The evolutionary history of human-specific gene duplications

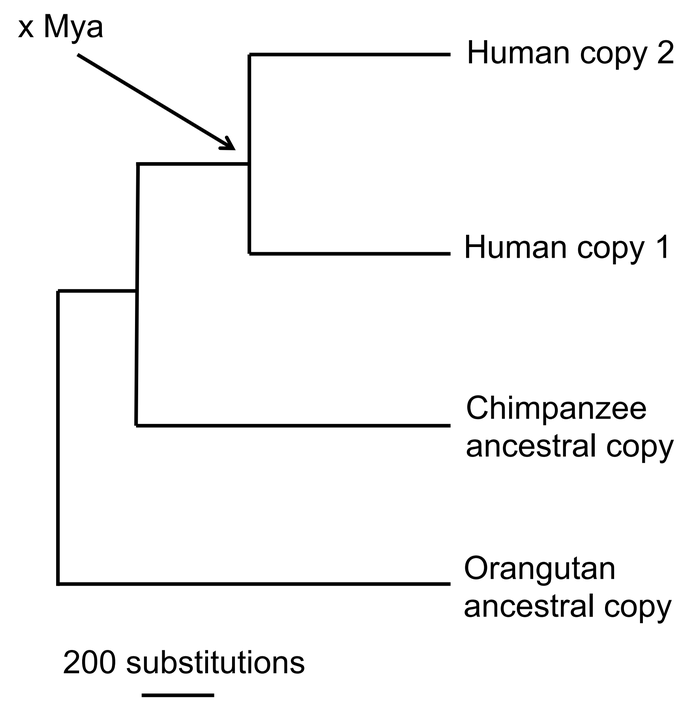
Priya Moorjani[1,2], Bob Handsaker[2,3,4], Franck Polleux[5], Steve McCarroll[2,3,4], Molly Przeworski[1,6]

[1] Department of Biological Sciences, Columbia University, New York, NY; [2] Program in Medical and Population Genetics, Broad Institute, Cambridge, MA; [3] Department of Genetics, Harvard Medical School, Boston, MA; [4] Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA; [5] Department of Neuroscience, Zuckerman Mind Brain Behavior Institute and Kavli Institute for Brain Science, Columbia University, New York; NY; [6] Department of Systems Biology, Columbia University, New York, NY

Introduction

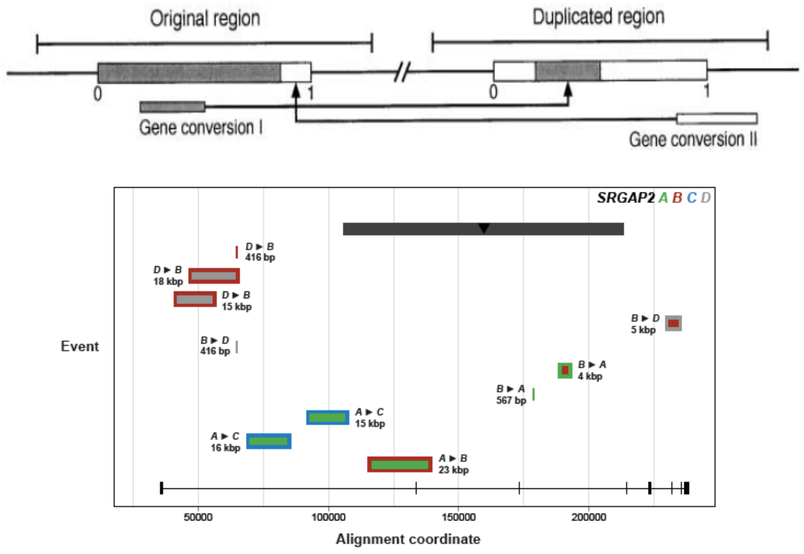
While gene duplications are known to have led to functional novelty in many species, the evolutionary history and functional impact of duplications that arose on the human lineage remains unknown. A subset (notably, *SRGAP2C*[1] and *ARHGAP11B*[2]) has been shown to have important functions in spine maturation and neocortex expansion and may contribute to the risk of neuropsychiatric diseases. Despite their importance for understanding human adaptation and evolution, most of the current research on human-specific duplications has been limited to biomedical applications and the few evolutionary studies to date have not accounted for the confounding effects of paralogous gene conversion that homogenizes the variation between the copies leading to inaccurate estimates of the age of the duplications and potentially of the evidence for selection[3,4]. To estimate the age of duplications, we develop an Approximate Bayesian Computation (ABC) method that models the pairwise divergence between duplicated copies as a function of the time, mutation rate and non-allelic gene conversion rate. Direct estimates for the rates of paralogous gene conversion are limited and hence we use an approach that accounts for the distance between duplicates, local recombination rate, and allelic gene conversion rate to infer these rates. Further we use the revised mutation rate that has recently emerged from studies of human pedigrees[5,6]. Application of our method to *SRGAP2* provides substantially older dates than had previously been reported[4], indicating that *SRGAP2C* occurred long before the beginning of the neocortex expansion. Together, these analyses allow us to reliably infer the chronology of human-specific duplications and to assess their potential role in critical periods of innovation in the human evolution.

Using the molecular clock for estimating the age of duplications



The standard approach to estimating the age of duplicates is based on the pairwise sequence divergence among the duplicated copies (in this example, Human copy 1 and Human copy 2). Using an estimate of the direct mutation rate from sequencing human pedigrees or assuming a divergence time to an outgroup such as chimpanzees, the pairwise sequence divergence between duplicated copies can then be translated into an estimate of the time of divergence between the copies. (Regional variation in mutation rate can also be taken into account, e.g., by using a distant outgroup such as orangutan or macaque.) This approach does not account for the impact of paralogous gene conversion that occurs between non-allelic segments that share high sequence similarity and homogenizes the variation between the copies. This omission leads to a systematic downward bias in estimated dates[7,8].

Non-allelic homologous gene conversion



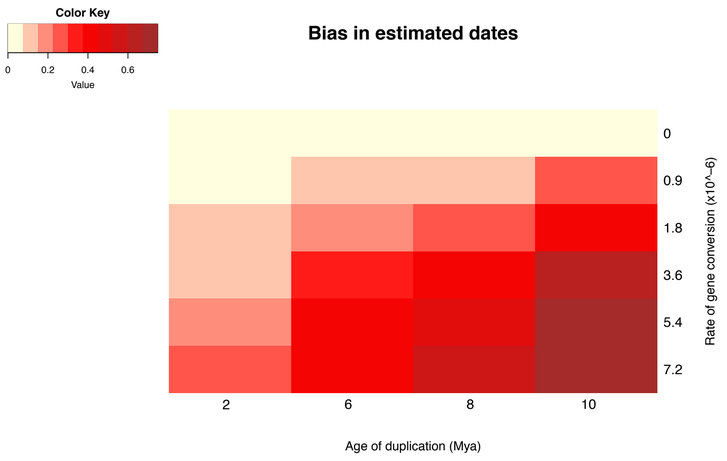
Figures from top: Teshima and Innan, 2004 and bottom: Nuttle et al., 2013

Non-allelic gene conversion acts as a copy paste mechanism leading to unidirectional transfer of short stretches of sequence between paralogous genes[7,8]. Tracks of shared sequences have been empirically documented in *SRGAP2* and other human-specific duplications[9], and thus this phenomenon needs to be taken into account in dating duplications[8].

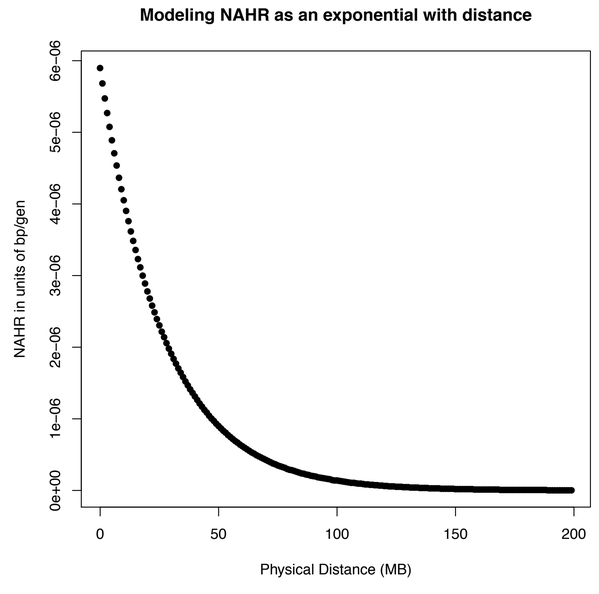
Impact of non-allelic gene conversion on the estimated age of duplications

To characterize the bias in estimated ages that arises when ignoring gene conversion, we performed coalescent simulations using *ms_dup*[8]. We assumed that the mutation rate $\mu = 1.2\times10^{-8}$ per bp per gen.[5,6], the rate of recombination $r = 10^{-8}$ per bp per gen., the rate at which a site is involved in a gene conversion track g is shown on the Y-axis, the gene conversion track length $l = 75$ bp[10] and the length of duplicated region is 200 Kb. Shown in the heat plot is the bias, computed as:

$$Bias = (T_{truth} - T_{estimated}) / T_{truth}$$



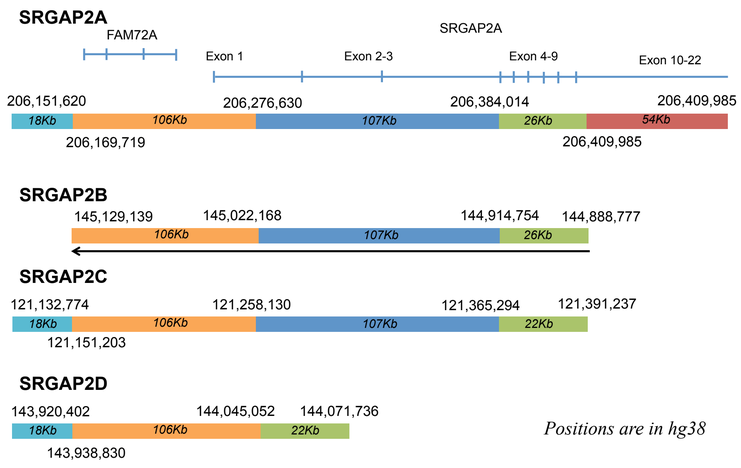
Model to infer regional estimates of non-allelic gene conversion



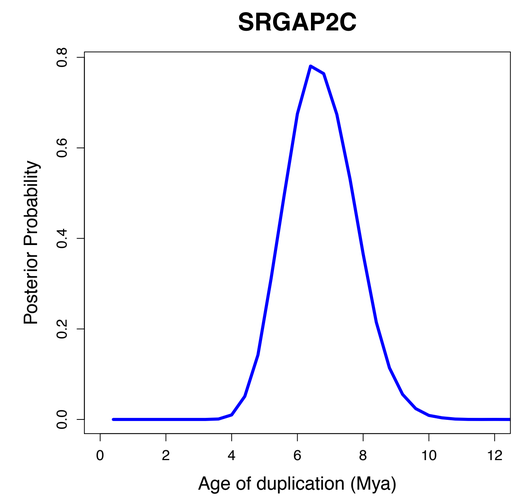
$$g = Ae^{-\lambda d}$$

Here, g is the rate at which a site is involved in a non-allelic gene conversion event, A is a regional estimate of the rate of allelic gene conversion[10,11], λ is the rate of decay of the rate of non-allelic gene conversion with physical distance. This model is motivated from observations in yeast[7] that the rate of ectopic exchange decreases as distance increases, and is highest when duplications are in tandem.

Estimating the age of human-specific duplication, SRGAP2



The *SRGAP2* gene has duplicated three times in human history. Most nonhuman primates only have the ancestral copy, *SRGAP2A*, but contemporary humans have between two to four copies. *SRGAP2C* is fixed in humans and has been functionally shown to play a role in neocortex expansion, by delaying spine maturation and increasing spine density[1]. In contrast, the *SRGAP2B* and *SRGAP2D* are present between zero and four copies[3]. To reconstruct the evolutionary history of human-specific duplications, we developed an Approximate Bayesian Computation (ABC) method that models the pairwise divergence between duplicated copies as a function of the time, regional mutation and gene conversion rates. We implement the approach as follows: (i) We sample parameters from prior distributions: g is the per bp per gen. rate of non-allelic gene conversion ~ U(0, 5.9e-6) and T is the time the duplication arose ~ U(1,12 Mya). The upper bound for T is based on the duplicate being human-specific and for g on the rate of allelic gene conversion in the region[10,11]. (ii) We simulate data using *ms_dup*[8] given a draw from the priors, and estimate the pairwise sequence divergence in the simulations (iii) We accept parameter values for those simulations in which the divergence in the simulated data is close to observed value (in practice <0.1%). The parameter values that we accept provide an estimate of posterior distribution on the parameters given the data.



Figures above shows the posterior probability of SRGAP2C for non-allelic *gene conversion rate of 1.2e-9 per bp per generation.*

Conclusions

Gene	% Pairwise divergence between ancestral and duplicated copy	Dates in Mya (based on the ABC method)§
SRGAP2B	0.567 ± 0.009	4.0 − 7.6
SRGAP2C	0.635 ± 0.009	4.4 − 8.4
ARHGAP11B	1.073 ± 0.007	8.8 − 12.8

§ assumes the revised pedigree based mutation rate of 0.5×10^{-9} per bp per year[5,6] and accounts for paralogous gene conversion and mutation rate in the region.

References

1. Charrier, C. et al. *Cell* 149, 923-935 (2012).
2. Florio, M. et al.. *Science* 347, 1465-1470 (2015).
3. Dennis, M. Y. et al. *Cell* 149, 912-922 (2012).
4. Antonacci, F. et al. *Nature genetics* 46, 1293-1302 (2014).
5. Moorjani P, Gao Z, Przeworski M. *PLOS Biology*; in press http://www.biorxiv.org/content/early/2016/08/05/058024 (2016).
6. Scally A *Current Opinion in Genetics & Development* 41, 36-43 (2016).
7. Sasaki, M., Lange, J. & Keeney, S. *Nature reviews Molecular cell biology* 11, 182-195 (2010).
8. Teshima, K. M. & Innan, H. *Genetics* 166, 1553-1560 (2004) .
9. Nuttle, X. et al. *Nature Methods* 10, 903-909 (2013).
10. Williams, A. L. et al.. *Elife* 4, e04637 (2015).
11. Halldorsson, B. V. et al. *Nature Genetics* (2016).