequence identity of consensus exons 2, 4,

| Table 1. Relative length of Xist elements. |
|-------------------------------------------|
| Species | Gene size | Length of homologous exons | Total exon length |
|         | Ex1 | pEx2 | Ex3 | Ex4 | pEx5 | Ex5 | Ex6 | Ex7 | Ex8 | Ex1 | pEx2 | Ex3 | Ex4 | pEx5 | Ex5 | Ex6 | Ex7 | Ex8 |
| C.f.    | 37592 | 15480 | 96  | 59  | 140  | 211 | 162 | 130 | 4561 | 195 | 370 | 16188 | 21404 |
| B.t.    | 34934 | 18693 | 90  | -   | 137  | 210 | 164 | 131 | 4524 | 156 | 372 | 10457 | 24477 |
| H.s.    | 32063 | 11333 | 90  | 64  | 137  | 209 | 161 | 164 | 4543 | 146 | 378 | 14838 | 17225 |
| P.t.    | 32050 | 11316 | 90  | 64  | 137  | 209 | 144 | 164 | 4541 | 146 | 378 | 14861 | 17189 |

Notes. Pseudoexon (pEx) means a part of an intronic sequence which is present in other species as an exon. Abbreviations: exons (Ex) and pseudoexons (pEx), in Canis familiaris (C.f.), Bos taurus (B.t.), Homo sapiens (H.s.), Pan troglodytes (P.t.), Rattus norvegicus (R.n.), Mus musculus (M.m.), Microtus rossiaemeridionalis (M.r.), and consensus Xist gene; (-) the exon is missing. Length is shown in base pairs.

doi:10.1371/journal.pone.0002521.t001
9, and 10 to mobile elements varies from 59 to 64%. All similarities found are statistically significant (Table 4). This is also the case for the degree of similarity between consensus exons 5, 6, and 7, and the exons of the \textit{Lnx3} gene. Thus, we have shown that \textit{Xist} consensus exons 2–4 and 8–10 share some homology with different classes of mobile elements (Fig. S3). Additionally, our data suggests that after the divergence of the main mammalian taxa, young species-specific repeats (Alu of primates, B1 and B2 of rodents, etc.) and pseudogenes were integrated in the \textit{Xist} gene (Fig. 1). Some young taxon-specific SINEs were inserted in exon 1 and are now present in the processed \textit{Xist} transcript. Interestingly, \textit{Xist} CpG islands show a unique distribution in each species and are predominantly coincide with the insertion of species-specific mobile elements (Fig. 1).

![Figure 1. Organization of the \textit{Xist} gene in seven eutherian species and the \textit{Xist} gene consensus.](image)

Grey rectangles represent exons; green, pseudoxons (pEx); blue arrows, LINES; and red, SINES; brown rectangles indicate tandem repeats (for details see table 3); black rectangles, CpG-islands. Note that in chimpanzee \textit{Xist} gene tandem repeats (B, Bh and C) are not detected possibly due to gaps in sequence. Pseudoxon (pEx) means a part of an intronic sequence that is present in other species as an exon.

doi:10.1371/journal.pone.0002521.g001

---

### Table 2. Correspondences between consensus, mouse and human \textit{Xist} elements and their putative origin.

| \textit{Xist} elements | Human \textit{XIST} | Mouse \textit{Xist} | Origin from \textit{Lnx3} | Origin from transposable elements |
|------------------------|---------------------|---------------------|---------------------------|----------------------------------|
| Pmin                   | Pmin                | Pmin                | Exon 1, Exon 2             | TR, TE                           |
| c1                     | h1                  | m1                  | Exon 2                    | LINE, L3CR1                      |
| c2                     | ph2                 | m2                  | Exon 3                    | LINE, L1MC3                      |
| c3                     | h2                  | -                   | Exon 4                    | LINE, L1                         |
| Intron 4               | Intron 3            | Intron 3            | Exon 3                    |                                  |
| c5                     | h4                  | m4                  | Exon 4                    |                                  |
| c6                     | ph5                 | m5                  | Exon 5                    |                                  |
| c7                     | h5                  | m6                  | Exon11                    |                                  |
| c8                     | h6                  | m7                  | Exon 7                    | TR, TE                           |
| c9                     | h7                  | pm7                 | Exon 8                    | Fragment of ERV                  |
| c10                    | h8                  | m8                  | Exon 10                   | Fragment of DNA transposon       |

Notes. Abbreviations: exons of the mammalian consensus gene – (c1–c10), human and mouse exons, respectively – (h1/m1 – h8/m8); pseudoxons – (ph, pm). Minimal promoter – (Pmin). Tandem repeats – (TR), transposable elements – (TE), endogenous retrovirus – (ERV). Pseudoxon (pEx) means a part of an intronic sequence which is present in other species as an exon.

doi:10.1371/journal.pone.0002521.t002
Tandem repeats in Xist originated from mobile elements

The majority of Xist exon 1 consensus consists of various tandem repeats (A–F), which may represent up to 78% of the exon, as observed in bovine Xist (Table S1). The beginning of exon h6/m7 also contains a tandem repeat E (Fig. 1, Table 3). Using the program TRF4, we deduced the consensus sequences for the main blocks of tandem repeats in the Xist gene and compared the consensus dimers with the database of dispersed repeats (RepBase). This analysis revealed that repeats A, C, and D display a similarity to different classes of endogenous retroviruses (Table 4, Fig. S4). Three similar variants of repeat F are detectable in the Xist gene consensus and they all display some similarity to DNA transposons. The ancient block of H repeats, which we identified in the Xist consensus sequence, displays a high similarity to the 3' end of marsupial (M. domestica) LINE L1-7_MD L1. Repeat B is a microsatellite tract (CCCCAG)n, which is found in dispersed repeats and could have been introduced, for example, with the same endogenous retroviruses (Table 4, Fig. S4). It should be noted that a

Table 3. Tandem repeats in the Xist gene.

| No | Repeat name, exon | Description | References | Possible origin |
|----|-------------------|-------------|------------|----------------|
| 1. | A exon 1 | Repeat A is present in all the species studied and is the most important region of the RNA for chromosome silencing. It has a monomer with an average size of about 50 bp. The copy number varies from six to nine. This repeat consists of two parts—constant GC-rich and variable AT-rich. | [11, 17, 26, 27, 28] | Endogenous retroviruses |
| 2. | F exon 1 | Repeat F is not purely a tandem repeat and is detectable in all species only as a 19-bp motif. Its copy number varies from two to five. This region in mouse carries the P2 promoter. | [15] | DNA transposons |
| 3. | B exon 1 | Repeat B is detectable only in primates; it is a small microsatellite tract, (C)5–8A | [15] | |
| 4. | C exon 1 | Repeat C is a block of tandem repeats with a monomer of 11 bp and copy number of 14 present only in mouse and rat. In other species including rather closely related voles, only one incomplete monomer is detected. | [15, 26, 27, 28] | Endogenous retroviruses |
| 5. | H exon 1 | Repeat H is a small block of tandem repeats with a monomer unit of 33 bp, which is detected in the consensus gene and is most pronounced in dog. The monomers in contemporary species have diverged beyond recognition. | L1 LINE |
| 6. | B exon 1 | Repeat B is also a microsatellite tract, (CCCCAG)n; it is present in all the species. | [15, 26, 27, 28] | Endogenous retroviruses |
| 7. | H exon 1 | Repeat H is a small block of tandem repeats with a monomer unit of 33 bp, which is detected in the consensus gene and is most pronounced in dog. The monomers in contemporary species have diverged beyond recognition. | L1 LINE |
| 8. | E exon h6/m7 | Repeat E is the most variable repeat present in all the species. Its monomer is 30–25 bp long, AT-rich, and not distinct. | [15, 26, 27, 28] | Intracisternal A particle (IAP) |

Figure 2. Comparison of the chicken Lnx3–Fip1l2 genomic region with the human XIST, 3'-flank and pseudo-TSX region. Homologous regions are indicated by arrows. Positions are shown in base pairs. Coloured boxes represent exons; yellow, pseudoexons; and blue, 5'-UTR. doi:10.1371/journal.pone.0002521.g002
weak similarity to the endogenous retroviruses is found between blocks of tandem repeats. The comparative analyses allow us to conclude that tandem repeats, which are the main part of exon 1, could indeed have originated from fragments of endogenous retroviruses and other mobile elements (Table 4, Fig. S4).

The ancestral Xist gene and its evolution

By compiling the Xist sequences representing four eutherian orders, we generated the Xist gene consensus and recognized a number of ancient elements which diverged considerably in the contemporary eutherian species (Fig.S1). We suggest that the consensus in some extent reflects the Xist structure that existed approximately 100 Mya in a common ancestor of Eutheria. We propose that the ancestral Xist gene contained ten exons (Table 1, Fig. 1). Part of exon 1, intron 3 and exons 5 to 7 are all remnants of Lnx3, whereas exons 2 to 4 and 8 to 10 originated from mobile elements of various classes (Fig. 3, Table 2). The promoter region of the ancestral gene evolved from exons 1 and 2 of the Lnx3 gene, flanked by an ancient L3 element at the 5' region (Fig. 3). It is likely that basic arrays of tandem repeats A, B, D, F, and H, which originated from endogenous retroviruses, were already present within exon 1 of the ancestral Xist gene as well as MIR3. It seems

| First sequence | Second sequence | PRSS P-value |
|----------------|----------------|--------------|
| G.gallus Lnx3 exon3 | H.sapiens Xist intron 3 | 2×10⁻⁵ |
| G.gallus Lnx3 exon5 | H.sapiens Xist intron 3 | 0.01 |
| G.gallus Lnx3 exon5 | M.musculus Xist gene exon 5 | 0.004 |
| G.gallus Lnx3 exons 1, 2 | H.sapiens Xist gene (ACC U80460) promoter | 0.007 |
| G.gallus Ras11C ex3 | H.sapiens Xic locus | 10⁻⁶ |
| G.gallus Wasf3 ex1 | H.sapiens Xic locus (FTX) | 10⁻⁵ |
| G.gallus Wasf3 ex2 | H.sapiens Xic locus (FTX) | 3×10⁻⁶ |
| G.gallus Uspl exon 1 | H.sapiens Xic locus (ENOX) | 0.006 |
| G.gallus Uspl exon 2 | H.sapiens Xic locus (ENOX) | 7×10⁻⁵ |
| G.gallus Uspl mRNA (ACC NM_001031123.1) | human ENOX EST BC071776 | 0.0007 |
| Consensus Xist exon 2 | L3 CR1 | 0.0001 |
| Consensus Xist exon 4 | L1 LINE | 3×10⁻⁵ |
| Consensus Xist exon 9 | ERV2 | 4×10⁻⁵ |
| Consensus Xist exon 10 | DNA transposons hAT-10_XT | 0.13 |
| Consensus Xist exon 10 | ZOMBI DNA | 2×10⁻⁵ |
| Monomer of consensus tandem repeat A | Endogenous Retrovirus ERV85 | 2×10⁻⁵ |
| Monomer of consensus tandem repeat B | versus Endogenous Retrovirus ERV18_MD_1 ERV1 | 0.0006 |
| 3-mer of consensus tandem repeat F | DNA transposones HAT2_MD hAT | 2×10⁻⁷ |
| Monomer of consensus tandem repeat H | L1 LINE | 0.001 |
| Monomer of consensus tandem repeat C | Endogenous Retrovirus ERV4_3-I_MM | 2×10⁻⁵ |

doi:10.1371/journal.pone.0002521.t004

Figure 3. The Xist gene evolved from a protein-coding gene and a set of transposable elements. Blue box – exons originating from Lnx3; red box – exons originating from transposable elements; dashed box – pseudoexons.
doi:10.1371/journal.pone.0002521.g003
likely that the ancestral Xist contained a single copy of the C repeat as it is present in a single copy in the majority of species analysed (vole, bovine, canine, human) and is tandemly repeated in mouse and rat only. The proportion of mobile elements in the ancestral Xist gene is relatively low, representing only 4.39% of the total gene length (Table S2). In the Xist gene of contemporary eutherian species the proportion of mobile elements is higher than in the consensus gene and varies from 7.9 to 15.9% (Table S2).

The promoter region of Xist is among the most conserved regions of the gene, showing a degree of homology between 82% and 98% for different species. Rodents are the exception, as the degree of homology for the promoter region is only 68% when compared with members of the other orders (Table S3).

Probably, exon 3 of the Xist consensus is retained in dog, chimpanzee and human but was deleted together with surrounding sequences in other species studied. The scenario for consensus exon 3 is consistent with evolutionary tree data that suggests that dog and primates diverged from the common ancestor earlier than primates and rodents [7]. The sequences corresponding to consensus exon 2 and 6 are retained as Xist exons in as evolutionarily distant taxa as rodent and insectivore [16], but are part of intronic sequences in other species. Conversely, consensus exon 9 which is retained in the majority of species is an intron in rodents (Fig. 1, Fig. 3).

During evolution the length of Xist RNA in different species changed dramatically. This occurred mainly due to formation of new exon–intron boundaries (for example, exon 2 of Xist in rat, exon 7 of Xist in mouse and vole), and variation in length of exon 1 by differential amplification of tandem repeats (Tables 1, 3, S1) and insertions of species-specific SINEs (Fig. 1, Table S2). However, despite almost twofold differences in exon size in various species, the total GC content in exons is conserved, varying from 41.4 to 42.6% (Table 5). The GC content in introns is more variable ranging from 37.2–41.9%.

The time of Xist gene emergence

To estimate the approximate time of Xist gene emergence, we aligned the nucleotide sequences of chicken and opossum Lnx3 genes (Ggl nx3 and Mdlnx3) with human and bovine Xist genes (Fig. S5). This alignment was constructed with preservation of the Lnx3 reading frame (all the deletions/insertions causing frameshift were discharged from the alignment). Xist is not a protein-coding gene; however, this alignment allows the number of potential synonymous and nonsynonymous substitutions to be assessed, as all the stop codons in the Xist gene were removed.

Figure 4 shows the phylogenetic trees constructed for the synonymous and nonsynonymous substitutions. As anticipated, the

---

Table 5. GC-content (%) in various elements of bovine, human, mouse and vole Xist genes.

| Species | 5'-flank | Exons | Introns | 3'-flank | Gene Xist |
|---------|----------|-------|---------|----------|-----------|
|         | size     | G+C%  | size     | G+C%  | size     | G+C%  | size     | G+C%  | size     | G+C%  | size     |
| B.t.    | 20000    | 41.3  | 109      | 53.2  | 24470    | 41.4  | 37.6     | 10464 | 40.0     | 16000 | 40.3     | 34934 |
| H.s.    | 18355    | 42.7  | 101      | 56.4  | 16974    | 42.1  | 37.2     | 15089 | 39.4     | 9582  | 39.8     | 32063 |
| M.m.    | 21461    | 42.1  | 104      | 54.8  | 15080    | 41.6  | 40.6     | 7706  | 45.3     | 15759 | 41.3     | 22786 |
| M.r.    | 11778    | 43.2  | 105      | 53.3  | 13357    | 42.4  | 41.9     | 7805  | 44.7     | 11452 | 42.2     | 21162 |
| R.n.    | 18165    | 41.5  | 101      | 54.0  | 15155    | 42.7  | 41.6     | 7743  | 44.3     | 14325 | 42.3     | 22898 |

Notes. Abbreviations: Ex, exon; In, intron; Pmin, minimal promoter; H.s., human; C.f., dog; M.r. vole Microtus rassiaemeridionalis; B.t., bovine; R.n., rat.

---

Figure 4. Tentative time frame of Xist origin. Minimum evolution trees constructed using (a) synonymous and (b) nonsynonymous substitutions (modified Nei-Gojobori method, p-distance) [24]. Arrow indicates an approximate time of Xist gene origin, assuming that this gene would remain a protein-coding gene and have a similar evolutionary rate as Mdlnx3 gene. Scale is shown nucleotide differences per site.

doi:10.1371/journal.pone.0002521.t005

doi:10.1371/journal.pone.0002521.g004
evolutionary rates of synonymous and nonsynonymous substitutions at positions within the Xist gene are approximately equal. For the synonymous substitutions (Fig. 4a), the lengths of branches approximately correspond to the anticipated time estimates, thereby indicating the uniformity of evolutionary rates over time. Nonsynonymous substitutions (Fig. 4b) give characteristic short branches leading to Ggnx3 and Mdlnx3 (as compared to synonymous substitutions). This results from the action of stabilizing selection, which is typical for the majority of protein-coding genes. As Xist is not a protein-coding gene, this limitation is removed. Therefore, the branch leading to Xist genes is very long as compared with the other branches. If Xist were a protein-coding gene and displayed approximately the same evolutionary rate as Mdlnx3, the length of these branches would be approximately equal (the suggested position of Xist gene is denoted with arrow; Fig. 4b). However, the rest of the length of the branch leading to the last common ancestor of bovine and human (B1x in Fig. 4b) approximately equals the lengths of the branches leading to human and bovine (B2x in Fig. 4b). This indicates that the Xist gene has an ancient origin. The long branch B1x might result from an acceleration in the evolution of the Lvnx3/Xist gene during the pseudogenization process. The fact that B1x is approximately 2 times longer (Table S4) in the Figure 4b (nonsynonymous sites) compared to B2 (Fig. 4a) can probably be attributed to the nonsynonymous sites being far from equilibrium with the mutational pattern at the moment of pseudogenization, such that for a while the number of substitutions per site would have been higher for these sites than for the synonymous positions (which would have been already closer to mutational equilibrium).

To test the reliability of the obtained results, we aligned nucleotide sequences of Ggnx3 and Mdlnx3 and all available mammalian Xist genes (Fig. S6). This alignment contained a much smaller number of sites (results not shown), however the results are similar to the results obtained for the four species alignment (Fig. S6). Similar estimates of branch lengths were obtained for different models of substitutions (Table S4) suggesting that the obtained results are reliable.

In general, a short alignment and a lack of good descriptors of the pseudogenization process prevent precise time estimates; however, taking into account that divergence of bovine and human is assessed as 90–110 Mya (branch B2x in Fig. 4b) [8], we can assume that the time of Xist gene emergence might be approximately 2 times older than this estimate (in the assumption that the evolutionary rate of nonsynonymous substitutions at the branch B1x is 2 times faster than B2x and B2, i.e., 180–220 Mya. This estimate exceeds the assumed divergence time of marsupials and eutherians (130–180 Mya). Taking into account the inevitable errors in the estimation of evolutionary rates and branch lengths in short alignments, the data obtained suggest that the inactivation of the Lvnx3 gene occurred during the divergence of marsupials and eutherians or in early eutherians.

The Origin of Ftx and Enox(Jpx) genes

Next we analysed the contribution of Usp1 and Wave4(Wasf3) genes to the formation of the XIC, as this question has remained unresolved in previous studies. We conducted a comprehensive search for similarity between the chicken protein-coding genes Wave4(Wasf3) (ACC XM_420299.1), Ras11c (ACC XM_420297.1), and Usp1 (ACC NM_001031123.1) and the genes of the mammalian XIC. A more detailed analysis with less stringent parameters in the search programs detected several regions homologous to the chicken Wave4(Wasf3) gene in the human XIC. Exons 1 and 2 of Wave4(Wasf3) correspond to exons 2 and 3 of FTX (ACC AK057701.1), displaying 58 and 53% similarity, respectively. The similarity is statistically significant in both cases (Table 4). Note that the exon-intron boundary between two homologous exons of Wave4(Wasf3) and FTX coincides (Fig. S2). Moreover, the considerably diverged regions homologous to Wave4(Wasf3) exons 3 and 4 are detectable at the same human loci (localized to 7q 5’ region). These data prompted us to search for homology to genes Usp1 and Ras11c.

The most pronounced homology was for exon 3 of Ras11c (63%), as described previously in [4]. The most likely explanation is that Enox(Jpx) is related to the chicken gene Usp1. As demonstrated previously, the main part of the Enox(Jpx) exons, except for exon 1, evolved from mobile elements of various classes [2]; moreover, different types of dispersed repeats gave rise to the exons of the gene in different mammalian species. When analyzing the human EST database, several alternatively spliced transcripts with common exons 1 and 2 were detected for the human Enox(Jpx) gene (data not shown). We found a statistically significant similarity (58%) between the putative Enox(Jpx) promoter region and exon 1 of chicken Usp1, between Enox(Jpx) exon 2 and Usp1 exon 5, and between Usp1 exon 7 and human EST BC071776, which is likely to be an alternatively spliced variant of the Enox(Jpx) gene (Table 4, Fig. S2). Thus, like Xist, Enox(Jpx) seems to have evolved from a protein-coding gene and a set of different transposable elements.

Discussion

The origin of the mammalian X-inactivation center

In this study we analysed the origin and evolution of the group of linked genes from which the mammalian XIC originated. The tight linkage of Lvnx3 (the precursor of Xist) with protein-coding genes Cdx4 and Chic1 (caudal type homeo box transcription factor 4; cysteine-rich hydrophobic domain 1), which flank Xist in eutherians, was detected in fish and, presumably, existed in the ancestor of vertebrates 450 Mya (Fig. 5). Unlike the evolutionary conserved syntenic group Cdx4, Chic1 and Lvnx3, the other genes that contributed to the sequence of the Xic, Ras11c, Usp1, and Wave4(Wasf3), are found within other linkage groups in fish (for details see Fig. 5 legends).

Tight linkage of all the precursor genes of the XIC and the flanking conserved protein-coding genes is also detectable in amphibians [4,10,11]. In pipid frog (Xenopus tropicalis), the genes Cdx4, Chic1, Fip122, Lvnx3, Ras11c, Usp1, Wave4 (Wasf3), Slc16a2 (solute carrier family 16), and Rnf12 (ring finger protein 12) are detectable within the same scaffold (Fig. 5). Note that in an ancestral amphibian, the Fip122 gene was inserted into the conserved syntenic group Cdx4-Chic1-Lvnx3, detectable in fish (Fig. 5). In chicken the order and orientation of the genes remains unchanged and they are integrated with genes which represent the main part of X conserved regions in mammals. Thus, it is likely that the formation of a proto XIC commenced about 450–560 Mya in ancestral amphibians.

An independent disruption of this linkage group took place in Prototheria and Metatheria [10,11,12]. It is interesting, that these genes map to different regions of an autosome, chromosome 6 [11] in monotremes, while in marsupials they map to the X chromosome [12]. This linkage group remained in the eutherian lineage, where the five protein-coding genes of the proto XIC underwent pseudogenization about 180 Mya to give different sequences of the X-inactivation center [4].

The origin of the genes in the mammalian XIC

Our comprehensive analysis together with previously published data [4] has enabled us to outline the origin of genes in the X-
inactivation center. Car2p is a retrotransposon protein-coding gene specific for eutherians [4]. Three exons of the protein-coding Tsx gene, which is expressed during spermatogenesis, originated from Fip12. The homology of Fip12 exon 17 is found in the region located 2 kb from the end of Xist exon 8. We have also found additional exons of this gene in the human sequence (Fig. 2). Two genes, Uspl and Wave4(Wav3), provided the basis for development of Exons(Jps) and Fixs, respectively. Both Exons(Jps) and Fixs, contain exons homologous to those in cognate protein-coding genes Uspl and Wave4(Wav3), respectively. Additionally Uspl contributed one exon to the promoter region of Exons(Jps). Presumably, Rass1lc was omitted from the evolutionary process and its remains are located between Xist and Exons(Jps). Analysis of Exons(Jps) and Fixs ESTs has revealed overlapping variants of their RNAs (BC345566 (ENOX), BM546361 and CN12172 (FTX)) similar to Xist and Tsx RNAs. It cannot be excluded that this is another pair of regulatory genes, however, their function and possible role in X inactivation is as yet untested. As for the origin of the Xist gene itself, the following scenario of ancestral gene formation seems the most probable. The basis was the Lnx3 gene, containing Pdz and RING finger domains. Lnx3 exons 4, 5, and 11 rise to the exons 5, 6 and 7 of the Xist consensus gene (Fig. 2). The promoter region originated from the 5'-UTR of Lnx3 exons 1 and 2. The remaining six Xist exons including those with simple tandem repeats detectable in their structure have similarity to different transposable elements and thus the Xist gene has originated from pseudogene Lnx3 and a set of various transposons. These mobile elements were integrated into the locus and subsequently their fragments were incorporated into the RNA.

Simple tandem repeats A, B, C, D, E, and F contributed essentially to the formation of exons in the ancestral gene; their monomer units presumably originated from endogenous retroviruses as weak similarity was observed in blocks of tandem repeats (Fig. 1). There are some cases of coding sequences entirely or partially derived from ancient transposable elements [13], and it has been reported that over 4.4% of Unigene transcripts contain significant matches to transposable elements within their coding regions [14]. Xist can be classified as an intermediate pseudogene, and represents a new type of ncRNA gene. Our data adds to what is already known regarding Xist evolution by showing that insertion of transposable elements within the ancestral protein-coding gene contributed to the creation of a novel ncRNA gene. In fact, other regions of the XIC have evolved similarly by transposable element insertion, mutating active protein-coding genes into non-coding pseudogenes. Our results constitute important findings into the origin of the unique Xist gene and suggest that acquisition of transposable elements by a protein-coding gene may be a more widespread phenomenon contributing to ncRNA gene origin and evolution than previously thought.

Species-specific mode of Xist gene evolution

Comparative analysis across different species shows that eutherian Xist evolved in a species-specific manner. As reported previously the unique sequence of the Xist gene is not conserved and evolves very rapidly [2,15]. The exon-intron structure of Xist is also not strongly conserved. The interspecific differences in the unique sequence, length and structure of exons suggest that the length of the RNA (and, consequently, the sequences, like mobile elements and tandem repeats, contributing to the RNA size in different species) is either non-essential for function (neutral sequences) or these sequences become selectively adapted in a specific manner to the conditions of functioning in the genome and the X chromosome of particular species. Certain core sequences common for all species are essential for the regulation of gene activity and its function.

The functional role of introns in the structure of the XIST/Xist gene is not well understood. The introns in protein-coding genes can separate structural or functional protein domains; however, it is unknown what their roles are in regulatory genes producing nuclear RNAs and how they differ from exons, as neither are protein-coding. Moreover, there are some cases in the evolution of the Xist gene when exons convert into introns and vice versa. We named the exons that are inactivated in some species but preserved in introns as pseudoxons (pExs). These pseudoxons, which are active in some species and inactivated in others, have been identified (Table 1, fig. 1).
Alongside the formation of pseudoexons, which arise due to mutation to intron-exon sites (ph2, man; pEx2 - chimpanzee), we observed exons shortening at their 3' ends due to the formation of new 3'-5' exon/intron junctions, for example m7 (mouse) and vole exon 2 [15]. A similar phenomenon was reported previously for Xist exon 4 in mole, dog and cow [16].

Thus, the exon-intron structure of the XIST/Xist gene is not stringently conserved or evolutionarily stable. It displays lability and exon–intron transitions, presumably connected with species-specific patterns related to its function.

A characteristic of the Xist gene is the presence of both ancient and young species-specific mobile elements in the processed transcripts. A similar feature is observed in other XIC genes, Enox(gp4), which also produce nuclear RNA. The processed transcript of the Enox(gp4) gene includes SINEs, LINEs, and even processed pseudogenes [9]. Presumably, the integration of mobile elements into exons accompanies the overall evolution of these genes and continues in contemporary Eutheria. Possibly, this is a general mechanism involved in the formation and rearrangement of genes encoding large nuclear regulatory RNAs.

Despite the drastic evolutionary rearrangements of the Xist gene structure in different species, the 42% GC content in exons is a preserved characteristic of this gene in mammals and, presumably, is maintained by selection. The introns contain less GC pairs compared to the exons and display a considerable variation between orders. The total GC content in Xist gene is similar to the overall XIC region and varies in the range of 39.8–42.2% (Tables 4). The region of the minimal promoter differs drastically—it is enriched with GC nucleotides (52.4–56.4%). The 3'-flanking region of rodents is richer in its GC content compared with both the gene itself and its 5'-flanking region. In contrast, the 3'-flanking region in human and bovine displays a lower GC content compared with the gene; in addition, a decrease in the GC content is evident from the 5' - to the 3'-flanking region (Tables 4). On the other hand, the localization and structure of Xist CpG islands in the species compared are different. The detected CpG islands in human and bovine are mainly associated with SINE and LINE repeats. The absence of conservation in the localization of CpG islands and their divergence in the case of similar localization in the Xist gene of different species may indicate that species-specific mechanisms are involved in the regulation of its activity.

Thus, we have proposed a mechanism whereby the Xist gene may have originated. We suggest that the Xist gene lost the function of the protein-coding gene Lnx3 and no longer contains any extended ORFs. However, due to transposon insertions and their partial subsequent amplification, new functional domains formed (for example, repeat A in exon 1 [17]). These domains then became necessary for the transcriptional silencing of X chromosome genes. We suggest that this example of how a protein-coding gene loses its protein-coding function by mutation and then gains a new function due to transposon integration is not an exceptional case, but is a more wide-spread phenomenon applying to other non-coding RNA genes and pseudogenes with new functions.

Materials and Methods

The following sequences extracted from the corresponding databases of sequenced genomes at the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu/) were used: human Mar. 2006 (hg18) assembly range = chrX:72494111-74160153; chimpanzee Mar 2006 (panTro2) assembly range = chrX:75160560-77392648; dog May 2005 (canFam2) assembly canFam1_dna range = chrX:60100000-60735000; mouse Feb 2006 (mm8) assembly range = chrX:99501937-100404904; rat June 2003 (rn3) assembly range = chrX:91358074-91899712; chicken Feb. 2004 (galGal2) assembly range = chr:11020349-12295219; opossum assembly MonDom 4.0. January 2006 from Ensembl (http://www.ensembl.org/index.html); vole Microtus rossiaemerdionalis ACC AJ310130; and bovine ACC AJ421481.

Computer analysis was conducted using modern versions of the following software packages: BLAST ([10], http://www.ncbi.nlm.nih.gov/) for searching for homologous sequences; Tandem Repeat Finder 4 [19] for searching for tandem repeats; RepeatMasker [20] for searching for mobile elements; Fasta [21], CLUSTALX [6] for aligning two and more sequences (the programs and data are available at http://genome.ucsc.edu/, http://www.ensembl.org/, and http://bio.cse.psu.edu/) and (PipeMaker, [5]) for genomic analysis of extended loci.

Comparative analysis of the Xist gene of ten eutherian species belonging to four orders (Rodentia, Cetartiodactyla, Carnivora, and Primates) was conducted. The structure of the Xist gene in four closely related vole species was determined earlier [15]. As vole Xist genes are very similar, the sequence of M. rossiaemerdionalis was used in the present study. We identified chimpanseee, dog, and rat orthologs of the Xist gene in the data available from genomic sequencing projects of the three species by pairwise alignments with the known human, mouse, and bovine sequences [2]. The exon–intron structures of dog, chimpanzee, and bovine Xist genes were determined by comparing with the human and mouse genes. Comparison with genomic sequences demonstrated that the exon–intron boundaries in bovine and chimpanzee genes coincide with numerous spliced ESTs. In dog only one EST is located in the region of Xist gene. Comparison with the genomic sequence demonstrated that it contained exons 2, 4, 5, and beginning of exon 6 (according to human XIST gene). This data cannot be regarded as final, because other species display various rare variants of EST splicing, including the loss of certain exons. The Xist sequences of different mammalian species were analyzed for the presence of both tandem and dispersed repeats using the TRF, RepeatMasker and the repeat database RepBase12.03. (http://www.girinst.org/repbase/index.html) [22]. To search for the homology between considerably divergent and evolutionarily distant sequences (for example, human–chicken), we used the programs WUBLAST with the parameter W = 3 and FASTA34 with the algorithm Smith–Waterman or SSEARCH34 and the parameters f = -14, g = -2. The similarity between Xist exons and mobile elements was searched for using the same programs and parameters. Each individual exon was compared with the database of chordate mobile elements, which was obtained by consolidating the human, primate, Mammalia, Rodentia, and Vertebrata databases (http://www.girinst.org/repbase/index.html). CpG islands of at least 200 bp and with a GC percentage exceeding 50% and with an observed/expected CpG ratio that is greater than 60 bp were detected using the FIND-CPG program from the PipTools [23]. To test the statistical significance of the obtained alignments the standard PRSS program from the FASTA software package was used (Table 4). As E-value and z-score estimated by the WUBLAST, FASTA and SSEARCH programs depend on sequence length and the size of search database, we assessed the significance of alignments by P-value calculated in the PRSS program from FASTA packed and selected the most significant alignments.

The modified Nei-Gojobori method as implemented in the MEGA program [24] and a maximum likelihood method as implemented in the CODEML program [23] were used to estimate the numbers of synonymous and nonsynonymous substitutions per site. Phylogenetic trees based on multiple alignments of nucleotide sequences were constructed using the
minimum evolution method as implemented in the MEGA program with default parameters.

Supporting Information

Figure S1 Nucleotide sequences of eutherian consensus gene Xist presented in GENE BANK format. Found at: doi:10.1371/journal.pone.0002521.s001 (0.10 MB DOC)

Figure S2 Alignment of the chicken proto XIC region with the human and mouse XICs. (a) Sequence alignment of part of human Xist intron 3 and a part of Lnx3 exon 3 (ACX XM_420296.1). (b) Alignment of part of human Xist intron 4 and part of Lnx3 exon 5. (c) Alignment of part of mouse Xist exon 5 and chicken Lnx3 exon 5. (d) Alignment of part of the human Xist promoter and Lnx3 exons 1 and 2. (e) Alignment of chicken Rasi11c exon 3 and the 5’-region of Xist. (f) Alignment of chicken Was3 exon 1, 2 and the 5’-region of Xist. (g) Alignment of chicken Uspl exon 1, 2 and the 5’-region of Xist. (h) Alignment of chicken Uspl mRNA (ACC NM_001031123.1) and human ENOX EST BC071776. (a-g) computed with SSEARCH34 from FASTA package; (h) computed with WUBLAST 2.0.

Found at: doi:10.1371/journal.pone.0002521.s002 (0.05 MB DOC)

Figure S3 Homology of the mammalian consensus Xist gene exons 2, 4, 9, 10 and various transposable elements.

Found at: doi:10.1371/journal.pone.0002521.s003 (0.03 MB DOC)

Figure S4 Homology of consensus monomer sequences of the main blocks of tandem repeats in Xist gene and various transposable elements

Found at: doi:10.1371/journal.pone.0002521.s004 (0.03 MB DOC)

Figure S5 Nucleotide sequences alignment of cDNA of chicken and opossum Lnx3 genes (Mlnx3 and Clnx3) and human and bovine Xist genes.

Found at: doi:10.1371/journal.pone.0002521.s005 (0.04 MB DOC)

Figure S6 Minimum evolution trees constructed using (a) synonymous and (b) nonsynonymous substitutions (modified Nei-Gojobori method, p-distance) [18]. The numbers for the interior branches refer to the bootstrap values with 1,000 pseudoreplicates.

Found at: doi:10.1371/journal.pone.0002521.s006 (0.03 MB DOC)

Table S1 Found at: doi:10.1371/journal.pone.0002521.s007 (0.04 MB DOC)

Table S2 Found at: doi:10.1371/journal.pone.0002521.s008 (0.04 MB DOC)

Table S3 Found at: doi:10.1371/journal.pone.0002521.s009 (0.03 MB DOC)

Table S4 Found at: doi:10.1371/journal.pone.0002521.s010 (0.03 MB DOC)

Table S5 Found at: doi:10.1371/journal.pone.0002521.s011 (0.06 MB DOC)

Acknowledgments

We would like to thank all members of the Developmental Epigenetics laboratory for helpful discussions and Claire E. Senner for valuable comments during preparation of this manuscript.

Author Contributions

Conceived and designed the experiments: NB SZ TN EE NK AS. Performed the experiments: EE. Analyzed the data: EE NK AS IR. Wrote the paper: NB SZ TN EE NK AS IR.
26. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, et al. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71: 515–526.

27. Brown CJ, Baldry SE (1996) Evidence that heteronuclear proteins interact with XIST RNA in vitro. Somat Cell Mol Genet 22: 403–417.

28. Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, et al. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71: 527–542.