SUPPLEMENTARY INFORMATION

Independent specialization of the human and mouse X chromosomes for the male germline

Jacob L. Mueller[1], Helen Skaletsky[1, 2], Laura G. Brown[1, 2], Sara Zaghlul[1], Susan Rock[4], Tina Graves[4], Katherine Auger[5], Wesley C. Warren[4], Richard K. Wilson[4], David C. Page[1, 2, 3]

[1]Whitehead Institute, Cambridge, Massachusetts, USA

[2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[4]The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA.

[5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

Table of Contents

Supplementary Table 5. Tallies of X-linked gene classifications depicted graphically in Figure 3

Supplementary Table 6. Tissue expression of X-linked genes shared between human and mouse

Supplementary Table 7. Expression patterns of independently acquired human X-linked genes

Supplementary Table 8. Expression patterns of independently acquired mouse X-linked genes

Supplementary Table 9. Documenting expression of individual members of independently acquired X-linked multicopy and ampliconic gene families in human and mouse

Supplementary Table 10. Expression patterns of all human autosomal genes

Supplementary Table 11. Expression patterns of all mouse autosomal genes

Supplementary Table 12. All OMIM X-linked phenotypes where the molecular basis is known

Supplementary Table 13. Expression patterns between the sexes, in the heart and kidney, of human genes that follow Ohno's law

Supplementary Table 14. Expression patterns between the sexes, in the heart and kidney, of mouse genes that follow Ohno's law

Supplementary Note

A. Problems with multi-haplotype assemblies of ampliconic regions

Here we highlight specific problems with the previous multi-haplotype assembly of the ampliconic region in Figure 2 and the steps taken to address these problems and reassemble the region via a SHIMS approach using full-length RP11 BAC sequences.

1. **Problem**: While the three RP11 BACs in the previous multi-haplotype assembly all derive from the same individual and therefore the same haplotype, the CTD and RP13 BACs derive from different individuals and therefore likely represent different haplotypes. **Solution**: We limited the SHIMS assembly of this region to RP11 BACs so as to eliminate the confounding effects of polymorphic differences between haplotypes.

2. **Problem**: BAC RP11-472D17 contains a large palindrome, with a nonduplicated "spacer" at the center of the palindrome. In the previous sequence assembly of BAC RP11-472D17, this spacer sequence was inverted, which contributed to the appearance of multiple palindromes in the previous multi-haplotype assembly. In the previous sequence assembly, RP11-472D17 is depicted as overlapping only modestly with RP11-485B17, and not at all with RP11-204I15, while in fact the three BACs overlap extensively. The unintended inversion of the central, non-duplicated spacer of a palindrome is a common problem in the assembly of spacer-spanning BACs. **Solution**: Correcting the orientation of the spacer sequence in RP11-472D17 revealed that RP11-472D17 overlapped extensively with BACs RP11-485B17 and RP11-204I15.

3. **Problem**: In the previous assembly, there were six miscalled nucleotides in the finished sequence of BAC RP11-485B17 (GenBank Accession # BX088602.6), which led to the erroneous conclusion that it did not overlap BACs RP11-472D17 and RP11-204I15. **Solution**: We generated full-length finished sequence of RP11-485B17 (GenBank Accession # BX088602.7) and found that it extensively overlapped RP11-472D17 and RP11-204I15.

4. **Problem**: In generating the previous assembly, it was standard procedure to truncate the finishing of each BAC sequence so as to provide 2-kb finished overlaps with neighboring BACs. While these 2-kb overlaps may suffice in assembling non-ampliconic regions, they provide insufficient information in ampliconic regions, where the investigator must distinguish one copy of an amplicon from another. It is our general practice in ampliconic regions to require overlaps of at least 20 kb, and we generally finish the entirety of each BAC. **Solution**: We generated finished sequence for the entirety of five BACs (all RP11) from the region to ensure that the BACs truly overlap. As with RP11-485B17 (see above), we generated full-length finished sequence for RP11-472D17 (GenBank Accession # BX293536.5) and RP11-204I15 (GenBank Accession # BX510359.5), replacing their respective truncated sequences (GenBank Accession #'s BX293536.4 and BX510359.4, respectively) on which the previous assembly of the region had been built. In addition, we generated full-length finished sequence for BACs RP11-651H2 and RP11-319K11, which had not been sequenced previously.

With accurate assemblies of six RP11-BACs from the region, we found that BACs previously depicted as spread across multiple palindromes, and roughly 500 kb, in fact collapsed over a single palindrome, and half the distance (Figure 2b). This highlights the importance of generating high-quality finished sequence in order to disentangle the complex nature of ampliconic regions. Indeed, the flipping of a palindrome spacer, or a handful of miscalled nucleotides, can generate sequence assembly artifacts not representative of any X chromosome.

B. Limitations of using whole genome shotgun sequence to infer the evolutionary history of a gene

For genes not shared between the human and mouse X chromosomes we used the genome sequences of the dog, horse and chicken as outgroups to infer whether a gene was lost, or gained via a lineage-specific duplication, or independently acquired. We chose these three species because of all outgroups to humans and mice, they have the deepest level of sequence coverage and independently generated dense genetic linkage maps were used to guide their assemblies. That being said, it is important to note that there are limitations in using these genome sequences and that caution should be used when inferring their evolutionary history. The genomes of dog, horse and chicken are not assembled to the same level of precision as the human and mouse X chromosome assemblies. Errors in these genomes could include omission of a gene, misassignment of a gene to a different location in the genome, or collapsing of multiple copies of a gene family into a single copy. Such miassemblies or misassignments could be present in any of these three genome's assemblies and confound our interpretations of a given gene's evolutionary history. Ampliconic sequences are particularly prone to misassembly in whole genome shotgun assemblies (She, X. *et al.* 2004), which was the primary sequencing strategy to assemble the dog, horse and chicken genomes.

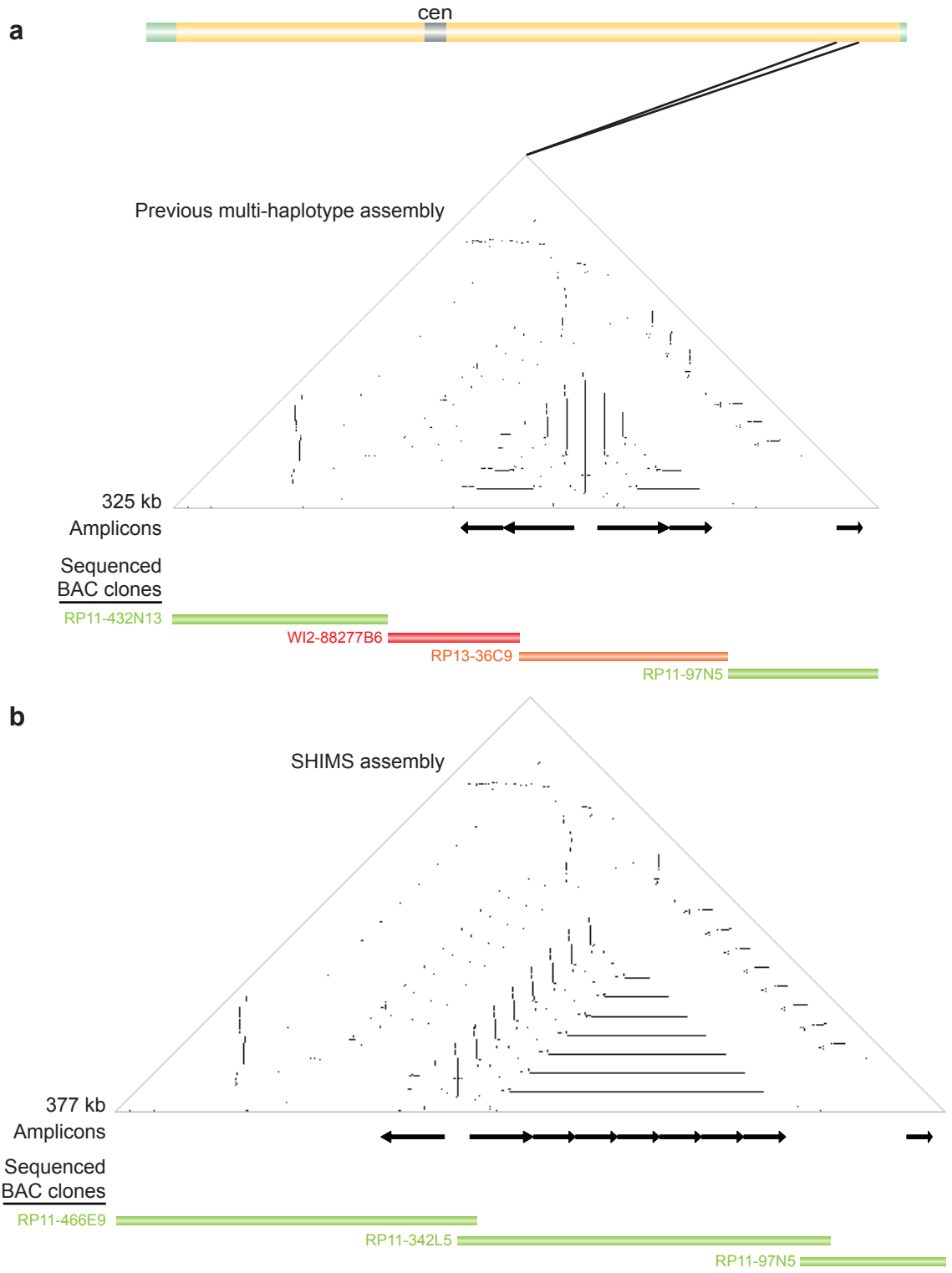
Examples of concerns of inferring evolutionary history are provided below and can be found in Supplementary Tables 3 and 4.

1. There are some cases where a gene is not detectable in syntenic regions of the chicken genome, but is present in syntenic regions of the dog and horse X chromosomes. This could be due to the given gene being added after mammals diverged from birds and then subsequently lost in humans or mice; a two-step evolutionary process. Alternatively, misassembly of the chicken genome could also account for the difference in assignment. In such cases, we assigned the gene as being lost or X-linked lineage-specific duplication, instead of independently acquired.

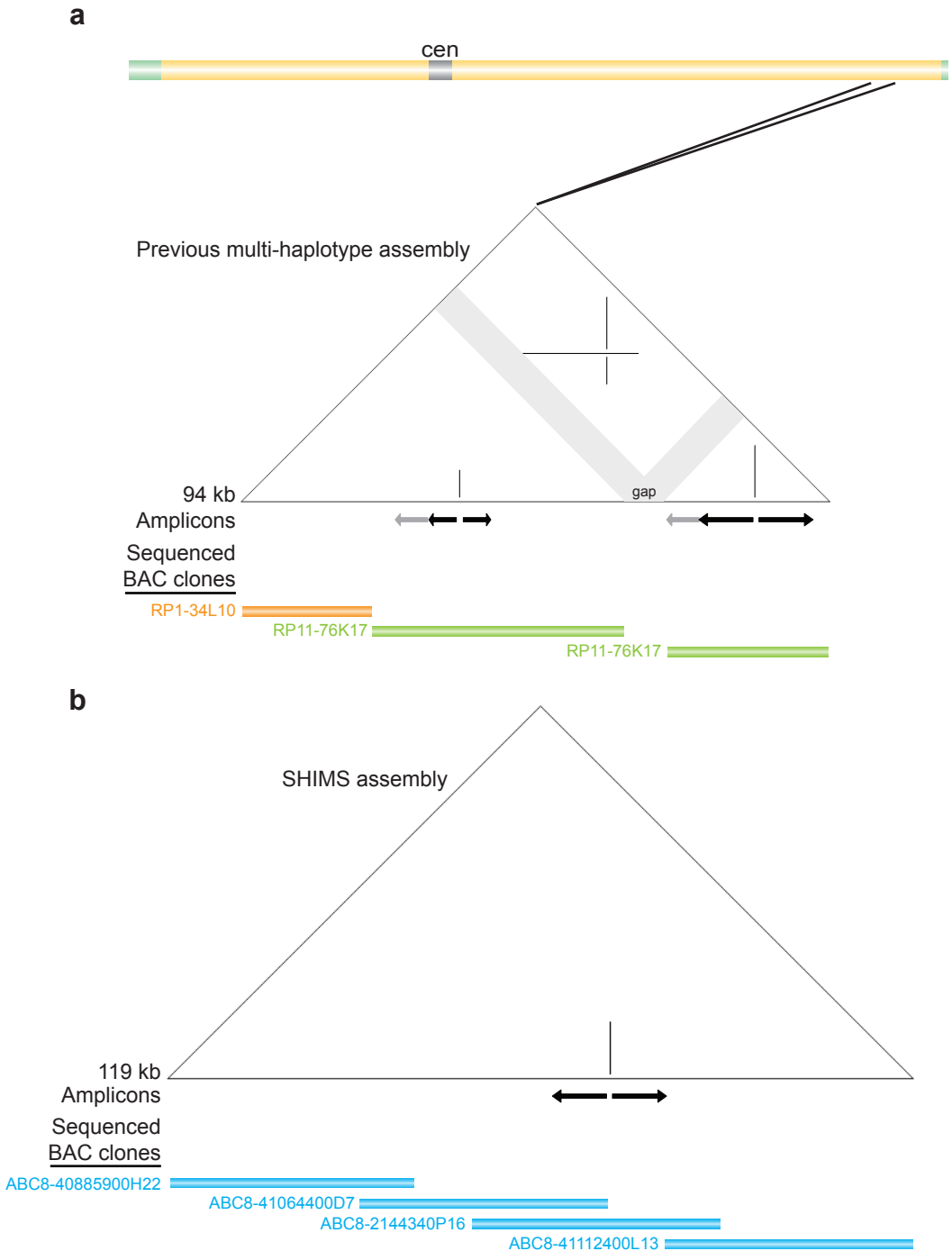
2. Some genes are not detectable in syntenic regions of the chicken genome, but detectable on the X chromosome of either the dog or horse and detectable on the X chromosome of either human or mouse. *RHOXF1* is an example of a human X-linked gene that is not detectable in chicken and horse, but is detectable in dog. *RHOXF1* could have been added to the X chromosome after mammals diverged from birds and then lost on the dog lineage and also lost on the mouse lineage. Alternatively, it could have been added to the X chromosome after mammals diverged from birds and due to miassemblies in the dog genome sequence is missing. We considered such cases as lost in the mouse lineage.

3. *FAM156* is ampliconic in human and only a single copy present in mouse. The second copy is not present in dog. Horse has two copies but they are very diverged and it is not clear that they neighbor each other, as in the case of the human *FAM156* ampliconic genes. *FAM156* is thus considered a X-linked human lineage duplication.

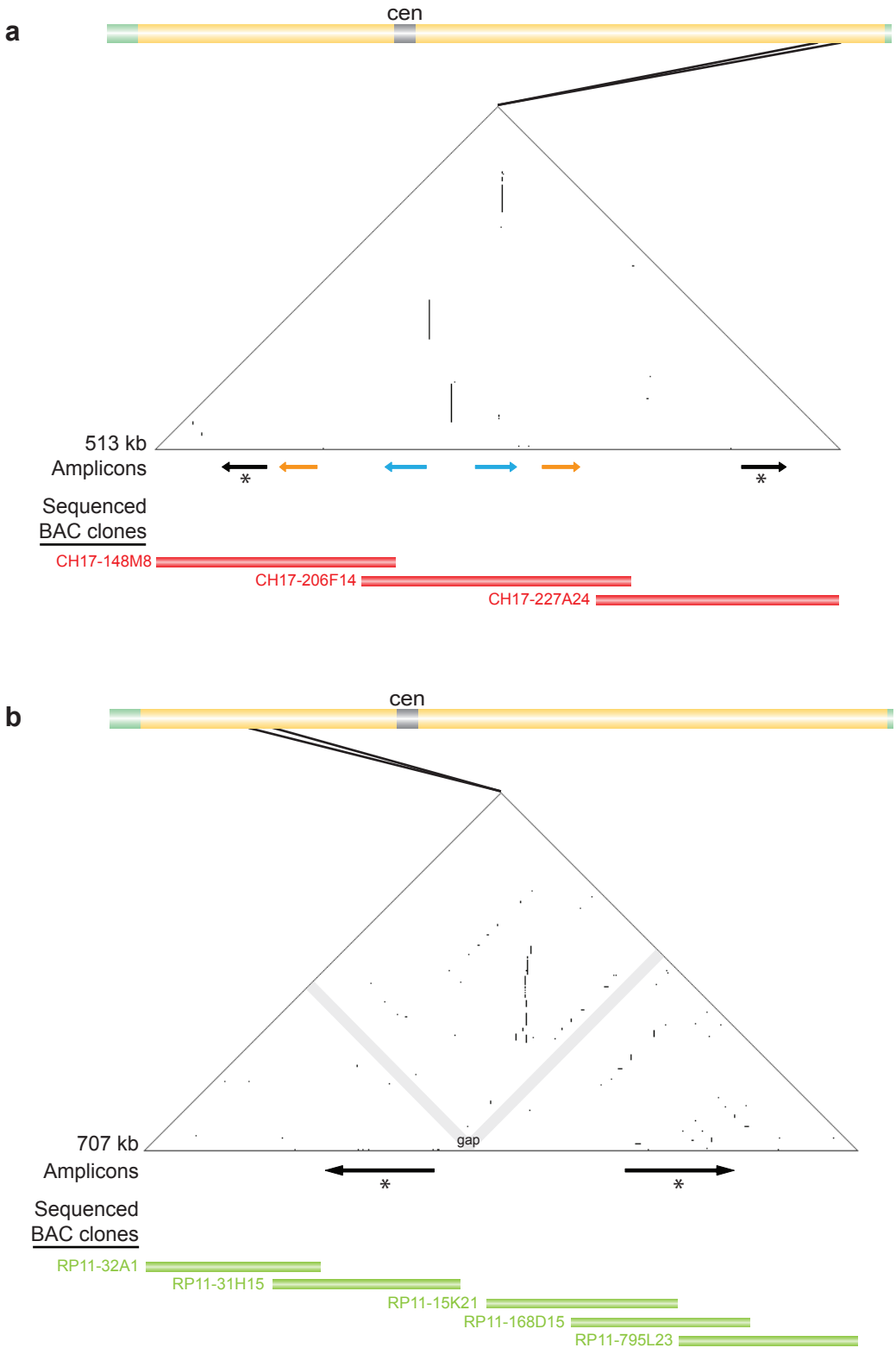
4. The *CT45* gene family represents a case where it is unclear whether the gene family was amplified multiple times or was lost multiple times. *CT45* is ampliconic in humans and has multiple neighboring copies in syntenic regions of the horse X chromosome. Syntenic regions of the dog and mouse X chromosomes do not have amplified copies of *CT45*. Thus, *CT45* could have been lost independently in the dog and mouse lineages or could have been independently amplified in the human and horse lineages. In such cases, we have chosen the more conservative approach and indicate that the *CT45* copies were lost in mice. *CT45* highlights how particular caution should be used when inferring the evolutionary history of ampliconic genes in outgroups assembled via whole genome shotgun sequence.



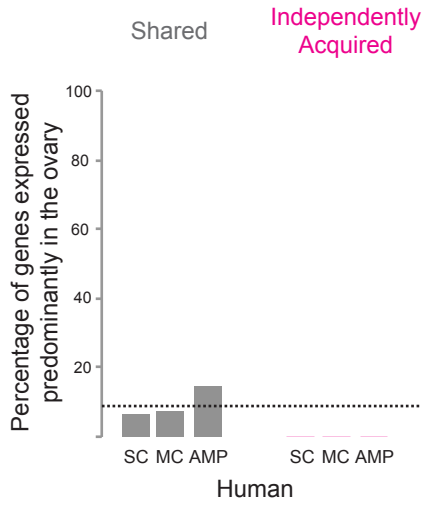
Supplementary Figure 1 Triangular dot-plots of a revised human X-ampliconic region containing the *CT45* gene family. **a**, The previous multi-haplotype reference sequence. Each dot represents 100% nucleotide identity within a window of 100 nucleotides; direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline. Black arrows immediately below plots denote positions and orientations of amplicons. Further below, sequenced clones from RP-11, WI-2, RP-13 libraries (each from a different individual) contributing to the assembly are depicted as green, red, and orange bars, respectively; each bar reflects the extent and position within the assembly of finished sequence for that clone. (As per the human genome assembly standard, finished-sequence overlaps between adjoining clones are limited to 2 kb.) Genbank accession numbers are in Supplementary Table 1. **b**, SHIMS assembly of same region. All BACs derive from RP-11 library (one male) and are fully sequenced; each BAC's finished sequence extensively overlaps those of adjoining BACs.



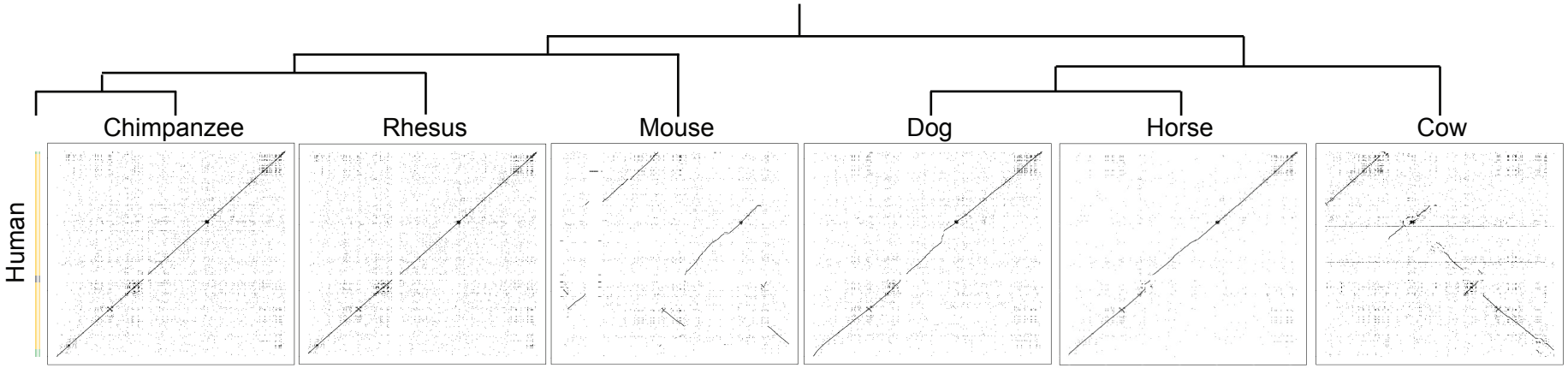
Supplementary Figure 2 Triangular dot-plots of a revised human X-ampliconic region containing the *PNMA6* gene family. **a**, The previous multi-haplotype reference sequence. Each dot represents 100% nucleotide identity within a window of 100 nucleotides; direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline; gaps are indicated by gray shading. Black arrows immediately below plots denote positions and orientations of amplicons. Further below, sequenced clones from RP-1 and RP-11 libraries (each from a different individual) contributing to the assembly are depicted as orange and green bars, respectively; each bar reflects the extent and position within the assembly of finished sequence for that clone. (As per the human genome assembly standard, finished-sequence overlaps between adjoining clones are limited to 2 kb.) Genbank accession numbers are in Supplementary Table 1. **b**, SHIMS assembly of same region. All clones derive from ABC-8 library (one male) and are fully sequenced and in blue; each clone's finished sequence extensively overlaps those of adjoining clones.



Supplementary Figure 3 Triangular dot-plots, based on SHIMS assemblies, of newly identified palindromic amplicons. **a**, An ampliconic region within chromosomal interval Xq28. Each dot represents 100% nucleotide identity within a window of 100 nucleotides. Direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline; gaps are indicated by gray shading. Black arrows immediately below plots denote positions and orientations of amplicons. Newly identified amplicons are indicated with asterisks. Further below, sequenced clones from the CH-17 (haploid genome) library contributing to the assembly are depicted as red bars. Genbank accession numbers are in Supplementary Table 1. **b**, An ampliconic region within chromosomal interval Xp21.1. Each dot represents 100% nucleotide identity within a window of 50 nucleotides. Sequenced clones from the RP-11 (one male) library contributing to the assembly are depicted as green bars.

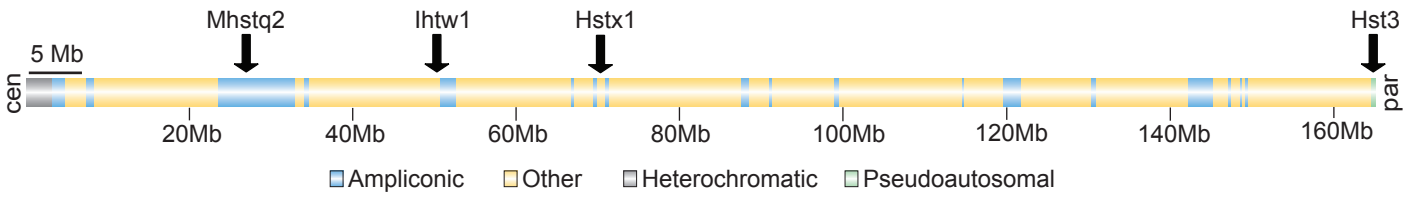


Supplementary Figure 4 Percentage of genes expressed predominantly in the ovary. The horizontal dotted line represents the proportion of genes on all autosomes exhibiting ovary-predominant expression. SC = single copy, MC = multicopy, AMP = ampliconic. The frequencies of shared MC, shared AMP, and independently acquired SC, MC, AMP genes are not statistically different (P-value >0.05) from either autosomal or X-linked single-copy gene frequencies (Chi-square test, with Yates' correction).



Supplementary Figure 5 Dot plots of human X chromosome versus six different mammalian X chromosomes. Dot plots were performed using BLASTZ nucleotide alignments, where each dot represents a high scoring stretch of nucleotide sequence with >70% identity (see methods). Chimpanzee, rhesus, dog, horse and cow X chromosome whole genome shotgun assemblies are comprised almost exclusively of single-copy sequence (Table 1), making it unclear if human and mouse X-ampliconic sequences are conserved in these species. We can see that the human X chromosome single-copy sequence is shared with other mammals, but it is unclear whether its amplicons are until SHIMS-based sequence is generated for the chimpanzee, rhesus, dog, horse, and cow X chromosomes. The phylogenetic tree is based upon a recent comparison of 29 mammalian genomes (Lindblad-Toh K. *et al*. 2011).

Reference:
Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476-82 (2011).



Supplementary Figure 6 Hybrid male-sterility loci in mice that map to the X chromosome. Based upon the mouse genome informatics phenotype database, three loci (Mhstq2, Ihtw1, Hstx1) are X-linked and map within or adjacent to independently acquired genes. Mhstq2 (male hybrid sterility QTL 2; Elliott et al., 2004) is genetically linked to chromosomal position 29.5 Mb -- within the *Slx* amplicon -- and is associated with low sperm production in hybrids. Ihtw1 (interspecific hybrid testis weight 1; Elliott et al., 2001) is genetically linked to chromosomal position 50.3 Mb (the map position of the DXMit23 marker) -- 500 kb proximal to the *Slx-like-1* amplicon -- and is associated with reduced levels of male fertility and reduced testis weight in hybrids. Hstx1 (hybrid sterility, X chromosome 1; Storchova et al., 2004) is genetically linked to chromosomal position 69.5 Mb (the map position of the DXMit119 marker) -- flanked by the *4930567H17Rik*-amplicon (67.6 Mb) and the *Xlr*-amplicon (70.5 Mb) -- and is associated with reduced levels of male fertility, reduced testis weight, reduced sperm count and increased abnormal sperm head morphology in hybrids. A fourth locus, Hst3, maps to the X chromosome pseudoautosomal region (par) and is considered to be due to differences in the PAR between *Mus spretus* and *Mus musculus*. The only other known hybrid sterility loci in the mouse genome map to chromosome 17. Of the six loci that map to chromosome 17, five loci (Hst4, Hst5, Hst6, Hst7, and Mhstq1) all map within close linkage at the proximal end of chromosome 17, near the t-complex, and the remaining locus (Hst1) maps to *Prdm9*.

Supplementary Table 1. Reassembling the human X chromosome: 33 regions scrutinized, 29 of which we sequenced using the SHIMS approach

Region #	Putative ampliconic region[a]	Region selected because of:	Insights gained from single haplotype sequence	GenBank accession numbers for SHIMS assemblies	GenBank accession numbers for BACs or fosmids spanning the region (ordered 5' to 3')[b]
1	chrX:36935232-37109891	Gap	Newly identified palindrome	JH720451, JH720452	**AC233304, AL606516, AC233310, AC233287, AL592156**, BX469939, BX842585, BX842588, AC006924, BX842568**, AC243371,** gap**, AC243975, AC245096, AC233283, AC233297, AC233292**
2	chrX:45392748-45475100	Misoriented fosmid ends		JH806589	AL034412, **AC234772**, AL031584
3	chrX:46653968-46756755	Misoriented fosmid ends		KB021648	**AL627143**
4	chrX:48087689-48177927	Amplicons		JH806590	**AC244636, AC245047, AL606490**
5	chrX:48863143-48949325	Amplicons		JH806590	**AC231533, AC233300, AC233294, AC231657**
6	chrX:48962948-49331797	Gap and amplicons	Gap due to tandem array with ~25 copies (data not shown)	JH806590	**AC232271, AC233302**, tandem array represented by flanking BACs, AC142497, **AC231643, AC231644**
7	chrX:49619684-50357383	Gap	Gap due to unclonable intervening sequence	JH806590, JH806587	**AC236430, AC231645, AC235789, AC243516,** gap**, AC239396,** AL357894**, AC233286,** AL121865, AL445491, AL359272**, AL359914, AC239367,** AL954833, AL391379**, AC233976,** AL158055, AL390060
8	chrX:51412207-51509602	Amplicons		JH806587	**AC234030**
9	chrX:51792300-51983269	Amplicons		JH806587	**AC241520,** AL929410**, AC239585, AC245177,** AL928717
10	chrX:52534857-53044111	Gap and amplicons	Closed gap and reassembled ampliconic region	JH806587	**AC231759, AC231532, BX088602, BX510359, AL450023, AC244505, AC234031, AL807736,** BX322635, AL591212, AL139396**, AC245102, AC233728, AC231658, BX323845, AC233279**
11	chrX:55480842-55590897	Amplicons			AL590410[c]
12	chrX:62252458-62412075	Amplicons		JH806591	**AL359854, AC246786, AC234780, AC158203**
13	chrX:70810842-70972407	Amplicons			BX276092[c]
14	chrX:71857884-72241800	Amplicons		JH806592	**AC240504, AC234776, AL662864**
15	chrX:76449373-76713483	Gap	Gap due to unclonable intervening sequence	JH720453, JH720454	**AC233982, AC233281, AC233981,** AL442646**, AL590789, AC233284, AC239601, AC233296, AC233277, AC240392,** AC017089**, AC233301, AC234032, AC243977,** gap**, AC243316,** BX510371**, AC233305, AC233293, AL138743**
16	chrX:100705379-100790633	Amplicons		JH806593	**AC234775**
17	chrX:101322434-101661047	Amplicons		JH806594	**AC234791, AC235565**
18	chrX:103081761-103248997	Amplicons		JH806595	**AC234782, AC234783**
19	chrX:105369795-105457848	Misoriented fosmid ends			**AL133271**
20	chrX:113234060-113655429	Gap	Gap due to unclonable intervening sequence	JH806588, JH806601	**AL442070, AC233289, AC233295, AL591842[d], AC236668, AC233303, AC233285,** BX510661**,** FP565586 gap, AC243413, AL953862, AL589677, AC233299, AL590097, AL355812, AL121878, AL445164**, AL589786, AC239923, AC239600,** AC003983**, AL589842,** AC005000
21	chrX:114862471-114932208	Amplicons			CR753863[c]
22	chrX:115410626-115831995	Gap	Gap due to unclonable intervening sequence	JH806602, JH806603	BX546444, BX323838**, AL772226, AL732602, AC233725, AC233291, AC241580, AC243535,** gap**, BX530410, BX510313, BX119904, BX284115, AL732586, AL732604, AL732637**
23	chrX:119024827-119237455	Amplicons	Reassembled ampliconic sequence	JH806596	**AC240732, AC240549**
24	chrX:119882751-119954883	Amplicons			AL670379[c]
25	chrX:134106841-134230880	Amplicons		JH806597	**AC234771**
26	chrX:134618182-134823894	Amplicons	Reassembled ampliconic sequence	JH806597	**AC240442, AC240441**, AL953870, AL732579
27	chrX:139874095-140551160	Amplicons		JH806598	AL451048**, AC234778, AC234777, AC234779, AC234774, AC235097, AC240443**
28	chrX:143132077-143553446	Gap	Gap due to unclonable intervening sequence	JH806599, JH806600	**AC239727, AC239395, AC239921, AL500522, AL135920, AL590424, AC231838, AC231661, AC234064, AC243369,** gap**, AC243412, AC231757, AC231840, AC231842**
29	chrX:148450053-148846052	Gap and amplicons	Closed gap and newly identified palindrome	JH159150	**AC231841, AC231760, AC231843, AC231656, AC233288, AC244197, AC244098, AC244099, BX322650, AC231839,** BX321867**, AC235697, AC235953**
30	chrX:151578519-151722859	Amplicons		JH159150	**AC244102**
31	chrX:151955838-152212442	Amplicons	Closed gap and reassembled ampliconic sequence	JH159150	**AC243591, AC243428, AC243374,** AC226403, AC152010**, AC236972**
32	chrX:153191585-153475817	Amplicons		JH159150	**AC244097, AC245140, AC244090, AC244107**
33	chrX:154217902-154417402	Amplicons		JH159150	**AC234781, BX571846**

a. Genomic coordinates are from human genome reference assembly version hg18 (NCBI37), the reference version from which we initiated our reassemblies. Each region's coordinates are encompassed within the collection of clones sequenced across the region (Column #6).
b. Accession numbers in bold indicate BACs or fosmids with newly generated or revised full-length finished sequence or contigs of SHIMS assemblies. Accession numbers not in bold represent pre-existing clone sequences which have not been re-examined.
c. Regions already spanned by single-haplotype sequence, which we therefore did not sequence across.
d. This clone's sequence is currently in the process of being finished.

Supplementary Table 5. Tallies of X-linked gene classifications depicted graphically in Figure 3

	Human				Mouse			
	Single-copy	Multicopy	Ampliconic	Total	Single-copy	Multicopy	Ampliconic	Total
Shared	548	75	33	656	548	75	33	656
Lost in reciprocal species	19	19	17	55	14	18	2	34
X-linked gene duplications	0	4	9	13	0	17	12	29
Independently acquired	7	21	48	76	23	9	102	134
Total	574	119	107	800	585	119	149	853

Supplementary Table 9. Documenting expression of individual members of independently acquired X-linked multicopy and ampliconic gene families in human and mouse

Mouse Gene Family #	Gene name	# of mRNA-seq reads[ab]
1	*Arxes2*	5
	Arxes1	0
2	*Gm15107*	19
	Gm15093	0
	Gm15114	3
	Gm15127	9
	Gm15080	1
	Gm10439	13
	Gm15097	4
	Gm15091	7
3	*Cypt1*	731
	Cypt7	158
	Cypt8	20
4	*Gm5934*	4
	Gm4297	19
	Gm5935	10
	Gm10230	0
	Gm10486	0
	Gm14632	0
	Gm14819	0
	Gm5169	44
	Gm1993	74
	Gm5168	22
	Gm2012	35
	Gm2030	11
	Slx	5
	Gm14525	5
	Gm6121	37
	Gm10487	22
	Gm10488	10
5	*Gm2933*	3
	Gm2799	49
6	*Gm10922*	3
	Gm10921	38
7	*Slxl1*	1511
	3830403N18Rik	30
8	*Gm6880*	95
	Gm6890	34

Human Gene Family #	Gene name	# of mRNA-seq reads[ac]
1	*CSAG1*	25
	CSAG2	21
2	*CT47A1*	0
	CT47A2	0
	CT47A6	0
	CT47A7	6
3	*VCX*	83
	VCX2	234
	VCX3A	34
	VCX3B	2
4	*PAGE2*	77
	PAGE2B	214
5	*SPANXN1*	3
	SPANXN2	35
	SPANXN3	148
	SPANXN4	66
	SPANXN5	24
6	GAGE1	6
	GAGE10	15
	GAGE12B	0
	GAGE12G	0
	GAGE12H	0
	GAGE12I	0
	GAGE12J	12
	GAGE13	0
	GAGE2A	0
	GAGE2B	0
	GAGE2C	0
	GAGE2D	0
	GAGE2E	0
	GAGE4	0
	GAGE5	0
	GAGE6	0
	GAGE8	0

a. The absence of mRNA-seq reads corresponding to a specific member of a gene family does not necessarily imply that that copy is transcriptionally inactive. For example, the absence of mRNA-seq reads corresponding to a specific copy may be due to insufficient read depth or reflect the absence of that gene variant in the genome of the sampled testis.

b. RNA-seq data from Brawand, D. et al. The evolution of gene expression levels in mammalian organs. Nature 478, 343-8 (2011).
c. RNA-seq data from Bradley, R.K. et al. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol 10, (2012).