Gene expression dominance in allopolyploids: hypotheses and models.

Samuel Bottani[1,2*], Nicolae Radu Zabet[3*], Jonathan F. Wendel[4] and Reiner A. Veitia[2,5,#]

[1] Matière et Systèmes Complexes, UMR 7057, Paris 75013, France.

[2] Université Paris Diderot-Paris VII, 75205 Paris Cedex 13, France.

[3] School of Biological Sciences, University of Essex, Colchester, CO4 3SQ, UK.

[4] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011 USA.

[5] Institut Jacques Monod, Université Paris Diderot, CNRS UMR7592, Paris 75013, France.

[*]These authors contributed equally to this work.

[#]Correspondence should be addressed to Prof. Reiner A. Veitia

Email: reiner.veitia@ijm.fr

Phone: 33 1 57 27 81 16

Keywords: paralogs, gene duplication, subfunctionalization, whole genome duplication.

Abstract

The classical example of non-additive contributions of the two parents to allopolyploids is nucleolar dominance, which entails silencing of one parental set of ribosomal RNA genes. This has been observed for many other loci. The prevailing explanation for this genome-wide expression disparity is that the two merged genomes differ in their transposable element (TE) complement and in their level of TE-mediated repression of gene expression. Alternatively, and not exclusively, gene-expression dominance may arise from mismatches between *trans* effectors and their targets. Here, we explore quantitative models of regulatory mismatches leading to gene expression dominance. We also suggest that, when pairs of merged genomes are similar from one allopolyploidization event to another, gene-level and genome dominance patterns should also be similar.

Gene expression dominance in allopolyploids

Increasing molecular evidence points to non-additive contributions of the two parents to gene expression in allopolyploids and hybrids. The classical textbook example of 'gene expression' dominance is nucleolar dominance, in which one parental set of ribosomal RNA genes is silenced in an interspecific hybrid or allopolyploid. This can be recognized even at the chromosome morphological level. For instance, early work in the 1930's documented that *Crepis* species have a single chromosome with a nucleolar organizing region (NOR) that forms a secondary constriction [1]. In interspecific hybrids, only the NOR derived from one species (the dominant NOR) forms the characteristic chromosomal structure. This phenomenon has been confirmed and further studied in the molecular era [2] and has also been observed in animals such as in hybrids of *Xenopus* [3].

The favoring of expression of only one of two ancestral duplicates is not restricted to rDNA genes. For example, an analysis by Adams et al. (2003) showed that the parental subgenomes in synthetic allopolyploids involving the ancestors of the most widely cultivated cotton, *Gossypium hirsutum*, do not contribute equally to the transcriptome of the polyploid. Interestingly, expression patterns in synthetic polyploids can recapitulate those found in the naturally-occurring cotton [4]. This shows that evolutionarily persistent patterns of expression can appear just after polyploidization [5].

More generally, in recent years it has become evident that gene-expression dominance (or genome dominance if at a genome-wide level) is a characteristic feature

of many, but not all allopolyploids (reviewed in [6,7]). The prevailing explanation for the preferential expression of homoeologs (duplicated genes) from only one of the two progenitor genomes is that the two parents differ in their number and distribution of transposable elements (TEs), with TEs in one genome being, on average, closer to genes than they are in the other co-resident genome. This difference in TE adjacency to genes is thought to lead to repression of gene expression, via localized heterochromatinization, in the genome descended from the progenitor with the greater TE load. For example, it has been shown that gene expression in *A. thaliana* is negatively correlated with the density of surrounding methylated TEs [8]. This suggests that host silencing of TEs decreases the expression of nearby genes, and, in the context of allopolyploids, would account for differences in homoeolog expression. In both *Brassica* [9] and *Gossypium* [10] polyploids, the density of small RNAs targeting TEs is higher in regions adjacent to the gene copy of homoeolog pairs that exhibit lower expression levels, consistent with their heterochromatizing effects and repression of expression.

Although this TE-mediated explanation of gene expression dominance in allopolyploids is attractive, there are additional factors that might play a role. In particular, and as suggested below, sudden and persistent changes in transcription and associated chromatin states can result from regulatory mismatches between effectors (transcription factors -TFs- and epigenetic modifiers) and their target genes contributed by the merging genomes. This mismatch can arise from the divergence of the relevant

77 *cis/trans* regulatory players [11], whose transcriptional effects can be subsequently
78 stabilized by epigenetic changes.
79
80 Of course, not all polyploids are anticipated to behave the same. Autopolyploids,
81 for example, are not expected to experience strong alterations in the concentrations of
82 gene products (compared to the parental diploid) because of the proportionality
83 between DNA content and cellular and nuclear volumes [12]. That said, these
84 proportionalities are imperfect and there is some evidence of
85 transcriptomic/proteomic/metabolomic alterations in autopolyploids relative to their
86 diploid antecedents [13–15]. At the other extreme, allopolyploids between divergent
87 species typically combine genomes with different sizes and suites of TFs and other
88 regulatory effectors (i.e. miRNAs) that may have different physico-chemical properties
89 (i.e. affinities, concentrations, etc.). Below we argue that, in addition to the changes in
90 cellular and nuclear volumes accompanying allopolyploidy, these different TF/effector
91 concentrations may play an important role in the expression differences between the
92 parental subgenomes.
93
94 **A simple model of gene expression dominance**
95 To get insights, let us consider more specifically how a regulatory mismatch can
96 affect transcription from a promoter containing multiple binding sites for an activator
97 (TF). For example, assuming that transcription is proportional to the occupancy fraction
98 of a promoter f, then the normalized transcriptional response (TR) follows the Hill
99 equation:

$$f = TR = \frac{[TF]^s}{K^s + [TF]^s} \qquad \text{Eq. 1}$$

where *K* and *s* are constants [16,17]. The exponent is proportional to the number of binding sites per promoter. Here *[TF]* represents the free concentration of TF, but we will assume that this in excess to the promoter 'concentration', so that free and total concentrations are practically equivalent. *K* is a constant that corresponds to a concentration of TF such that f or TR is 50% of the maximum ($[TF]_{TR0.5}$). This implies that there is a dynamic range of the response around $[TF]_{TR0.5}$ and that there must be a relationship between the effective concentration of a TF and its affinity for its targets to ensure a proper regulation. Thus, evolutionary divergence may have resulted in different but 'suitable' threshold settings in the two parents, by modifying *[TF]* and/or *K* (which may be due to changes in the TF, in the target *cis* regulatory sequences or most likely both). Before polyploidization, the homologous transcription factors TF_1 and TF_2 are expressed in each parent at levels ensuring proper gene regulation. After allopolyploidization, the concentrations of TF_1 and TF_2 may be different from the parental values. Note that not only are concentrations different, but the promoters are in different contexts, sometimes with varying adjacencies to TEs, which might also limit TF accessibility. Moreover, TEs can also contribute TF binding sites. For instance, as many as 85% of the sequences that match the E2F consensus site in some *Brassica* species map to TEs. When such TEs are located close to genes they may directly participate in gene regulation, whereas those located far from genes may have an indirect TF-titrating effect [18].

After the genome merger, the actual TF concentration in the newly formed polyploid can be enough to activate the promoters with the smallest activation threshold (i.e., $[TF]_{TR0.5}$), whereas the other ones will be either less active or even silent (Figure 1). Obviously, the promoters having the lowest activation thresholds are more likely to be active after polyploidization. These biochemical considerations apply also to epigenetic mechanisms, as similar mismatches can occur for epigenetic modifiers (activators or repressors of gene expression). One can object that this reasoning holds for a limited set of genes. However, it is worth noting that some TFs undergoing the regulatory mismatch might work as global regulators of transcription. For instance, the vertebrate TF c-Myc acts as a universal nonlinear amplifier of the expression of active genes [19,20]. Thus, homoeologs of cMyc-like factors having different affinities for their *cis*-regulatory sequences contributed by merging genomes might account at least in part for genome-wide patterns of expression dominance.

Connecting gene-expression dominance to epigenetic changes.

As noted above, gene-expression dominance has been proposed to be connected to differential TE adjacency and an associated heterochromatization mediated by the small interfering RNA machinery [9]. This hypothesis is not mutually exclusive with the models presented here. A decrease in the gene activation frequency of the poorly expressed homoeolog, influenced by differential TE adjacency inherited from the diploid ancestors, may render it even more tolerant to mobilized TE insertion thus leading to a further decrease in chromatin accessibility. Box 1 outlines a conceptual model linking changes in gene expression to chromatin modifications using elementary stochastic processes acting according to binary conditions (i.e. the

146 homoeologs are accessible or not; the accessible homoeologs are induced or not,
147 Figure I). The parameters of the model can also take into account the effect of relative
148 TE adjacency, although this is not attempted here. It is worth noting that this epigenetic
149 mechanism holds for both individual homoeologous pairs and for genome-wide biased
150 expression of only one of the two parental genomes.

151

152 **More sophisticated models of gene expression dominance**

153 To further explore what may happen right after the shock [21] induced by the
154 genome merger, we next consider more complex models taking into account the
155 existence of DNA fostering non-functional interactions with the TFs. In the first situation,
156 the TFs or other regulatory agents, such as chromatin remodeling factors, will continue
157 to specifically regulate their original target genes (there is no cross-regulation after the
158 merger). Note that this possibility is extreme and is presented as a useful 'ground state'
159 assumption. The second, much more likely scenario, involves target cross-regulation in
160 the allopolyploid, as previously suggested to explain the poorly appreciated
161 phenomenon of gene expression dominance, where aggregate duplicate gene
162 expression mimics the level of one of the two parents [22,23].

163

164 More specifically, we assume that the two merging genomes have different sizes.
165 Homoeologs targets G_1 and G_2 in each genome will be regulated by the paralogous TFs,
166 TF_1 and TF_2 with different affinities (TF_i regulates/interacts with G_i in subgenome *i* with a
167 dissociation constant K_i). Depending on the scenarios, allowing cross-regulation or not,

168 TF_1 will also bind to the promoter of G_2 and *TF*$_2$ to the promoter of G_1. For simplicity, in
169 the following we deal with target promoters having only one binding site for the TFs.
170
171 A TF can be involved in two main types of binding events: specific and functional
172 binding to a *cis*-regulatory region with a transcriptional impact and nonfunctional
173 binding, specific or not, without transcriptional impact [24]. Nonfunctional binding is a
174 key ingredient of the model, which was not taken into account in Equation 1. In such
175 conditions, the fractional occupancy of the promoter f, which can be used as a proxy of
176 transcriptional activity, can be calculated using a classical approach [25,26] described in
177 Box 1 and in the *Supplementary Material*. We assume for simplicity the existence of
178 similar fractional occupancies of the promoters of the homoeologs before the merger.
179 To achieve similar promoter occupancies, a cell with a large euchromatic genome
180 fraction may have TFs displaying stronger affinities for functional sites to compensate
181 for the titration exerted by non-functional binding and to ensure better discrimination
182 between functional and non-functional binding or, alternatively, scale the concentrations
183 of the TFs with the size of euchromatic genome or, more likely, a situation in-between.
184 We have focused on the euchromatic region of the genome because it is the most
185 readily available to provide accessible binding sites. However, it is possible that the
186 heterochromatic and repeat fractions of the genome, which change from species to
187 species in size and composition, may also bind TFs [18]. The discussion that follows is
188 based on the genome euchromatic fractions but it is clear that for TFs interacting with
189 both euchromatin and heterochromatin the most relevant parameter would simply be
190 genome size (i.e., C-value). This is so because the effect of changing repeat number

and the heterochromatic compartment could potentially have more extreme consequences than changing the euchromatic space. We can speculate that this may also help explain why expression dominance patterns can vary among tissues, in which the eu-/heterochromatic compartments would also be different.

Although *cis* variation between species is amply documented, differences in the TFs can more easily account for potential genome-wide expression pattern changes. Obviously, both *cis* and *trans* changes are supposed to explain gene-specific regulatory mismatches. Here, for simplicity, we speak of TFs as if they were monomers. However, they frequently function as dimers or higher order structures, which can also explain differences in DNA recognition [24]. This is so because the assembly of such dimers/oligomers may be characterized by different physico-chemical parameters (specific assemble/disassembly rates in the parental species), and because they recognize bi-/multi-partite DNA binding sites with varying spacing and orientations, often different from the individual TF's motifs [27].

To begin our theoretical exploration, we will consider two limiting conditions: (A) when the concentrations of the two paralogous TFs are the same and the dissociation constants (K) regarding functional sites relative to non functional binding are inversely proportional to the euchromatic fraction of the genome and conversely, (B) when the dissociation constants are identical ($K_1 = K_2 = K$) and the TF concentrations are proportional to euchromatic genome size. Both cases are limiting situations and in reality both TF affinities and concentrations might be scaled with the size of the

214 euchromatic genome fraction. As well, and not exclusively, the two homoeologs may
215 differ in terms of promoter accessibility due to chromatin-level features. That is, even
216 with the same affinities, one homoeolog may be less transcribed than the other due to,
217 say, nearby heterochromatinization.
218
219 Figure 2A shows the occupancy probability of the promoters of the
220 homoeologous targets G_1 and G_2 in the allopolyploid as a function of the ratio of the
221 sizes of the euchromatic fraction of the merging genomes when the concentration is the
222 same for the two TFs (case A), but when affinity is stronger in the subgenome
223 containing a higher euchromatic fraction. In the absence of cross-regulation, we
224 observe that an increase in the ratio of non-functional sites (which is basically the ratio
225 of euchromatic genome sizes), NS_2/NS_1, leads to a predominant expression of the
226 genes from the genome containing a larger euchromatic fraction and hence to
227 expression dominance. The gene from the subgenome with a larger euchromatic
228 fraction also dominates when there is full cross-regulation. Note that this follows from
229 the assumption that a cell whose genome has a large euchromatic fraction would have
230 TFs with higher affinities. Regarding the exaggerated span of the NS_2/NS_1 ratios in
231 Figure 2, it is clear that most often the progenitor diploids of allopolyploids have
232 relatively small differences in genome sizes (e.g. 1.1-2, as in *Brassica* or
233 *Tragopogon*/goatsbeard and *Gossypium*/cotton). However, given the prevalence of high
234 infrageneric variation in genome sizes among diploids (e.g., 13-fold in *Vicia*) and in
235 allopolyploids (http://data.kew.org/cvalues/), we anticipate that examples of greater
236 differences in progenitor diploid genome sizes will be documented.

In Figure 2A, we considered only the case of a promoter occupancy of 50% in the parental genomes (f = 0.5). We next tested the effect of parental promoter occupancy level on our findings. Figure 2B shows the occupancy probability in the allopolyploid as a function of the promoter occupancy in the parental genomes (still assuming that $f = f_1 = f_2$), when one of the genomes has a euchromatic fraction 5 times larger than the other (i.e., $NS_2 = 5 \times NS_1$). The response of the promoters right after the merger is non-linear with respect to parental f, both in the absence or in the presence of cross-regulation. Indeed, there is a value of parental promoter occupancy for which the difference between the occupancy levels of the promoters from the two subgenomes reaches a maximum. In Figure 2B, this happens somewhere between 0.4 and 0.8 and can be explained by the fact that this interval represents the dynamic range of the promoters, so that small differences in TF concentrations/affinities are amplified and lead to high differences in occupancy.

Thus far, we have observed that for equal TF concentrations in the two parental cells, gene expression dominance could emerge from a co-adaptation between the dissociation constants and euchromatic genome size. Next, we consider the case (B), where we do not assume identical concentrations of the TFs in the original genomes but, instead, identical values of the dissociation constants $K_1 = K_2 = K$. In contrast to the previous case, the target promoters would be occupied to similar extents in the polyploid and there would be no dominance.

The driving factor of gene dominance in scenario A is the difference in the affinities of the two TFs. To further explore this effect, we have also studied an intermediate situation in which $K_2 = \beta K_1$, so that $0 \leq \beta \leq \frac{N_1}{N_2}$ (i.e., the inverse proportionality between *Ks* and euchromatic genome fraction is less strict). In the *Supplementary Material*, we show that the expression of the gene from the genome with the largest euchromatic fraction dominates as long as $\beta < 1$, thus, confirming under less extreme conditions that the adaptation in binding affinities of the TFs may explain at least in part the cases where gene expression in the larger euchromatic genome is dominant.

It is important to keep in mind that these models apply only to what happens right after genome merger. After that, genes encoding the TFs mentioned above may themselves undergo the gene-dominance effect, because they too are the targets of TFs. This cascading or snowballing effect will foster a positive feedback that will translate into a reinforcement of the initial situation, that is, the dominant homoeolog would dominate even more. This suggests that situations less extreme than those depicted above can in principle lead to expression dominance.

Gene dominance, genome dominance and replaying the tape of evolution.

A genome-wide extension of the gene dominance models outlined above is genome dominance, that is, when one subgenome dominates over the other in terms of gene expression. Genome dominance is frequently observed in allopolyploids [6], and this observation is conceptually connected to the longer-term process of biased

283 homoeolog deletion and the phenomenon of biased fractionation of two progenitor
284 genomes. Indeed, evolutionary analyses show that in (paleo)polyploid plants the
285 parental subgenome whose genes have higher expression levels, compared to the
286 other parental subgenome, lose a lower fraction of homoeologs [6,9,28–30].
287
288 In closing it is interesting to mention the thought experiment of ‘replaying life’s
289 tape’ popularized by Gould [31]. That is, one presses the rewind button and runs the
290 tape again, to see if the repetition reiterates the original or if “any replay of the tape
291 would lead evolution down a pathway radically different from the road actually taken”. In
292 our opinion, gene dominance is a candidate example of a phenomenon that leads to a
293 replaying of the evolutionary tape. Whenever the two merged genomes in an
294 allopolyploid are similar, from one polyploidization event to another, patterns of non-
295 additivity and the degree of genome dominance are likewise expected to be similar.
296 Thus, evolutionarily persistent, non-additive gene expression patterns in allopolyploids
297 and their presence in synthetic polyploids provide examples of a deterministic tendency
298 for repeatability in the playing the tape of evolution, across multiple polyploidy events. In
299 this sense, the models presented here may contribute to our understanding of why such
300 events are predictable based on a detailed understanding of genome history and
301 architecture and of the physiochemical properties of the key players involved in
302 transcriptional regulation. In this context, it would be interesting to assess the extent to
303 which affinities and abundances of TFs or epigenetic regulators correlate with the
304 amount of accessible chromatin or even with genome size and how this translates into
305 better discrimination of signal (i.e. a true binding site) from noise (non-functional

306 binding) [32]. It will also be interesting to assess the extent of the contribution of TE
307 adjacency to gene/genome dominance and the contribution of primary regulatory
308 mismatches leading to decreased expression of homoeologs, a phenomenon which is
309 subsequently consolidated by epigenetics and TE insertions around such homoeologs.
310 Thus, a challenge for the future is to better connect biased homeolog expression and
311 subgenome dominance to their corresponding genomic contexts and to the relevant
312 parameters of models similar to those presented here. Finally, we have considered here
313 what happens immediately following merger of two genomes; in this respect it will be of
314 interest to compare synthetic versus natural allopolyploids, to understand more
315 precisely how evolution shapes gene expression following allopolyploidy. It has been
316 argued that all models are wrong [33]. The ones presented above will not be an
317 exception but we hope they will be useful.
318
319 **Acknowledgements**
320
321 We warmly thank M. Freeling and J. Birchler for their helpful comments on a previous
322 version of this manuscript and for pointing to interesting references. We are indebted to
323 S. Caburet and the three referees of this manuscript for their enthusiasm and their most
324 helpful comments and suggestions.
325

References.

1 Navashin, M. (1934) Chromosome alterations caused by hybridisation and their bearing upon certain general genetic problems. *Cytol. Tokyo* 5, 169–203

2 Chen, Z.J. *et al.* (1998) Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in Arabidopsis allopolyploids. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14891–14896

3 Maciak, S. *et al.* (2016) Nucleolar Dominance and Repression of 45S Ribosomal RNA Genes in Hybrids between Xenopus borealis and X. muelleri (2n = 36). *Cytogenet. Genome Res.* 149, 290–296

4 Adams, K.L. *et al.* (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4649–4654

5 Adams, K.L. and Wendel, J.F. (2005) Novel patterns of gene expression in polyploid plants. *Trends Genet. TIG* 21, 539–543

6 Steige, K.A. and Slotte, T. (2016) Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Curr. Opin. Plant Biol.* 30, 88–93

7 Wendel, J.F. *et al.* (2018) The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* in press,

8 Hollister, J.D. and Gaut, B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19, 1419–1428

9 Cheng, F. *et al.* (2016) Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. *New Phytol.* 211, 288–299

10 Renny-Byfield, S. *et al.* (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Mol. Biol. Evol.* 32, 1063–1071

11 Veitia, R.A. (2005) Paralogs in polyploids: one for all and all for one? *Plant Cell* 17, 4–11

12 Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76, 65–101

13 Stupar, R.M. *et al.* (2007) Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics* 176, 2055–2067

14 Dai, F. *et al.* (2015) Phenotypic and Transcriptomic Analyses of Autotetraploid and Diploid Mulberry (Morus alba L.). *Int. J. Mol. Sci.* 16, 22938–22956

15 Wang, Z. *et al.* (2017) Implications of polyploidy events on the phenotype,

microstructure, and proteome of Paulownia australis. *PloS One* 12, e0172633

16 Chu, D. *et al.* (2009) Models of transcription factor binding: sensitivity of activation functions to model assumptions. *J. Theor. Biol.* 257, 419–429

17 Bost, B. and Veitia, R.A. (2014) Dominance and interloci interactions in transcriptional activation cascades: models explaining compensatory mutations and inheritance patterns. *BioEssays News Rev. Mol. Cell. Dev. Biol.* 36, 84–92

18 Hénaff, E. *et al.* (2014) Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. *Plant J. Cell Mol. Biol.* 77, 852–862

19 Nie, Z. *et al.* (2012) c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* 151, 68–79

20 Lin, C.Y. *et al.* (2012) Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* 151, 56–67

21 McClintock, B. (1984) The significance of responses of the genome to challenge. *Science* 226, 792–801

22 Grover, C.E. *et al.* (2012) Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* 196, 966–971

23 Yoo, M.-J. and Wendel, J.F. (2014) Comparative evolutionary and developmental dynamics of the cotton (Gossypium hirsutum) fiber transcriptome. *PLoS Genet.* 10, e1004073

24 Todeschini, A.-L. *et al.* (2014) Transcription factors: specific DNA binding and specific gene regulation. *Trends Genet. TIG* 30, 211–219

25 Bintu, L. *et al.* (2005) Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* 15, 116–124

26 Zabet, N.R. and Adryan, B. (2015) Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Res.* 43, 84–94

27 Jolma, A. *et al.* (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527, 384–388

28 Flagel, L.E. and Wendel, J.F. (2010) Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol.* 186, 184–193

29 Schnable, J.C. *et al.* (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4069–4074

30 Woodhouse, M.R. *et al.* (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci. U. S. A.* 111, 5283–

5288

31 Gould, S.J. (1991) *Wonderful life—the Burgess shale and the nature of history. New York*, WW. Norton & Company.

32 Wunderlich, Z. and Mirny, L.A. (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet. TIG* 25, 434–440

33 Box, G.E.P. (1976) Science and Statistics. *J. Am. Stat. Assoc.* 71, 791–799

34 Scruggs, B.S. *et al.* (2015) Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* 58, 1101–1112

35 Simicevic, J. *et al.* (2013) Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nat. Methods* 10, 570–576

410 **Figure Legends**

411 **Figure 1.** A) Effect of an allopolyploidization event involving the parental cells 1 and 2
412 on transcription from the homoeologous promoters of genes G1 and G2, responding to
413 the transactivators TF_1 (or TF2, the homoeolog). The green circles represent TF_1 (or
414 TF2), which have different affinities for their cognate *cis*-regulatory sequences. B) We
415 assume that the merging cells have the same volume (that K_1 = 1nM, K_2 = 2 nM and
416 that parental TR=0.5). Right after the merger, let us suppose that cell volume doubles.
417 In this case, a genome merger will lead to a TR in which the alleles of Parent 1 will be
418 expressed at more than 80% of saturation, whereas those from parent B will be working
419 at less than 25% (i.e., parent 1 would be dominant).

420

421 **Figure 2.** Promoter occupancy probability in the allopolyploid as a function of the ratio
422 of the euchromatic genome sizes. A) Promoters 1 and 2 refer to the promoters of genes
423 G_1 and G_2 in the allopolyploid according to their parental origin. f represents the
424 homoeologous promoter occupancy probability in the parents (identical for simplicity).
425 Note that for genes (red curves) in genomes with very different euchromatic fractions
426 $NS_2 > NS_1$ the occupancy levels can be very different, leading to gene expression
427 dominance. **B**) The graph displays the promoter occupancy probability in the
428 allopolyploid as a function of occupancy in the parental genomes (assumed to be
429 identical in both parents). Upper panel: no cross-regulation; Lower panel: full cross-
430 regulation, for NS_2=5 x NS_1 (see supplementary figure for NS_2=2 x NS_1). The difference
431 between the homoeologous promoter occupancies in the allopolyploid reaches a
432 maximum for a particular value of f. C) Effect of allopolyploidization on transcription

433 from the homoeologous sub-genomes. Right after the merger there is an epi/genomic
434 shock [21] whose effects can be further strengthened by cascading effects, as
435 described in the main text, leading to gene expression dominance.
436
437 **Box 1.**
438 **Epigenetic 'use it or lose it' model.**
439 The process of epigenetic gene inactivation can be summarized in a two-step model
440 (Figure I). Accordingly, both homoeolog genes may be initially accessible (G_a) to the
441 transcription machinery. They may stochastically shift towards either an epigenetically
442 inaccessible (G) state or to an 'induced' state (G_I), allowing transcription. The constants
443 k_a, γ_a represent gain and loss of chromatin accessibility. A similar reasoning applies to
444 transcription induction and arrest in the context of open chromatin. Translation and
445 protein degradation are also represented. Under irreversible conditions (k_a=0, rather
446 extreme situation), there will be an exponential decay of the expression of the
447 homoeologs considered over the entire population. Just after genome merger both sets
448 of homoeologs (from each subgenome) would be accessible. Importantly, accessibility
449 for the more weakly expressed homoeolog subsequently decreases because of the
450 reduced level of induction/activation. Played out over a long time frame (within an
451 individual or within a population), depending on the parameters, there may be extinction
452 of the expression of such 'recessive' homoeologs in the majority of cells/individuals, as
453 shown in Figure IB. Some degree of expression of the 'recessive' homoeologs can
454 remain in reversible conditions ($\text{for } k_a > 0$), Figure IC. Recent work in mammalian

455 systems has shown that the presence of Pol II at active genes keeps the local
456 chromatin accessible and further permits TF binding at the promoter regions [34].
457
458 **Figure I.** A) Model of gene expression with a first order chromatin accessibility kinetics.
459 G denotes the alleles of the gene of interest in an inaccessible state due to, for
460 instance, the presence of repressive chromatin marks; G_a represents the alleles of the
461 gene in an accessible state and G_I is the accessible allele 'induced' for transcription
462 (according to a set of parameters described in the Box). The repressive effects of TE
463 adjacency can be taken into account in γ_a. (B and C) The graphics represent typical
464 time-courses of the percentage of highly or poorly expressed homoeologs in the cell
465 population but the idea could be extrapolated to a population of individuals over longer
466 periods of time. We assumed that the proportions do not change as they proliferate. (B)
467 when there cannot be recovery of chromatin accessibility $k_a = 0$, the proportion of
468 recessive alleles in G_a state is higher and their shift to state G favored. The decay of
469 accessibility is faster for the poorly expressed homoeologs. C) when k_a is proportional to
470 allele expression rate, at later time-points, there is a stable fraction of poorly expressed
471 homoeologs that remains active.
472
473 **Box 2**
474 The fractional occupancy of the promoter (which can be considered as a proxy of
475 transcriptional activity) in the parental species can be calculated using a classical
476 approach described in more detail in the *Supplementary Material*. In short,
477 $$f_i = \frac{[TF_i]}{[TF_i]+[NS_i]K_i} \qquad \text{Eq.2}$$

where $[NS_i]$ are the concentrations of 'non-functional' binding sites in the merging genomes (*i = 1* or *2*). We can obtain an expression for the TF concentrations from Eq. 2. This model can be generalized by assuming cross-regulation, where TF_1 can bind to promoter G_2 with a dissociation constant K_{12} and TF_2 can bind to promoter G_1 with a dissociation constant K_{21}.

We consider different nuclear volumes for the merging organisms (correlating with their genome sizes [12]): V_1 and V_2 for the original organisms and $V_3 = V_1 + V_2$, for the allopolyploid. This simplifying assumption can be relaxed to 'V_3 is proportional to V_1+V_2' without changing the conclusions drawn from our models.

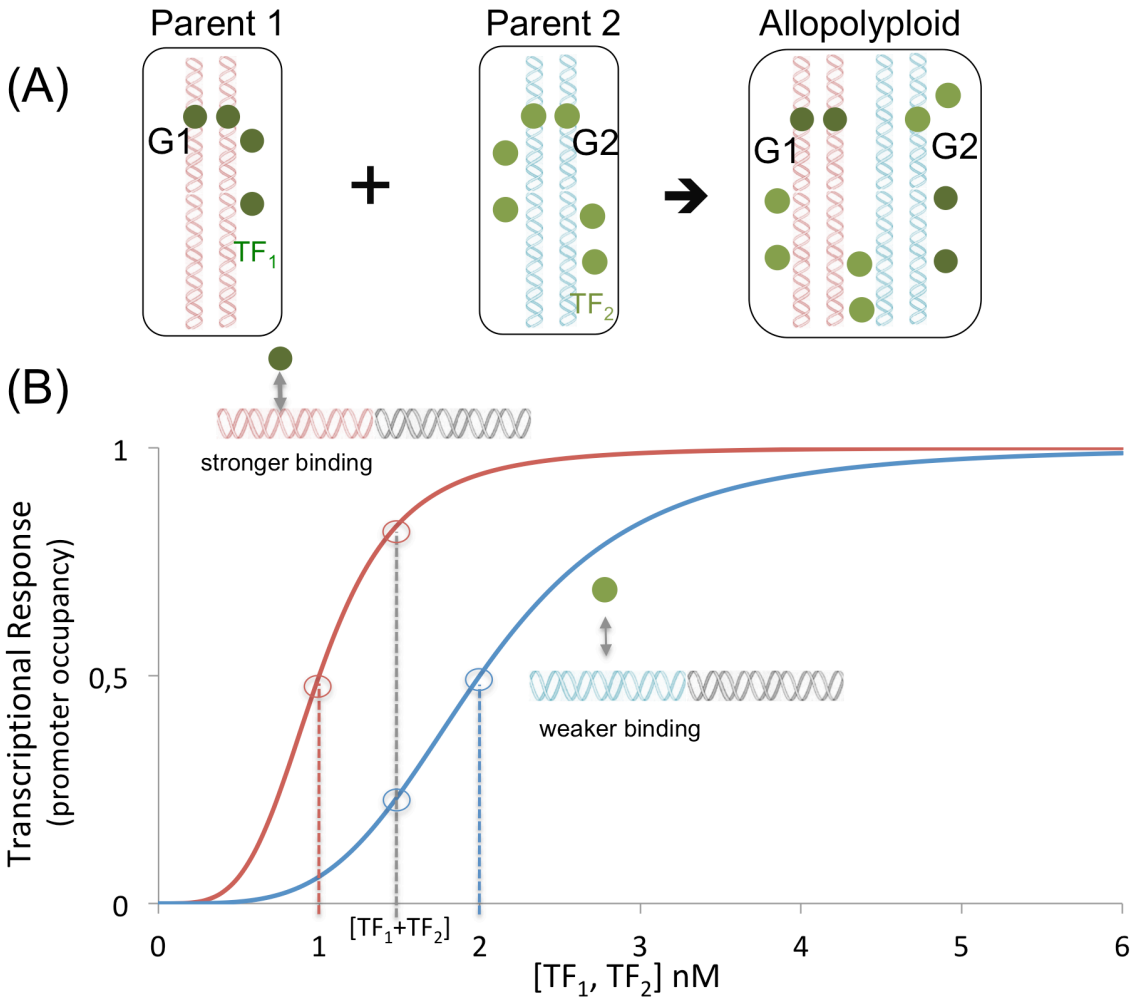
Fractional occupancies of the promoters in the allopolyploid (f'_i) can then be calculated after the introduction of cross-regulation coefficients (C_1 and C_2), which depend on the cross-regulation binding constants K_{12} and K_{21}. This coefficient is such that C_1=1 for full cross-regulation, and C_1=0 in the unrealistic limit of no cross-regulation (see *Supplementary Material*). Thus,

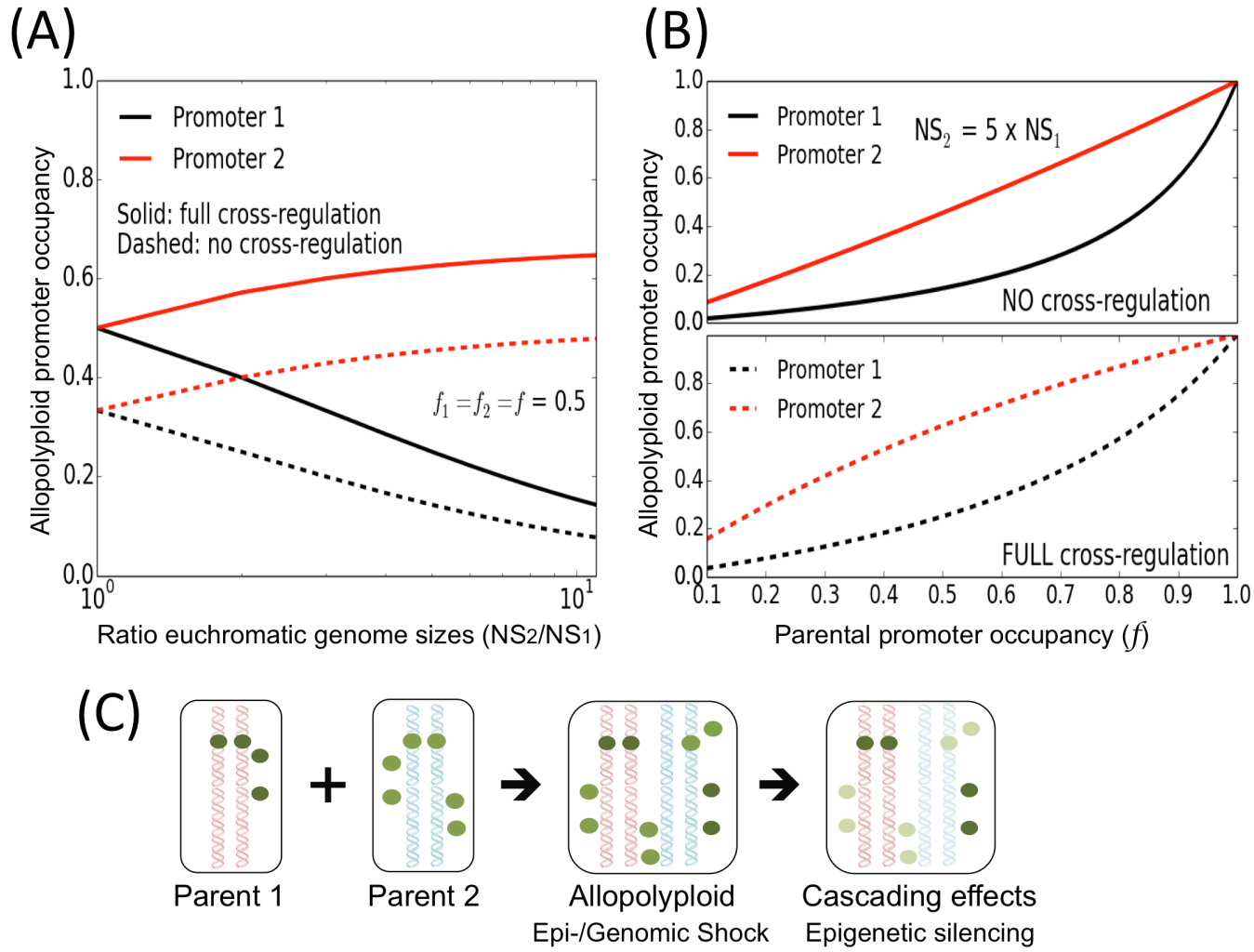
$$f'_1 = \frac{[TF_1]' + [TF_2]'C_1}{[TF_1]' + [TF_2]'C_1 + [NS]K_1} \quad \text{Eq. 3}$$

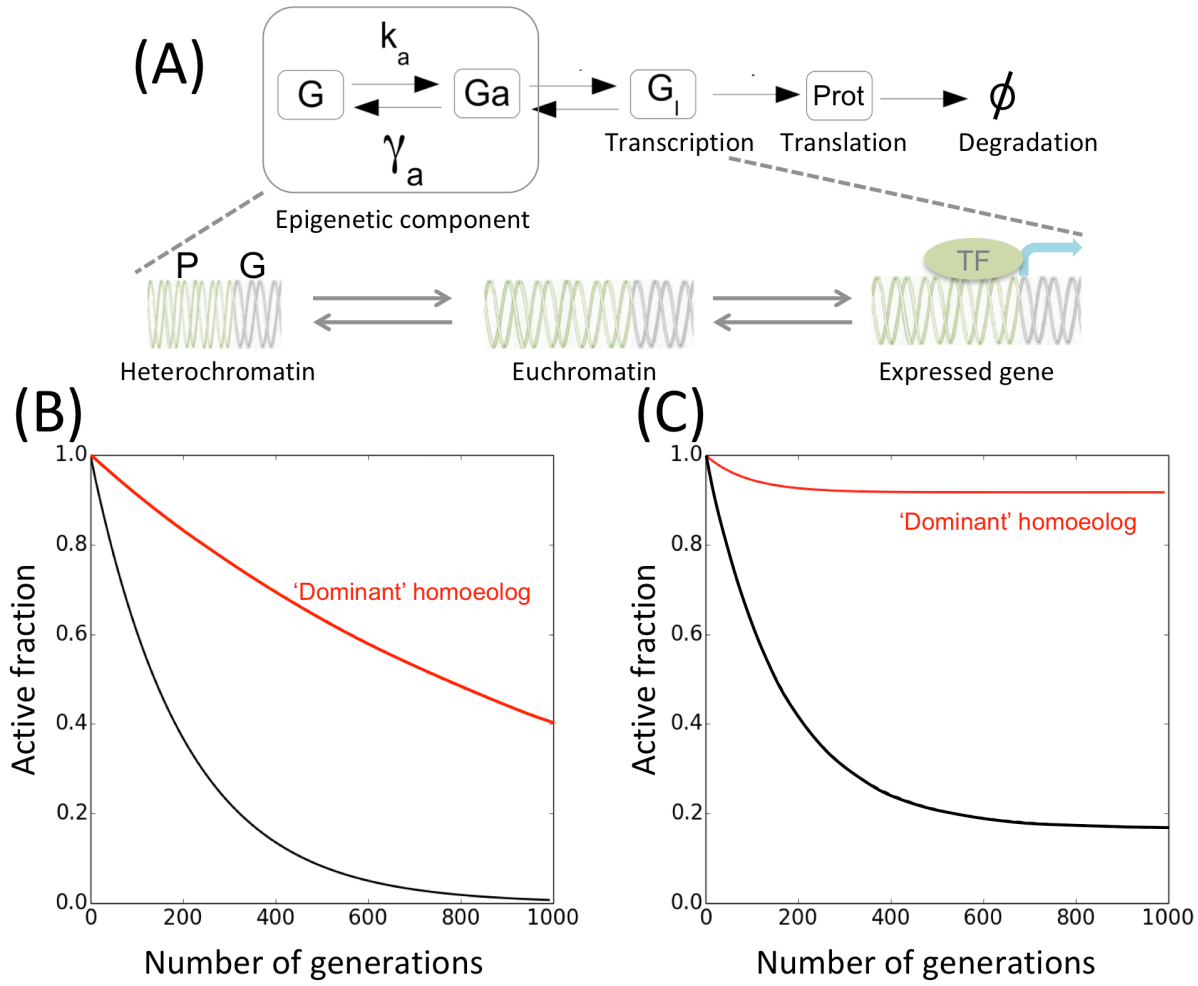
$$f'_2 = \frac{[TF_2]' + [TF_1]'C_2}{[TF_2]' + [TF_1]'C_2 + [NS]K_2} \quad \text{Eq. 3}'$$

where the total $[NS] = (NS_1 + NS_2)/V_3$.

We assume for mathematical simplicity the existence of similar fractional occupancies of the promoters of the homoeologs before the merger (i.e., $f = f_1 = f_2$). As outlined in the main text, to achieve the same promoter occupancy, a cell with a large euchromatic genome fraction may scale (i) the affinities of TFs for functional sites and/or (ii) the concentrations of the former with the size of euchromatic genome (or even with genome size, depending on the TFs). Here, we consider two limiting conditions: (A) when $[TF_1] = [TF_2]$ and the dissociation constants are inversely proportional to the euchromatic fraction of the genome and conversely, (B) when the dissociation constants are identical ($K_1 = K_2 = K$) and the TF concentrations are proportional to euchromatic genome size. In the first case (A), the cell would not pay an extra cost of TF production, and is mathematically captured by the relationship $K_2 = \frac{N_1}{N_2} K_1$ [25,26,35]. As already noted, reality is expected to be somewhere in-between these two extreme situations.


(A)
Parent 1
Parent 2
Allopolyploid
G1
G2
G1
G2
TF1
TF2
(B)
stronger binding
weaker binding
Transcriptional Response
(promoter occupancy)
1
0,5
0
0
1
[TF1+TF2]
2
3
4
5
6
[TF1, TF2] nM


(A)
Promoter 1
Promoter 2
Solid: full cross-regulation
Dashed: no cross-regulation
$f_1 = f_2 = f$ = 0.5
Allopolyploid promoter occupancy
Ratio euchromatic genome sizes (NS2/NS1)
(B)
Promoter 1
Promoter 2
NS_2 = 5 x NS_1
NO cross-regulation
FULL cross-regulation
Allopolyploid promoter occupancy
Parental promoter occupancy (f)
(C)
Parent 1
Parent 2
Allopolyploid
Epi-/Genomic Shock
Cascading effects
Epigenetic silencing


(A)
k a
G
Ga
γ a
G l
Prot
ϕ
Transcription
Translation
Degradation
Epigenetic component
P
G
TF
Heterochromatin
Euchromatin
Expressed gene
(B)
Active fraction
'Dominant' homoeolog
Number of generations
(C)
Active fraction
'Dominant' homoeolog
Number of generations

Appendix

A model of genome dominance in allopolyploids.

The Model

We assume that the two merging genomes have different sizes (and thus different euchromatic compartments). Homoeologous target genes G_1 and G_2 in each genome will be regulated by the paralogous TFs 1 and 2. TF_1 regulates T_1 in (sub)genome 1 with binding energy E_1; TF_2 regulates G_2 in (sub)genome 2 with binding energy E_2. According to the cases considered below, TF_1 can also bind to the promoter of G_2 and TF_2 to the promoter of G_1. We also consider different nuclear volumes for both parental organisms, respectively V_1 and V_2 and V_3 for the allopolyploid. For simplicity, we consider $V_3 = V_1 + V_2$. In the following, we deal with target promoters with only one TF binding site.

The fractional promoter occupancy or binding is calculated as follows:

$$f_i = \frac{[TF_i]e^{-E_i/k_BT}}{[TF_i]e^{-E_i/k_BT} + [NS_i]e^{-E_{NS}/k_BT}} = \frac{[TF_i]}{[TF_i] + [NS_i]K_i}, \quad Eq.1$$

Where $[NS_i]$ is the concentration of non-functional binding sites in genomes *i* (*i*=1 or 2). This corresponds to the euchromatic genome fraction. Note that the dissociation constants K_1 and K_2 can be different for the two paralogous factors TF1 and TF2.

As the target genes in the merging genomes are homoeologs, we assume for simplicity the existence of similar fractional occupancies of the homoeologous promoters before the merger, $f_1 = f_2 = f$. The corresponding TF concentrations for this specific occupancy fraction can be derived from the above equation, which can also be expressed in terms of numbers of TF_i molecules as follows:

$$TF_i = V_i \frac{[NS_i]K_i f}{1 - f} = \frac{NS_i \times K_i f}{1 - f}, \quad Eq.2$$

In the allopolyploid, we must consider a larger volume and TF cross-regulation. Factor TF_1 binds with energy E_{12} on the paralogous target on the second genome (and similarly for TF_2, whose cross-regulation binding energy will be E_{21}).

Fractional promoter occupancies of the promoters in the allopolyploid f_1' and f_2' (of targets from subgenomes 1 and 2) are then given by:

$$f_1' = \frac{[TF_1]'e^{-E_1/k_BT} + [TF_2]'e^{-E_{21}/k_BT}}{[TF_1]'e^{-E_1/k_BT} + [TF_2]'e^{-E_{21}/k_BT} + [NS]e^{-E_{NS}/k_BT}}, \quad Eq.3$$

To describe cross-regulation we introduce a coefficient C_1, so that $e^{-E_{21}/k_BT} = C_1 e^{-E_1/k_BT}$. This coefficient is such that C_1=1 for full cross-regulation, and C_1=0 in the limit of no cross-regulation. For instance, when the TF_2 binds its target in genome 1 with 90% of its affinity for its cognate target on genome 2 ($E_{21} - E_1 = 0.1E_1$), the cross-regulation coefficient is of the order $C \sim 0.8$. Thus,

$$f_1' = \frac{[TF_1]' + [TF_2]'C_1}{[TF_1]' + [TF_2]'C_1 + [NS]K_1}, \quad Eq.4$$

where the concentrations are calculated in the volume V_3: $[TF_i]' = \frac{TF_i}{V_3}$ and the number of molecules TF_i are the same as in the respective original genome (Eq. 2), and $[NS] = (NS_1 + NS_2)/V_3$.

Symmetrically we get :

$$f_2' = \frac{[TF_2]' + [TF_1]'C_2}{[TF_2]' + [TF_1]'C_2 + [NS]K_2}, \quad Eq.5$$

We can now used Eq. 2 to express f_1' and f_2' as a function of parental promoter occupancy probability f in the original genomes this leads after simplifications to:

$$f_1' = \frac{1 + \frac{C_1 NS_2}{NS_1}\frac{K_2}{K_1}}{1 + \frac{C_1 NS_2}{NS_1}\frac{K_2}{K_1} + \left(1 + \frac{C_1 NS_2}{NS_1}\right)\frac{1-f}{f}}$$

and symmetrically for f_2'.

We now consider three cases: (A) in the original genomes both the occupancy probability of the homoeologous promoters and the TF concentrations are identical ($f = f_1 = f_2$ and $[TF_1] = [TF_2]$). This situation may result from the co-adaption between the dissociation constants and euchromatic genome fraction as discussed in the main text, (B) the dissociation constants are identical and the TF concentrations are scaled with euchromatic genome size and (C) a mixture of the previous cases.

Case A: when $[\boldsymbol{TF_1}] = [\boldsymbol{TF_2}]$

This implies the following relationships between the dissociation constants and the genome sizes $NS_1 \times K_1 = NS_2 \times K_2$ leading to:

$f_1' = \frac{1+C_1}{1+C_1+\left(1+\frac{N_2}{N_1}\right)\frac{1-f}{f}}$ and symmetrically for f_2'

In this case there may be cross-regulation or not (i.e. $C_{1,2}$=1 or $C_{1,2}$=0, respectively).

Case B: when $K_1 = K_2$

Here, we assume a co-evolution of the TF concentrations (i.e., proportional to euchromatic genome size). This could well be the case of highly conserved TFs binding to very similar promoter sequences. This implies of course full cross-regulation after the merger.

For simplicity, we also assume identical promoter occupancy in the original genomes $f_1 = f_2 = f$, which together with $K_1 = K_2 = K$ leads to:

$$[TF_i] = \frac{[NS_i].K.f}{1-f},$$

and

$$\frac{[TF_1]}{[NS_1]} = \frac{[TF_2]}{[NS_2]} = \frac{K.f}{1-f}$$

This amounts to say that ratio of TF abundances is the same as the ratio of the euchromatic genome sizes.

Regarding the abundances of the paralogous TFs in the allopolyploid, we consider a situation involving identical numbers of protein molecules in a higher volume. Accordingly, since we have full crosstalk, both promoters will be identically occupied implying the absence of dominance.

Case (C). Finally we consider an intermediate case where $\boldsymbol{K_2 = \beta K_1}$, that is the dissociation constants in the original organisms have an arbitrary ratio ($1<\beta<NS_1/NS_2$). The identical original promoters occupancy f and the respective TF abundances are constrained.

$$TF_1 = \frac{NS_1 \times K_1 f}{1-f}$$

$$TF_2 = \frac{NS_2 \times K_2 f}{1-f} = \frac{\beta NS_2 \times K_1 f}{1-f} = \frac{\beta NS_2}{NS_1} TF_1$$

In the allopolyploid, this gives

$$f_1' = \frac{[TF_1]' + [TF_2]' C_1}{[TF_1]' + [TF_2]' C_1 + [NS]K_1}$$

$$f_1' = \frac{1 + \frac{\beta NS_2}{NS_1} C_1}{1 + \frac{\beta NS_2}{NS_1} C_1 + \frac{NS}{NS_1} \frac{1-f}{f}}$$

and

$$f'_2 = \frac{\frac{\beta NS_2}{NS_1} + C_2}{\frac{\beta NS_2}{NS_1} + C_2 + \beta \frac{NS}{NS_1}\frac{1-f}{f}}$$

In such conditions, in absence of cross-regulation $\boldsymbol{C_1 = C_2 = 0}$

$f_i' = \frac{1}{1+\frac{NS}{NS_i}\frac{1-f}{f}}$

f'₁ and *f'₂* are independent of β. Thus, $f'_2 > f'_1$ when $N_2 > N_1$ (i.e., the genome with the largest euchromatic fraction dominates).

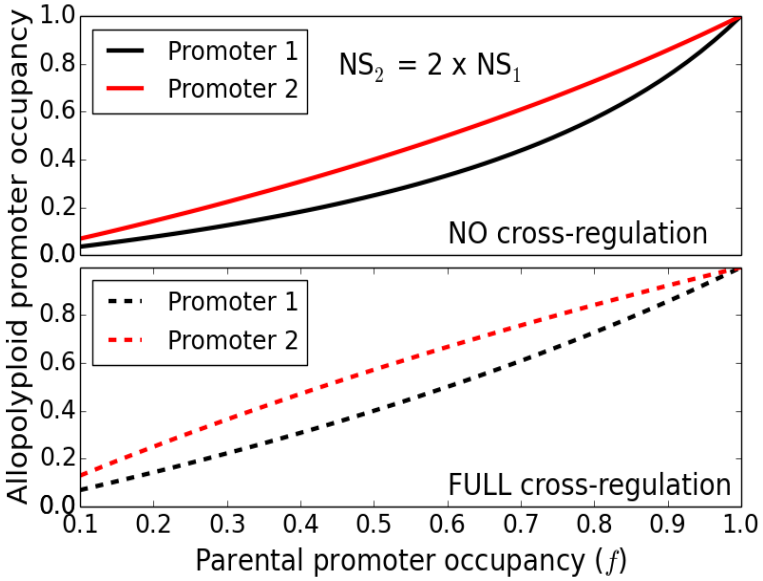
When there is full cross-regulation $C_1 = C_2 = 1$

$$f_1' = \frac{1 + \frac{\beta NS_2}{NS_1}}{1 + \frac{\beta NS_2}{NS_1} + \frac{NS}{NS_1}\frac{1-f}{f}}$$

and symmetrically for *f₂'*.

As we can safely assume $\beta < 1$, this leads again to dominance of the genome with the largest euchromatic fraction ($f'_2 > f'_1$).

Supplementary figure



The graph displays the promoter occupancy probability in the allopolyploid as a function of occupancy in the parental genomes (assumed to be identical in both parents). Upper panel: no cross-regulation; Lower panel: full cross-regulation, for NS_2=2 x NS_1