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.

Nature Genetics: doi:10.1038/ng.2705

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.

Nature Genetics: doi:10.1038/ng.2705