Chapter 4

POInT: Modeling Polyploidy in the Era of Ubiquitous Genomics

Gavin C. Conant

Abstract

Thirteen years ago, we described an evolutionary modeling tool that could resolve the orthology relationships among the homologous genomic regions created by a whole-genome duplication. This tool, which we subsequently named POInT (the *P*olyploid *O*rthology *In*ference *T*ool), was originally only useful for studying a genome duplication known from bakers' yeast and its relatives. Now, with hundreds of genome sequences that contain the relicts of ancient polyploidy available, POInT can be used to study dozens of different polyploidies, asking both questions about the history of individual events and about the commonalities and differences seen between those events. In this chapter, I give a brief history of the development of POInT as an illustration of the interconnected nature of computational biology research. I then further describe how POInT operates and some of the strengths and drawbacks of its structure. I close with a few examples of discoveries we have made using it.

Key words Polyploidy, Evolutionary model, Synteny

Abbreviations

POInT	Polyploid Orthology Inference Tool
WGD	Whole-genome duplication

1 Polyploidy and the Advent of Genomics

Very shortly after the rediscovery of Mendel's work [1], geneticists started to consider the role of polyploidy, the doubling (or more) of an organism's chromosome complement, in both genetics and evolution [2–4]. That interest continued, such that, when the first eukaryotic genome was released nearly a century later, it was very quickly shown to have the remnants of an ancient polyploidy encoded within it [5, 6].

Yves Van de Peer (ed.), *Polyploidy: Methods and Protocols*,
Methods in Molecular Biology, vol. 2545, https://doi.org/10.1007/978-1-0716-2561-3_4,

Due to the startlingly rapid improvements in sequencing technologies and the associated tools for assembly and genome comparisons [7], there are now hundreds of available genomes from species that underwent polyploidy at some point in their history [8]. For convenience, these polyploid lineages are often divided based on their age into young neopolyploids and old paleopolyploids [9], with an intermediate category of mesopolyploids used in some cases (e.g., [10, 11]). The precise distinction between these types arguably varies depending on the system and questions in play; for the purposes of this chapter, the most salient feature of the polyploidy is whether it occurred sufficiently long ago that both duplicate gene loss and speciation events have occurred since. Hence, unless otherwise qualified, in what follows polyploidy should be understood to refer to meso- or paleopolyploidy events.

These postpolyploidy gene losses are predicted by evolutionary theory, because, in the most straightforward framework, the genetic redundancy created by the polyploidy protects the organism from the deleterious effects of function-abolishing mutations in one copy of the duplicate pair [12, 13], allowing that copy to be lost through a combination of random mutation and genetic drift. This expectation is largely empirically confirmed by the observation that most duplicate genes are short-lived [14, 15], although it is notable that duplicates produced by polyploidies are longer-lived than are others [16].

2 Gene Loss, Comparative Genomics, and the Need for Models

The story of POInT begins with such duplicate losses, and I think it is instructive to give a brief history of how it developed. I do so less for the intrinsic interest of POInT and more as a reminder that many analysis tools, including POInT, are natural, if unexpected, developments of existing ideas and algorithms.

When a gene duplication of any type is shared among more than a single species, the tools of molecular phylogenetics can be used to model and understand its history [17]. In the special case of a polyploidy, however, there is information beyond the gene sequences themselves that can provide a great deal of assistance in understanding that history: the location of the duplicates in relation to the other genes in the genome. We can refer to groups of homologous genes that occur in the same order in two different genomes as being in synteny (although this conserved order is also sometimes referred to as colinearity, with synteny used instead to describe conserved gene content between genomes). Synteny is also of course useful in comparative genomics more generally (see Chen and Zwaenepoel; Berthelot et al.; in this volume; [18]), but, as I will argue below, it is absolutely essential to understanding the history of a paleopolyploidy event.

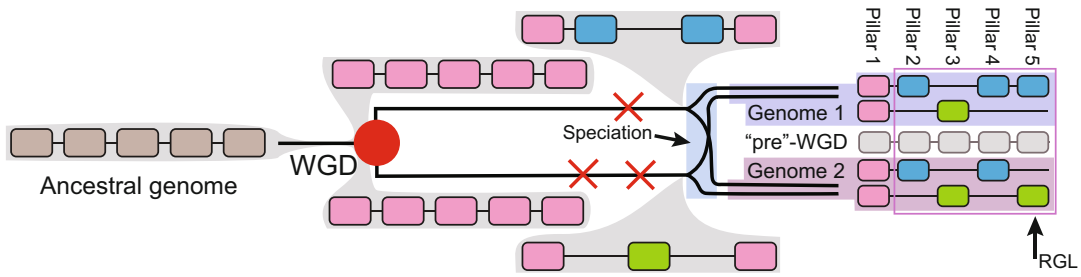

Fig. 1 Schematic of the evolutionary processes modeled with POInT, including gene losses and speciation after a whole-genome duplication. Immediately after the WGD, all five genes are present in two homoeologous copies. Three homoeologous gene losses occur prior to the split of the two species (red "X"s), one in the less fractionated subgenome (Track "0;" yielding the green gene in the lower window) and two from the more fractionated subgenome (Track "1;" yielding the two blue genes in the upper window). After the speciation event, Genome 1 loses a homoeolog from the more fractionated subgenome and Genome 2 loses one from the less fractionated subgenome, resulting in a case of reciprocal gene loss (RGL). The result is five "pillars" of duplicated or lost duplicated genes for the two genomes. The boxed region illustrates the principle that even polyploidies where most or all duplicates have been lost still show detectable patterns of DCS relative to a nonpolyploid outgroup

I was introduced to this argument during my postdoctoral work with Ken Wolfe: Ken and Kevin Byrne had just completed the Yeast Gene Order Browser (YGOB; [19]), a manually-verified set of homologous genes from many yeast genomes, depicted relative to their orders on their chromosomes (http://ygob.ucd.ie). YGOB illustrates the *double-conserved synteny* (DCS) preserved in the yeast genomes after their shared paleopolyploidy by comparing those genomes to other yeast genomes lacking that polyploidy (Fig. 1). A key aspect of DCS is that it is evident not merely among the genes that survive in duplicate from the polyploidy but also among those genes where one of the two duplicate copies has been lost. In principle, with a sufficiently closely related nonpolyploid relative, a polyploidy event could be identified using DCS even if every single duplicate gene pair created by it had lost one of its members (as shown in the boxed region at the extreme right of Fig. 1).

At the time YGOB was created, yeasts were unusual in that genome sequences for many closely related species were available, whereas, in most other groups of eukaryotes, the sequenced genomes were phylogenetically widely spaced. Using YGOB, the Wolfe laboratory made a number of discoveries about the yeast polyploidy and those genomes more generally. They documented the independent loss of alternate duplicate copies in different lineages [15]: These *reciprocal gene losses* (RGLs) could potentially reproductively isolate the lineages in question from each other [20]. They also described a biosyntentic gene cluster that had recently been "born" in bakers' yeast [21] and showed that the earliest phases of duplicate gene losses after the yeast polyploidy had been quite rapid [15].

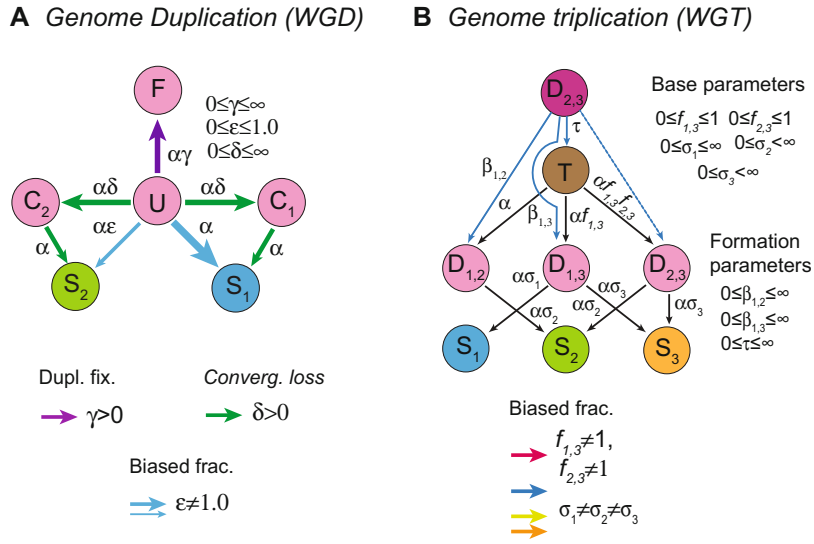

Fig. 2 Models of gene loss after polyploidy for a whole-genome duplication (WGD) and a whole-genome triplication (WGT, hexaploidy). (**a**) WGD: All pillars start in duplicated state **U** (*U*ndifferentiated), from which they can transition to either the three other duplicated states, $\mathbf{C_1}$ (*C*onverging state 1), $\mathbf{C_2}$ (*C*onverging state 2) and **F** (*F*ixed) or to the two single-copy states $\mathbf{S_1}$ (*S*ingle-copy *1*) and $\mathbf{S_2}$ (*S*ingle-copy *2*). We define subgenome 1 ($\mathbf{S_1}$) to be the preferred subgenome with more surviving gene copies: Thus, $\mathbf{C_1}$ and $\mathbf{S_1}$ are states where the genes from the less-fractionated parental subgenome will be or are preserved, and $\mathbf{C_2}$ and $\mathbf{S_2}$ the corresponding states for the more-fractionated parental subgenome. Duplicate fixation is inferred when $\gamma \neq 0$, convergent losses when $\delta \neq 0$, and biased fractionation when $\varepsilon < 1.0$. (**b**) WGT: All pillars start in state **T** (*T*riplicated) and transition first to duplicated states ($\mathbf{D_{x,y}}$) and hence to the single-copy states ($\mathbf{S_x}$). Subgenome 1 is assumed to be favored (fewer losses) and the identity of that subgenome inferred in the POInT computation. Losses from the triplicated state are then increasingly disfavored first to $\mathbf{D_{1,3}}$ (parameter $f_{1,3}$) and $\mathbf{D_{2,3}}$ (parameter $f_{2,3}$). There are also individual rates of loss from the duplicated to single-copy states (σ_x). As described in the text, we add a formation step to this model, with the initial tetraploidy being represented with the $\mathbf{D_{2,3}}$ state, which transitions to state **T** unless losses occur prior to the second allopolyploidy (parameters $\beta_{1,2}$, $\beta_{1,3}$, and τ)

This last question of gene losses reminded me of work by my undergraduate advisor, Paul Lewis, on models for the phylogenetic analysis of nonsequence data [22]. Presented with the comparative data in YGOB, it was not difficult to use its encoded presence/absence data as the states of a phylogenetic model (Fig. 2).

Armed with a simple version of this model, I was able to assist with the next project in the lab: adding a new genome to YGOB. Ken had realized that the genomes that would be most informative as to the history of the yeast polyploidy were not the close relatives of bakers' yeast but rather those like that of *Vanderwaltozyma polyspora* (then known as *Kluyveromyces polysporus*), which had diverged from bakers' yeast shortly after the polyploidy.

It became clear that, because *V. polyspora* and bakers' yeast (*S. cerevisiae*) had diverged so soon after the polyploidy event (Fig. 3), there was a significant challenge in inferring *orthology* for

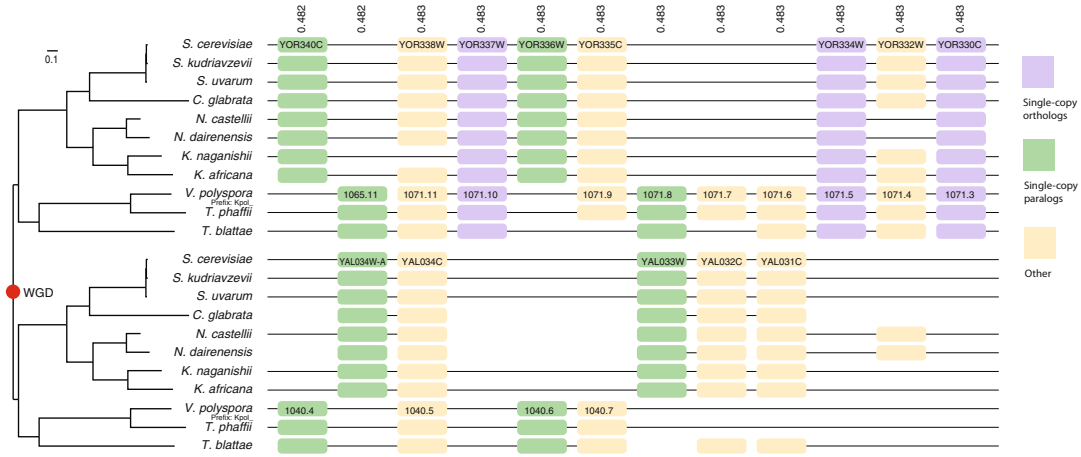

Fig. 3 Inferred orthologous regions from 11 species produced in the yeast WGD event: For clarity, gene names are shown only for *S. cerevisiae* and *V. polyspora*. Pillars shown in purple are cases of single-copy orthologs for these two species: green pillars are single-copy paralogs. Cases where either or both genomes retain homoeologous genes are shown in tan. As discussed in the text, the confidence in the orthology estimate (P_j) is given above each column. Because the yeast WGD has limited evidence for biased fractionation [32], the relationships depicted are degenerate, with the swap of the bottom and top blocks having an identical probability (hence doubling this number gives a sense of the "real" confidence). The mirrored topologies have branch lengths scaled by POInT's estimates of the relative number of duplicate losses ($\alpha \times$ time with a model without biased fractionation, i.e., $\varepsilon = 1.0$ in Fig. 2)

these two genomes. Orthologs, of course, are homologous genes in different species that last shared a common ancestor at the speciation of those species [23]. When two species split almost immediately after a polyploidy, the few *shared* duplicate gene losses that occurred before that split will necessarily produce single-copy orthologs in those species. However, the losses that occur after the speciation will be independent (Fig. 1 illustrates these processes). If we confine ourselves to the genes that are single-copy in both genomes, these independent losses will result in effectively equal numbers of paralogs and orthologs. We show an example of this phenomenon in Fig. 3, where, for the region shown, more of the single-copy genes in *S. cerevisiae* and *V. polyspora* are paralogs (green) than are orthologs (purple).

Under these circumstances, the problem becomes phasing the DCS blocks from *S. cerevisiae* with the orthologous block from *V. polyspora*. YGOB made its orthology phasing by pairing syntenic blocks so as to maximize the number of syntenic positions where all of the polyploid genomes had the same gene status (e.g., present or missing). With this approach, we estimated that about 55% of single-copy genes in *S. cerevisiae* and *V. polyspora* are orthologs and 45% paralogs [24].

This high degree of dissimilarity between the two genomes raises some concerns (to be considered shortly), but we were able to use YGOB's inferences to fit our duplicate loss model and infer

the phylogenetic timing of the gene losses. We found that more than 80% of the duplicate pairs produced by the WGD were still duplicated at the split of *S. cerevisiae* and *V. polyspora*, consistent with the observed near equality in numbers of orthologs and paralogs. Comparing the loss rates to sequence divergence allowed us to show that duplicate loss was rapid in the earliest postpolyploidy period.

3 One Polyploidy or Two?

Thus, the first "proto"-POInT analyses used orthology inferences made heuristically. In principle, such an approach can be quite accurate and could be thought of as analogous to multiple sequence alignment in molecular phylogenetics, where appropriately scaled dissimilarity penalties and hill-climbing methods are used to identify the homologous bases/residues that are then analyzed with probabilistic models of sequence evolution [25–27]. However, there were a few potential concerns. Firstly, could we be sure that *S. cerevisiae* and *V. polyspora* were actually products of the same polyploidy event? A heuristic approach will assume a common polyploidy and produce the syntenic blocks most consistent with that hypothesis. But how could we know that independent polyploidies analyzed with such a tool might not produce a similar 55%/45% split of inferred orthologs to paralogs? The heuristic approach would also work best for large synteny blocks, where even those few shared prespeciation losses will overwhelm any chance effects and give correct orthology phasing. But what is "large" in this context and how serious are such chance effects likely to be? Clearly what was needed is a modeling approach that accounted for the uncertainty in the orthology phasing.

The core of this new model came from adapting an unrelated project from my doctoral work. That project was converting the linkage analysis package Genehunter [28, 29] to run on a supercomputer [30]. Genehunter uses the linkage analysis algorithm of Lander and Green [31] to locate a disease-causing genetic variant in the genome by taking a number of genetic markers from several members of a family afflicted with that disease. The algorithm computes the odds of the disease-causing allele being at each of a number of points within the genetic map, given the known genotypes for each individual at each marker and the recombination rates between those markers. What is important about this algorithm is that while the markers for each individual are known, it is not known whether any given marker *allele* is derived from the individual's mother or father. As a result, the calculation carries this parental uncertainty throughout the analysis, meaning that, given n genotyped individuals, at each marker there are (essentially) 2^n inheritance patterns that must be considered.

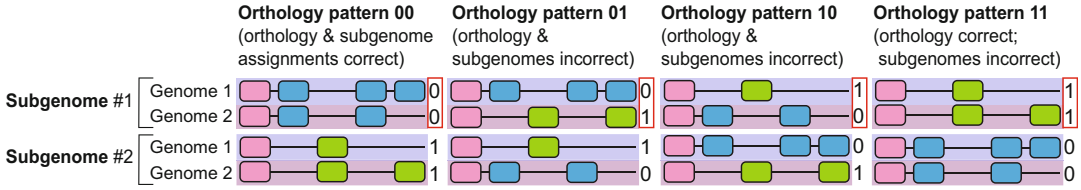

Fig. 4 Four possible orthology relationships (i.e., subgenome assignments) for two genomes sharing a WGD, using the track identifiers (0,1) and colors from Fig. 1, showing how different orthology assignments give rise to different patterns of orthology and paralogy in a column. Hence, the first block depicts the correct orthology taken from Fig. 1, while the remaining three swap subgenome assignments for one or both of the sections of the two extant genomes. POInT computes the likelihood of each and then conditions them on pillars to the left and right (see text)

I realized that the unknown orthology states in the polyploidy model could be treated with the mathematics of the Lander and Green algorithm: It was this insight that made POInT a useful tool. The starting point is the model of duplicate loss along a phylogeny already discussed (Fig. 2). In our original analysis, we assumed that the orthology relationships were known and simply computed the likelihood of the resulting loss data under the model. However, for a given duplicated locus (pillar, Fig. 1), it is perfectly possible to compute a likelihood for any or all possible orthology relationships (Fig. 4). Indeed, for the special case of a duplicate pair fully preserved in each genome, that likelihood is the same for all possible orthology relationships.

4 The POInT Computation

For clarity, let us first assign identifiers 0 and 1 to the two genomic regions produced by the WGD in each of the n genomes considered (Fig. 4). These assignments are purely for bookkeeping purposes: We do not assume that section "0" in one genome is orthologous to section "0" in another. We must compute the likelihood for a set of presence/absence data for every possible combination of orthology relationships: For n genomes, there are 2^n such relationships. POInT can make similar computations for hexaploidies and octoploidies, but the notation is more cumbersome, so I confine my examples here to the case of a genome duplication/tetraploidy. We represent an orthology relationship as a binary vector $\overline{O}$ of length n. For example, for two taxa, the four possible values of $\overline{O}$ are `00` (0 in decimal notation), `01` (1), `10` (2), and `11` (3). The value at the gth position of the vector (0 or 1) indicates which of the two genomic sections is assigned to the "top" track for the gth genome, with the other assigned to the lower track (Fig. 1). When modeling *biased fractionation*, the

unequal preservation of duplicate copies from the two subgenomes, this top track corresponds to the subgenome with more surviving genes [32].

The first step of the computation is, for each of the m pillars in the dataset (the origins of which we will explain briefly in the next section), to compute the likelihood of the presence/absence data at that pillar under each of the 2^n orthology relationships, given a phylogeny and set of model parameters. These likelihoods are stored in a vector $\overline{L}$, indexed by the 2^n orthology relationships in binary format.

The insight of the Lander and Green algorithm comes in at this point. We would like to assess if there is an orthology relationship that has relatively high likelihood for a number of pillars. To do so, we take advantage of the synteny relationships between neighboring pillars, if such relationships exist. We therefore design a transition probability matrix $\Theta^{i-i,i}$ that describes the synteny transition between pillars $i-1$ and pillar i. This matrix gives the probability of orthology relationship j at pillar i given that pillar $i-1$ has orthology relationship k, where both j and k represent orthology relationship vectors when read in binary format. The elements of this matrix $\Theta^{i-1,i}_{j,k}$ have the form:

$$\Theta^{i-1,i}_{j,k} = \prod_{l=0}^{n-1} \theta_{i,l}^{((j\wedge k)\gg l)\&1} \cdot (1-\theta_{i,l})^{1-(((j\wedge k)\gg l)\&1)} \qquad (1)$$

where "^" represents the bit-wise *exclusive-or* operator, "$\gg$" the bit-wise shift operator, and " & " a bit-wise *and*. Hence, as horrifying as this equation looks, it simply implies that we consider the genomes $l = [0\ldots n-1]$ and construct a product that uses $\theta_{i,l}$ at positions where $\overline{O_j}$ and $\overline{O_k}$ differ and $(1-\theta_{i,l})$ at positions where they are the same. Here, $\theta_{i,l}$ is the probability that the orthology assignments change between a pair of pillars for genome l. Its value, however, differs depending on the synteny information for the two pillars. If, between $i-1$ and i, synteny is maintained in either the upper or lower track or both, $\theta_{i,l} = \theta$, a global constant estimated from the data by maximum likelihood. Otherwise, $\theta_{i,l} = 0.5$ for genome l, meaning position $i-1$ conveys no information on the orthology relationships at i. We will return to the interpretation of θ shortly.

We can now construct expressions for our orthology estimates at a position, given the information at one or more other neighboring positions. Recall that $\overline{L}^i$ contains the likelihood of the data at i for each possible orthology state. We now define $\overline{L}^{i-1|D_0\ldots D_{i-1}}$, the likelihood of each possible orthology state at $i-1$, given the data at pillars $0\ldots i-1$. We can write a recurrence equation for $\overline{L}^{i|D_0\ldots D_i}$ using $\overline{L}^{i-1|D_0\ldots D_{i-1}}$, $\overline{L}^i$, and the transition probability matrix $\Theta^{i-1,i}$:

$$\bar{L}^{i|D_0\ldots D_i} = \bar{L}^i \odot \left(\Theta^{i-1,i} \cdot \bar{L}^{i-1|D_0\ldots D_{i-1}}\right) \qquad (2)$$

where "$\odot$" represents an element-wise vector product. When we define the base-case $\overline{L}^{0|D_0} = \overline{L}^0$, we can apply this equation sequentially to the m pillars, with the total likelihood of the data L being

$$L = \sum_{l=0}^{2^n-1} L_l^{m-1|D_0 \ldots D_{m-1}} \tag{3}$$

In other words, the total likelihood is just the sum of the likelihoods of ending up in each of the 2^n orthology states at the last pillar (pillar $m - 1$). Using eq. 3, we optimize the model parameters by maximum likelihood using standard numerical approaches [33]. POInT can either perform this optimization for a user-supplied phylogenetic topology or search across all possible rooted topologies and return the one with the highest likelihood.

To obtain the orthology inferences themselves, as well as associated confidence estimates, we need to add what is known in the HMM literature as a posterior decoding step. This step is analogous to obtaining the odds ratios for the disease allele placement in the Lander and Green algorithm; a similar approach has been applied to estimating correlated rate variation across sites in a sequence alignment [34]. For a given pillar i, we compute, as before $\overline{L}^{i-1|D_0 \ldots D_{i-1}}$. We then compute the analogous probabilities working from the last pillar $m - 1$ down to pillar $i + 1$: $\overline{L}^{i+1|D_{i+1} \ldots D_{m-1}}$ Conceptually, this step corresponds to applying eq. 2 to the reversed sequence of pillars. We then obtain $\overline{L}^{i|D}$, a vector of conditional likelihoods for all of the 2^n orthology states at pillar i, given all of the observed gene presence/absence data, denoted as D:

$$\bar{L}^{i|D} = \left(\bar{L}^{i} \odot \left(\Theta^{i-1,i} \cdot \bar{L}^{i-1|D_0 \ldots D_{i-1}}\right)\right) \odot \left(\Theta^{i,i+1} \cdot \bar{L}^{i+1|D_{i+1} \ldots D_{m-1}}\right) \tag{4}$$

The elements of $\overline{L}^{i|D}$ sum to L, and the conditional probability P_j of any given state j is just the jth element of that vector divided by that total likelihood L:

$$P_j = \frac{L_j^{i|D}}{L} \tag{5}$$

These conditional probabilities give us estimates of our confidence in a given orthology inference. In Fig. 3, the values at the top are such conditional probabilities. Because the yeast WGD is not marked by biased fractionation, the model is completely symmetrical, with states C_1 and C_2 and S_1 and S_2 being equivalent (Fig. 2). As a result, the orthology relationships shown have the same conditional probabilities as those where the top and bottom tracks are swapped, which accounts for the fact that P_j has values near to ½ rather than to 1 for these examples.

4.1 Comments on the POInT Computation

In the next section, I will give some examples that I hope illustrate the value of POInT for analyzing polyploidies. But I think it is first appropriate to note a few weaknesses.

First, POInT is only as accurate as the presence/absence data used as input. We have described a pipeline for producing these pillars for cases when manual sets like those of YGOB are not available [32, 35]. Briefly, each polyploid genome is searched against an outgroup genome lacking the event using BLAST. The resulting homologs are then placed in DCS blocks relative to that outgroup using simulated annealing [36]. The DCS blocks for individual genomes are merged, and a global pillar order that minimizes synteny breaks is inferred and used by POInT. This approach works especially well for more recent polyploidies with relatively few rearrangements observed between the genomes. We have shown that even for the ancient teleost genome duplication, it is adequate [35] but have found it to be problematic for the very old vertebrate 2R events (data not shown).

A second concern is with the θ parameter from eq. 1. Unlike the model parameters in Fig. 2, θ is an error or nuisance parameter in the model without a biological interpretation. Hence, unlike the Lander and Green algorithm, where the corresponding parameter is a recombination rate, θ only indicates how often our model "changes its mind" about the orthology assignments from pillar to pillar. Fortunately, in practice, θ tends to be quite small: In a recent analysis of seven tetraploidies, we found $0.002 \leq \theta \leq 0.009$, meaning that the inferred orthology was almost invariably maintained between adjacent pillars. It is hence a useful diagnostic tool: If a large (>0.1) value of θ is inferred, the resulting orthology inferences are of low quality.

4.2 Example Uses of POInT

Testing for a shared polyploidy. One of the first uses we made of POInT was testing whether the polyploidies found in *V. polyspora* and *S. cerevisiae* were in fact a single shared event. In our modeling framework, two independent polyploidies can be easily represented as a phylogeny like that in Fig. 3 where the root branch has zero length. In that case, the clades on either side of that root have no common losses from their respective polyploidies. The question of how many *apparently* shared losses we would infer in the case of independent polyploidies can be addressed with simulation: giving POInT a phylogeny with a zero-length root and set of model parameters and asking it to simulate new polyploid genomes under that model. For each simulation, we ask how much smaller the likelihood seen under a model with a forced zero length root (an independent polyploidy model) is than is the likelihood of a model where the root length can take on any value (e.g., a model of a shared polyploidy; [37]). We then ask whether the difference in these two likelihoods for the real data is much greater than the differences seen among the simulations. A large improvement in

likelihood for the shared polyploidy model relative to the independent one when analyzing the real data as compared to analyzing the simulations suggests significant evidence of a shared polyploidy between these lineages, and was indeed what we found [37]. We conducted an identical analysis more recently to show that three species of parasitic nematodes also descend from a common triploid ancestor [38].

Confirming the existence of biased fractionation. As with the question of the shared *S. cerevisiae/V. polyspora* genome duplication, inferences of biased fractionation have the potential to be influenced by the methods used to detect them. Hence, while biased fractionation is inferred when heuristic approaches for orthology inference are used on polyploid genomes [39, 40], there remained the slight concern that such methods, in seeking to maximize genomic similarity, might infer a bias in gene losses where none exists. Our models incorporate bias in fractionation with model parameters such as ε in Fig. 2a. Using likelihood ratio tests [41], we initially showed [32] that models with biased fractionation ($\varepsilon < 1.0$) were preferred over those without it ($\varepsilon = 1.0$). However, one might still argue that the maximization approach inherent in making these parameter estimates was giving rise to spurious inferences of bias. In the end, I again adopted a simulation approach, creating simulated polyploidies without biased fractionation and showing that analyzing those simulated genomes using models that included biased fractionation did not give estimates of the ε parameter that were anything like as small (e.g., high bias) as was observed for real data [35]. We can therefore be quite certain that biased fractionation is a common pattern in paleopolyploid genomes [32, 35, 38, 42].

Modeling hexaploidy. We recently completed an analysis of the hexaploidy shared by the *Brassica* crops *B. oleracea* (broccoli, cabbage, cauliflower) and *B. rapa* (kale) and their relatives [42]. The 3:1 pattern of conserved synteny between these genomes and that of *Arabidopsis* is clear [11], but there is not an obvious single mutational step that could produce such hexaploids. Instead, it is believed that they were formed when two diploids hybridized to form an allotetraploid, which subsequently underwent an allopolyploid hybridization with another diploid to form an allohexaploid [43]. The extant genomes show signs of strong biased fractionation, such that one of the three subgenomes shows many more surviving genes (the LF, or "least fractionated" subgenome) than the other two (MF1 and MF2, "more" and "most" fractionated, respectively). As a result, it was argued that the progenitor of the LF genome was most likely the "last arriving," i.e., the diploid progenitor contributing to the second allopolyploidization. We developed a new type of model for a two-step hexaploidy (Fig. 2b) that

includes the transition from a genome duplication (state $D_{2,3}$) to a triplication (state T) prior to gene losses and speciation events after the second allopolyploidy [42]. If the two allopolyploidies were separated in time, not all of the initially duplicated genes would have survived in duplicate to become triplicated. Denoting LF as subgenome 1, we can model this effect with transitions from state $D_{2,3}$ to $D_{1,2}$ (loss of a duplicate from MF2 prior to LF arriving) or from $D_{2,3}$ to $D_{1,3}$ (loss from MF1). Because all of these transitions occurred prior to the first speciation in our data, we cannot estimate whether any given pillar underwent a loss prior to the arrival of LF. However, treating the pillars as a whole allows us to estimate the frequency of such losses, as well as to compare our inferences to models where the LF subgenome was one of the original progenitors, rather than the last arriving. The POInT framework gives us strong confidence both in there having been a number of gene losses prior to the arrival of LF and that LF was indeed the last arriving subgenome, confirming the two-step model for this event.

5 Future Directions and Concluding Remarks

POInT is a complex and slightly idiosyncratic program, and running a full analysis with it is unnecessary for most users. Instead, we have made our inferences from POInT publicly available in a graphical format with $POInT_{browse}$ (wgd.statgen.ncsu.edu). Here, users can first visually explore the DCS blocks and orthology inferences and then download the associated coding regions and inferred gene trees, as well as gain access to larger datasets from our analyses. Our main future goal for this portal is to add a "batch download" feature that will allow users to request sets of pillars that meet specified criteria, such as orthology confidence and number of surviving homoeologs. The synteny visualizations can also be downloaded in publication quality: such was the origin of Fig. 3.

We also plan to further extend POInT to handle nested polyploidies, using the framework of Fig. 2b. While computationally expensive, such models could allow the analysis of, for instance, the recent soybean polyploidy simultaneously with the more ancient shared legume WGD. With such an addition, as well as a few other polyploidies we have not yet considered, such as that in the Solanaceae [44], we will be able to provide something close to a comprehensive reference for paleopolyploidies in the sequenced genomes. From there, questions such as the degree of convergent evolution inherent to polyploidy, the sources of biases in duplicate losses and the timing of losses after hybridization can be addressed in detail using POInT's inferences as a starting point (pun intended).

Acknowledgments

I would like to thank J. Thorne and A. Zwaenepoel for assistance with clarifying the mathematical description of POInT's computations.

References

1. Simunek M, Hoßfeld U, Wissemann V (2011) "Rediscovery" revised–the cooperation of Erich and Armin von Tschermak-Seysenegg in the context of the "rediscovery" of Mendel's laws in 1899–1901. Plant Biol 13(6):835–841
2. Clausen R, Goodspeed T (1925) Interspecific hybridization in Nicotiana. II. A tetraploid glutinosa-tabacum hybrid, an experimental verification of Winge's hypothesis. Genetics 10(3):278
3. Kuwada Y (1911) Maiosis in the pollen mother cells of Zea Mays L. (with plate V.). 植物学雑誌 25(294):163–181
4. Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615–643
5. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274(5287):546, 563–567
6. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387(6634):708–713
7. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333
8. Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat Rev Genet 18(7):411–424
9. Hilu K (1993) Polyploidy and the evolution of domesticated plants. Am J Bot 80(12): 1494–1499
10. Mandáková T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA (2010) Fast diploidization in close mesopolyploid relatives of Arabidopsis. Plant Cell 22(7):2277–2290
11. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Yu J, Meng J, Min J, Poulain J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Z, Li Z, Xiong Z, Zhang Z (2011) The genome of the mesopolyploid crop species Brassica rapa. Nat Genet 43(10):1035–1039
12. Li W-H (1980) Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fish. Genetics 95: 237–258
13. Nei M, Roychoudhury AK (1973) Probability of fixation of nonfunctional genes at duplicate loci. Am Nat 107:362–372
14. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155
15. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341–345
16. Walsh B (2003) Population-genetic models of the fates of duplicate genes. Genetica 118(2–3):279–294
17. Li W-H (1997) *Molecular evolution*. Sinauer Associates, Sunderland
18. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. Brief Bioinform 12(5):379–391
19. Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15(10):1456–1461
20. Werth CR, Windham MD (1991) A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. Am Nat 137(4): 515–526

21. Wong S, Wolfe KH (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet 37(7):777
22. Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol 50:913–925
23. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338
24. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A 104:8397–8402
25. Henikoff S, Henikoff JG (1992) Amino-acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22): 10915–10919
26. Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205–217
27. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
28. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
29. Kruglyak L, Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. J Comput Biol 5(1):1–7
30. Conant GC, Plimpton SJ, Old W, Wagner A, Fain PR, Pacheco TR, Heffelfinger G (2003) Parallel Genehunter: implementation of a linkage analysis package for distributed-memory architectures. J Parallel Distrib Comput 63: 674–682
31. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84:2363–2367
32. Emery M, Willis MMS, Hao Y, Barry K, Oakgrove K, Peng Y, Schmutz J, Lyons E, Pires JC, Edger PP, Conant GC (2018) Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. PLoS Genet 14(3): e1007267em
33. Press WH, Teukolsky SA, Vetterling WA, Flannery BP (1992) Numerical recipes in C. Cambridge University Press, New York
34. Felsenstein J, Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13(1): 93–104
35. Conant GC (2020) The lasting after-effects of an ancient polyploidy on the genomes of teleosts. PLoS One 15(4):e0231356
36. Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
37. Conant GC, Wolfe KH (2008) Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. Genetics 179:1681–1692
38. Schoonmaker A, Hao Y, Bird D, Conant GC (2020) A single, shared triploidy in three species of parasitic nematodes. G3 10:225–233
39. Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16:805–814
40. Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A 108(10):4069–4074
41. Sokal RR, Rohlf FJ (1995) Biometry, 3rd edn. W. H. Freeman and Company, New York
42. Hao Y, Mabry ME, Edger P, Freeling M, Zheng C, Jin L, VanBuren R, Colle M, An H, Abrahams RS, Washburn JD, Qi X, Barry K, Daum C, Shu S, Schmutz J, Sankoff D, Barker MS, Lyons E, Pires JC, Conant GC (2021) The contributions of the allopolyploid parents of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. Genome Res 31:799–810
43. Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC (2012) Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190(4):1563–1574
44. Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry CS, Bliek M, Boersma MR, Borghi L, Bruggmann R (2016) Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. Nat Plants 2(6):1–9