Resurrected protein interaction networks reveal the strong rewiring that leads to network organisation after whole genome duplication

Zhicheng Zhang [1¶], Heleen Coenen[1¶], Philip Ruelens[1¶], Rashmi R. Hazarika[2], Tareq Al Hindi[1], Georgianna K. Oguis[1], Vera van Noort[2], Koen Geuten[1*]

[1] Department of Biology, Division of Molecular physiology of Plants and Micro-organisms, Section of Molecular Plant Biology, University of Leuven, Leuven, Belgium.

[2] Department of Microbial and Molecular Systems, Division of Microbial and Plant Genetics, University of Leuven, Leuven, Belgium.

* Corresponding author

E-mail: koen.geuten@kuleuven.be

[¶] These authors contributed equally to this work

Abstract

The evolution of plant genomes is characterized by several rounds of polyploidization or ancient whole genome duplication. While the consequences of these major events for genome structure and transcriptome expression have been investigated, the effects at the protein level remain unknown and yet will be functionally important. To understand how a plant protein-protein interaction network organizes itself after whole genome duplication, we studied the evolution of MADS-domain transcription factors. We accurately inferred, resurrected and tested the interactions of their ancestral proteins before and after the gamma triplication at the origin of core eudicots and directly compare these ancestral networks to the networks of *Arabidopsis* and tomato. We find that the gamma triplication generated a network constrained in size and saturated in possible number of interactions, which strongly rewired by the addition of many new interactions. The new interactions are surprisingly often established with related proteins, something we call neo-redundancy. The evolved networks are organized around hubs and into modules. The direct observation of preferential attachment of existing interactions to hubs through gene duplication explains the scale-free organization. The evolutionary optimal modular organization is favored by the addition of new interactions to the network and by the avoidance of mis-interactions, as shown by simulations. The resurrection of ancestral networks and the direct observation of ancestral rewiring events allowed us to elucidate the role of whole genome triplication, elementary processes and evolutionary mechanisms in the origin of a biological network.

Author summary

The complement of all DNA in an organism is called the genome, while the complement of all proteins is called the proteome. The cooperation between proteins in a proteome is organised in particular ways of which the origin and significance is not fully understood. Plants and — to a lesser extent — animals have undergone ancient and recent duplications of their genomes and corresponding proteomes in the course of evolutionary history. These are considered important events that may have coincided with the origin of important groups of species. However the consequences of such duplications for the organism are not well understood. Here we study ancient and current sample proteomes by resurrecting the ancestral proteins and comparing how they interacted with other proteins millions of years ago and now. We find that a genome duplication dramatically reorganised the proteome and the proteome acquired its organisation by the addition of new interactions and avoiding mis-interactions.

Introduction

The field of interactomics promises to generate insight in the molecular modules that establish the functions of genomes. One intensively studied type of interactome is the protein-protein interaction network (PIN) and such large-scale networks have been described for several model organisms, allowing the comparison of functional networks between species [1,2]. The challenges in interactomics are orders of magnitude larger than for genomics. While the scale of the problem is in principle smaller because a genome produces many more regulatory molecules than only proteins, the experimental determination of protein interactions is more tedious and different methods are not always consistent [3–5]. The co-localization of proteins varies between cell types and between stages of development calling for differential or comparative interactomics [6,7]. Moreover, the effective concentrations of proteins vary, which together with their structurally determined capacities to establish interaction affinities influences whether or not proteins interact *in vivo* [8]. Despite these challenges, interactomics revealed that biological networks, like man-made networks, exhibit complex modular topological organisation [9,10], which is fascinating and evokes questions about its origin and reason of existence. These organisations can be mathematically described by identifying a network as *scale-free*, *small world* or *hierarchical* [11] (Box 1). While it might be clear which generative processes of such networks can theoretically explain their topology, these mechanisms have not been directly observed and the evolutionary dynamics leading to such modular networks in biological networks have not been experimentally investigated because they occur on an evolutionary time scale. One elementary process in network dynamics to achieve scale-free networks is preferential attachment [11,12]. In the context of biological networks, preferential attachment posits that proteins with more interacting partners acquire new partners faster [11]. This is very plausible when considering the important role of gene duplication in evolution: duplication of a well-connected protein will result in many more new duplicate edges than duplication of a relatively isolated protein. Specifically in plants, but of general biological importance, the role of massively concerted gene duplication at the genome level is well documented [13,14]. Such whole-genome duplication events could also have a major

effect on the rewiring of protein interaction networks as predicted by the duplication-divergence model [15].

Phylogenetic and synteny analyses of MADS-box genes in combination with comparative analysis of transcriptomes placed an ancient whole genome triplication named the γ triplication, in the stem lineage of core eudicots [16]. The γ triplication was estimated to have occurred at 117 Mya, predating the divergence of Gunnerales yet before the divergence of Buxales [16] (Fig 1). The origin of core eudicots marks an important event in plant evolution as today the core eudicots comprise approximately 75% of all species of flowering plants [17]. The duplication patterns of MADS-domain proteins — a conserved class of transcription factors that act as key regulators of reproductive development in flowering plants — indicate that many gene lineages present in extant core eudicots derive from this whole genome triplication with most being retained in duplicate or triplicate copies after the triplication event [16]. Their molecular function requires them to be nuclear localised and form specific multimeric transcriptional complexes to regulate their downstream targets [18]. Considering the critical role of the specific protein binding affinities among these floral homeotic proteins in floral organ specification [19,20], the expansion of MADS-box genes may have played an important role in the origin of core eudicot floral apomorphies.

Here we study the intricate ancestral protein interaction network of MIKC^{c} MADS-domain transcription factors. We resurrected ancestral proteins immediately before this genome triplication and 10 million years later, at the diversification of rosid and asterid flowering plants. By directly comparing ancestral networks with extant networks from *Arabidopsis* and tomato in a single experimental setup, we are able to go beyond theoretical models and comparative analysis of present-day networks and instead pinpoint directly how this network diverged and which processes were responsible for its origin and divergence.

BOX 1 | Topological concepts of protein-protein interaction network

Scale-freeness: In network theory, the degree k of a node represents the number of edges that connect this node with other nodes. The degree distribution $P(k)$ describes the probability of a randomly selected node having degree k [11]. In a scale-free network, the probability $P(k)$ that a randomly selected node with degree k follows a scale-free or power-law distribution is proportional to k^{γ}, where γ is the degree exponent. A scale-free PIN is formed by many proteins with few interactions and a few highly connected proteins or hubs which may play a significant role in the network.

Hierarchy: Apart from scale-freeness, many biological networks also display a high degree of inherent modularity or clustering corresponding to the interconnected sub-networks between nodes, which makes the global network hierarchical [10,12,21]. The clustering coefficient C_k is used to quantify the extent of clustering for each node in a PIN. In a strongly hierarchical network, the distribution of the clustering coefficients of a network $C(k)$ follows $k^{\square\beta}$ where the scaling exponent β equals 1 [12].

Small world networks: Numerous real life networks, including biological networks ranging from metabolic networks to protein interaction networks, were observed possessing the small world property [10,21–23]. The small world property refers to the characteristic that most nodes can be reached from any other randomly chosen node by a small number of intermediate nodes, comparable to random networks of the same size and average degree. Yet unlike random networks, real life networks with the small world property typically have a larger clustering coefficient as their respective random networks and are referred to as small world networks.

Fig 1. Evolution of selected MADS-box proteins of interest. (A) Simplified phylogenetic tree of eudicots. The positions at which protein interaction networks were inferred are indicated; (1) PIN just before γ triplication or Pre-PIN, (2) PIN at the Asterid-Rosid split or Post-PIN, (3) PIN of *Arabidopsis thaliana* (Ara-PIN) and (4) PIN of *Solanum lycopersicum* (Sol-PIN). (B) Phylogenetic relationships between ancestral and extant proteins present in each reconstructed PIN. Reconstructed ancestral proteins are named after their *Arabidopsis* or tomato descendants preceded by 'anc' for ancestral. Greyed out proteins were not included in this study. White circles indicate the absence of a specific protein due to gene loss. Abbreviations: AP: APETALLA. FUL: FRUITFULL. FUL-LIKE:

FRUITFULL-like gene. CAL: CAULIFLOWER. MC: MACROCALYX. SLMBP: *Solanum lycopersicum* MADS-box protein. DEF: DEFICIENS. TM: Tomato MADS-box gene. TAP: Tomato APETALLA. GLO: GLOBOSA. PI: PISTILLATA. AG: AGAMOUS. SHP: SHATTERPROOF. TAG: Tomato AGAMOUS. TAGL: Tomato AGAMOUS-like gene. STK: SEEDSTICK. SEP: SEPALLATA. FBP: FLORAL-BINDING PROTEIN. SVP: SHORT VEGETATIVE PHASE. SOC1: SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1.

Results

Accurately resurrected ancestral MADS-domain proteins reveal the origin of extant networks

To understand the mechanisms driving evolution and innovation of protein-protein interactions (PPIs) in the context of networks, we analyzed the interaction patterns of MADS-box proteins at ancestral and extant nodes along the angiosperm phylogeny. We aimed to observe changes in these protein interaction networks across several hundred million years of evolution and following the hexaploidization event at the origin of core eudicots. More specifically, we reconstructed protein-interaction networks (PINs) between representatives of nine MADS-box gene subfamilies at three distinct points in time for a total of four PINs; (1) just before the γ triplication event coinciding with the origin of the core-eudicots, (2) following the γ triplication at the Asterid-Rosid split and present-day (3) from *Arabidopsis thaliana* and (4) *Solanum lycopersicum* (tomato) (Fig 1). These ancestral and extant interaction networks are respectively referred to as Pre-PIN, Post-PIN, Ara-PIN and Sol-PIN throughout this study.

Reconstruction of ancestral proteins that form the Pre-PIN and Post-PIN was performed by inferring the maximum-likelihood protein sequences at the ancestral nodes of interest for each subfamily separately (see methods, S1 Fig). MADS-box genes are a good model system for ancestral sequence resurrection since there are many sequences available throughout the angiosperm phylogeny that are well conserved within their subfamilies. The accuracy of reconstructed ancestral protein sequences is

represented by the posterior probability of each inferred amino acid in the ancestral sequence (S2 Fig). Both the ancestral proteins before and after γ triplication were reconstructed with on average 92.8% and 94.6% of sites obtaining a posterior probability higher than 0.95. Previous studies utilizing ancestral proteins to characterize evolutionary transitions defined ambiguously reconstructed sites as those sites for which the most likely amino acid has a posterior probability lower than 0.80 and that have an alternative amino acid state with a posterior probability higher than 0.2 [24–27]. Surveying ambiguous sites in the ancestral proteins reconstructed in this study revealed that ambiguously inferred sites were mostly located outside of the highly conserved K-domain (S2 Fig), which plays a prominent role in interactions between MADS-box proteins [28,29]. Out of the 26 reconstructed ancestral proteins, only 11 ambiguous sites in the K-domain had an alternative amino acid not biochemically similar to the most likely amino acid. Given the scale of our study, this represents only 0.5% of all reconstructed sites in the K-domain. Following inference, codon optimized ancestral DNA sequences were synthesized and, analogous to their *Arabidopsis* and *Solanum* descendants, cloned into yeast two-hybrid (Y2H) expression vectors. Subsequently, all pairwise interactions for each set of MADS-box protein constructs at an ancestral or extant node were determined using a high-throughput yeast two-hybrid system. In total 582 binary protein-protein interactions were tested (Fig 2A).

As yeast two-hybrid is prone to false positive results, we first determined the reliability of the Y2H system used in this study and consequently our interaction data. In the absence of a curated interaction data set for MADS-box proteins, we compared Ara-PIN and Sol-PIN to *Arabidopsis* and *Solanum* MADS-box protein interaction networks described in the literature (S3 Fig) [30,31]. Our analysis shows a high overlap between Ara-PIN and Sol-PIN described here and previously constructed networks. We determined the accuracy or similarity of Ara-PIN and Sol-PIN to be 0.80 and 0.74 respectively. The overall similarity between both networks is 0.77 (S3 Fig). The discrepancy between studies is likely caused by the different two-hybrid systems utilized, e.g. different yeast strains and/or expression vectors [32]. Yet, the overall high similarity between Ara-PIN or Sol-PIN and previously described *Arabidopsis* and *Solanum* MADS-box protein interaction networks strengthens confidence in the ancestral Pre-PIN and Post-PIN networks.

Fig 2. Evolution of MADS-box protein interaction networks reveals accelerated rewiring following γ triplication. (A) MADS-box protein interaction networks just before the γ triplication (1), at the Asterid-Rosid split (2), for *Arabidopsis* (3) and for *Solanum* (4) as assayed by Y2H. Same color nodes are coding proteins following the same MADS-box gene lineage mentioned in Fig 1B. Conserved and gained protein interactions are indicated as black and green solid lines, respectively and are determined by comparing Post-PIN to Pre-PIN and Ara-PIN or Sol-PIN to Post-PIN. (B) shows a number of network topological measurements for reconstructed Y2H MADS-box PINs. Values between parentheses denote the average network topological measurements of 10,000 randomly generated networks of same network size and average degree. (C) The rate of interaction gains and losses as determined from Pre-PIN to Post-PIN and from Post-PIN to Ara-PIN or Sol-PIN for Y2H. Gain and loss rate was defined as gained or lost edges divided by number of potential interactions (self-loops included) times the divergence time.

Constrained network growth and increased rewiring rate after genome triplication

One important force driving the evolution of protein interaction networks is the introduction of new nodes through duplication, either whole-genome or segmental duplications [15]. Especially throughout plant evolution, series of whole-genome duplications have expanded the number of agents in major regulatory networks including the MADS-box gene family [16,33,34]. Following the γ triplication event, theoretically all genes were triplicated, however, quickly afterwards redundant genes would be silenced and lost through pseudogenization [17]. Indeed, from Pre-PIN to Post-PIN, not all genes retained three copies, generating an ancestral Post-PIN with a network size only two fold larger than the original Pre-PIN [33]. For proteins that function in multimeric complexes, gene retention after whole genome duplications is often explained by the gene dosage balance hypothesis, which states that specific types of interacting proteins, such as transcription factors, are retained in similar dosage not to disturb dosage-sensitive processes [35]. Since our approach provides us with ancestral interactions, it allows us to directly test this hypothesis. To test for dosage balance, we plotted post-γ gene dosages of the proteins that interacted before γ by Y2H (S4A Fig). In line with previous findings

[36], we did not observe any evidence for dosage-balance in MADS-networks as only two out of nine proteins (22%) are retained in balance. After the Asterid-Rosid split, two additional rounds of ancient whole genome duplications occurred along the lineage towards Arabidopsis (β, 77.5 Mya and α, 23.3 Mya) and one ancient whole genome triplication towards *Solanum* (T, 71 Mya) [37,38]. Here, the expansion of the network size was limited as only 22 proteins are present in Ara-PIN and 23 in Sol-PIN as compared to 19 proteins in the Post-PIN (Fig 1B), illustrating that unlike the γ triplication, the more recent ancient whole genome duplications did not result in the expansion of the MADS-box PIN. This suggests that MADS-box PIN growth through the addition of proteins was constrained after the γ triplication in the sense that a maximal network size may exist that is established by the physical properties of the proteins.

Duplication is not the only force driving changes of PINs. Also edge rewiring results in the diversification of networks and rewiring is thought to occur following duplication events as proposed by the duplication-divergence model [39–42]. Surprisingly, network density, i.e. the ratio between the number of actual observed interactions and the number of potential interactions, did not change notably from Pre-PIN to the current PINs, Ara-PIN and Sol-PIN (Fig 2B). Although the network did not become more compact, our results do show dynamic interaction rewiring with numerous interactions being gained and lost (Fig 2A). To investigate whether edge rewiring happened at a faster rate following gene duplication, we calculated and compared the average rate of interaction gain and loss from Pre-PIN to Post-PIN with the rates from Post-PIN to Ara-PIN and Post-PIN to Sol-PIN (Fig 2C). Our results show that from Pre- to Post-PIN the rate in interaction gain is 1.5E-02 gained edges per total possible edges per myr, approximately 1.5 fold higher than the rate in interaction loss (1.0E-02/edge/myr). In addition, from Post-PIN to Ara- and Sol-PIN new interactions evolved at a rate of 1.4E-03 and 1.1E-03/edge/myr respectively, while interactions were lost at 2.1E-03 and 1.2E-03/edge/myr, respectively. These results indicate that following the γ triplication the MADS-box network rewired at a rate approximately tenfold higher than between Post-PIN and Ara-PIN or Sol-PIN. Moreover, from the origin of core eudicots up to the Asterid-Rosid split, the network rewired mainly through the gain of new interactions. By contrast, from the Asterid-Rosid split until present-

day, not only did the overall rewiring rate decrease, interaction loss became higher than interaction gain (Fig 2C). Together, these results indicate that shortly following the γ triplication, the MADS-box PIN underwent extensive rewiring in agreement with the duplication-divergence model [39,42].

The ancestral protein interaction networks before and after the γ triplication also allow us to directly test specific duplication-divergence models, such as neofunctionalization and subfunctionalization. By comparing the interaction patterns of ancestral Pre-PIN proteins to the interaction patterns of their duplicated or triplicated descendants, we can distinguish between these models [43]. The neofunctionalization model proposes that following gene duplication, one paralog retains all interactions while the other is released from functional constraints and undergoes rapid diversification resulting in the evolution of novel interactions. The subfunctionalization model posits that paralogous proteins rewire by redistributing their ancestral interactions among the different paralogs without new interactions emerging. In agreement with the rapid rewiring after the γ triplication, our data show many more instances of neofunctionalization than of subfunctionalization when comparing Pre-PIN to Post-PIN or Post-PIN to Ara-PIN and Sol-PIN. Rewiring following the γ triplication, however, can not be explained by a strict interpretation of sub- or neofunctionalization (S4B Fig). Rather, the data shows either rapid rewiring of all descendant paralogs by acquiring novel interactions or a combination of sub- and neofunctionalization, in which novel paralogs both acquire new interactions while maintaining a set of ancestral interactions. This corresponds to Voordeckers et al. [24] where they find that the evolution of MALS genes in yeast could also be explained by a combination of the two rather than by strict sub- or neofunctionalization. Interestingly, we also observe many cases in which redundancy originated through the γ triplication; i.e two newly emerged paralogs interact with the same protein while their ancestor did not. This rather surprising observation, that we refer to as neo-redundancy, might be explained by the fact that new paralogs are highly similar and as a consequence a protein evolving to interact with one of these paralogs will likely also interact with the other paralog.

Modular organisation in reconstructed MADS-box networks

Although the inferred PINs are small, our method provides an unprecedented opportunity to explore the evolution of modular organisation through time and between ancestral and descendant networks instead of horizontal comparisons between extant species. To this end, we calculated a number of topological network parameters commonly used to describe real-world networks for the inferred MADS-box PINs. For comparison, we also determined these parameters for random networks of equivalent size and average degree (Fig 2B). A common characteristic of many real networks is that they tend to form clusters or modules. The extent to which a network consists of clusters is described by the average clustering coefficient of a network, which for most real-world networks is higher than comparable random networks and indicates their modular structure [44]. For the MADS-box PINs, all Y2H networks have a higher average clustering coefficient than their corresponding random networks (Fig 2B). Moreover, the average clustering coefficient increases following the γ triplication from $C_{Pre\text{-}PIN}$ = 0.311 to $C_{Post\text{-}PIN}$ = 0.555, while both descendant networks, Ara-PIN and Sol-PIN, again have a lower clustering coefficient of $C_{Ara\text{-}PIN}$ = 0.378 and $C_{Sol\text{-}PIN}$ = 0.281.

In addition to a high clustering coefficient, real-world networks also have a short average path length. The average shortest path length is defined as the average minimal number of edges that connect all possible pairs of nodes in a network. A short average path length ensures an efficient and fast transmission of information throughout the network [11,22]. Networks that have a higher clustering coefficient than a comparable random network, yet also have an average shortest path length similar to random networks are referred to as small world networks (Box 1). The average shortest path length of each Y2H MADS-box PIN was consistently smaller, albeit only slightly, than their random networks and remained relatively stable throughout their evolution (Fig 2B). Together with their high clustering coefficient, this indicates that all Y2H PINs meet the requirements of small-world networks.

Another important characteristic of most real networks is their heterogeneous degree distribution. A highly heterogeneous degree distribution is suggestive for the presence of hubs in a network as some nodes or hubs will have many connections while others have only a few connections. By contrast, in random networks all nodes have approximately the same number of connections with only a small

deviation [44]. As compared to random networks, all MADS-box PINs exhibit a high network heterogeneity suggesting the presence of hubs (Fig 2B). Indeed, in Pre- and Post-PIN more than 40% of all connections involve SEP3-like proteins with a degree of 7 and 16 approximately three-fold the networks' average degree of 2.6 and 4.7 respectively. However, both in Ara-PIN and Sol-PIN, there is no single prominent node with a high maximum degree. Rather, in the latter two networks, multiple proteins exhibit a moderately high degree between 8 and 12 only two-fold the average degree, which are FUL, AP1, SEP3, SOC1 and SVP in Ara-PIN and JOINTLESS, TM5, RIN and SLMPB21 in Sol-PIN. SEP3 therefore lost its prominence as the sole hub protein because multiple hubs originated in more recent time. These novel intermediary hubs might represent novel genetic signal integrators required to specify development of more complex derived morphologies associated with core eudicots

Parameterized simulations reveal scale-free and hierarchically modular networks generated through edge addition

While we observe characteristics of modular organisation in our Y2H networks (Fig 2B), their relatively small size does not allow for a statistical analysis of their potential scale-free and hierarchical network topologies. We therefore simulated large-scale networks for all actual Y2H networks in which the observed network properties and dynamics are applied. The simulation procedure is schematised in Fig 3A. Based on this procedure, we implemented separate modifications to understand the underlying elementary processes that drove the evolution of the MADS-box PIN and contributed to the organisation the networks acquired. For each simulation run, we carried out a series of at least 10 stochastic simulations with an initial network size of 1,000 nodes, simulated from the actual Pre-PIN (see Methods).

In the first simulation of the initial upscaled Pre-PIN, an approximately 2.7 fold edge addition frequency (73%) compared to edge deletion frequency (27%), measured from the actual Pre-PIN to Post-PIN, facilitated the generation of numerous novel clusters in the simulated Post-PIN, while not many such newly created clusters could be eliminated. A significant negative linear correlation between clustering coefficient $C(k)$ and degree k was observed from upscaled Pre-PIN to simulated

Post-PIN (Fig 3C, (ii) to (iv)). Such hierarchy seems subsequently to be retained from simulated Post-PIN to simulated Ara-PIN (Fig 3C, (v) to (viii)) and simulated Sol-PIN (Fig 3C, (ix) to (x)) even though the edge deletion frequency was found 2 and 1.4 fold higher than the edge addition frequency respectively (Fig 3A). Upon fitting the scaling law, a high scaling β exponent was obtained in our simulation (Fig 3C), which is consistent with earlier studies where several biological networks such as yeast PINs have exhibited higher scaling exponents [45].

To understand whether evolution can generate network hierarchy *de novo* or is dependent on a pre-organised network, we initialised our simulations using a homogeneous random network (initial Erdős–Rényi (ER) random Pre-PIN model, S5A Fig). Application of the measured node and edge dynamics to such a homogeneous random network did not yield a good fitting negative linear correlation between the variables $C(k)$ and k in the simulated extant networks (S5B Fig) and showed relatively lower β exponent values (S6D-6H Fig) compared to these from the initial upscaled Pre-PIN model. This implies that a scale-free heterogenous initialization network with its organisation was necessary for the sustenance of network modularities. We also traced the degree distribution $P(k)$ at each stage for both models. In the initial upscaled Pre-PIN model, we found that there was a gradual transition from a pure Power Law degree $P(k)$ distribution in the simulated Pre-PIN to a bell shaped distribution with a tail in the simulated Post-, Ara- and Sol-PIN (Fig 3D). By contrast, when the initialization network was random, the observed Poisson distribution at the beginning was retained in all descendant simulated networks (S6B Fig).

We subsequently evaluated the role of the γ triplication by constructing networks devoid of this major event. In this simulation the network size from Pre-PIN to Post-PIN was retained at 1,000. Here, edge dynamics created clusters in the simulated networks of Post-, Ara- and Sol-PIN but no significant negative linear $C(k)$ distribution was observed and the values of scaling β exponent dropped close to or even beneath zero (S6C-6G Fig). Indeed, the effect of edge dynamics was too drastic on such a reduced network size. This suggests that an adequate initial network size is crucial for edge dynamics

to operate and retain modular structures. This also identifies the γ triplication as the key event that increased the number of nodes to facilitate the formation of new modules through edge dynamics.

Finally, we applied discrete edge deletion and addition frequencies via pure random attachment to the simulated Pre-PIN and examined which elementary process played a dominant role in driving the network evolution towards the hierarchical topology as observed in the simulation of the initial upscaled Pre-PIN model. In a first simulation, the edge deletions preceded the edge additions and in a second one, this order was reversed. This illustrates that edge addition played the dominant role in creating modularity in the simulated Post-PIN (S6I-6J Fig) and such dominance of edge addition is not dependant on the relatively higher edge addition frequency because the hierarchical organisation could already be observed in the simulated Post-PIN when the frequency ratio (edge addition frequency to edge deletion frequency) was 40 to 60 (S6K-6L Fig). Edge additions increase the *C(k)* of a node which in turn increases the overall *C(k)* of the whole network (S6A Fig). In our simulation, random addition of edges drove the highly abundant low degree nodes to acquire more links than the lowly abundant hubs, resulting in the low degree nodes to become parts of a highly clustered neighbourhood (hub-like high degree intermediates). The network visually emerged as a relatively complex mesh of interrelated modules which were maintained by several high degree nodes (hubs) from simulated Post-PIN to simulated extant Ara-PIN and Sol-PIN (Fig 3B, (iv), (viii) and (x)) [46]. Higher edge deletion frequencies which occurred from simulated Post-PIN to simulated extant Ara- and Sol-PIN lead to hubs getting eroded of their existing links constantly as a result, the hubs did not gain as many interactions as compared to the low degree nodes which rapidly acquired more edges than in the simulated Post-PIN. The simulations illustrate that edge addition has led to an increase in the degree of low degree nodes in the extant PINs resulting in the emergence of many hub-like high degree intermediates.

Fig 3. Tracing of elementary process in network evolution from Pre-PIN until extant Ara-PIN and Sol-PIN. (A) The flowchart shows the initial upscaled Pre-PIN model with the simulation process based on the actual ancestral and extant MADS-box PIN parameters (calculated percentages of node deletion, edge addition and deletion). The initializing network is upscaled to the size of 1000 nodes

and 1005 edges based on the actual Pre-PIN with 7 connected nodes where 2 isolated nodes are excluded. Each simulation step reflects bins of 0.001 myr for random edge addition or deletion excluding the steps of ancient whole genome triplication or duplication and node deletions. (B) The visualized networks with scale-free and hierarchical topological characters are shown for each final simulated network. (C) Exponent β distribution for the initial upscaled Pre-PIN model. The main plots show the average value of the β exponent together with the standard deviation (plots in black, yellow and red and error bars in grey) for a total 10 replicated simulations. Blue arrows indicate positions of several important simulated networks following the simulation steps corresponding to the simulation process from A. There are total of ten intrinsic log-log plots showing the relationship between degree *k* and *C(k)* of each node in simulated MADS-box PINs following the simulation steps (for the complete plots, S5 Fig). (D**)** Degree *P(k)* distribution for the initial upscaled Pre-PIN model.

Possible mechanisms behind evolutionary edge and node dynamics: preferential attachment, selection and hub essentiality

To understand the underlying mechanisms behind the observed evolutionary edge dynamics, we investigated the extent to which node degree can explain conservation, gain or loss of edges (Fig 4). We define conserved interactions as those interactions that are retained between single nodes at different points in time or those interactions that are added directly by node duplication (see methods). Edges added through single node or whole genome duplication is a form of preferential attachment because in absolute terms, high degree nodes will grow much stronger than low degree nodes through duplication (Fig 4B). We indeed observe this effect for the conserved interactions (Fig 4A, 1 and 2). Gained interactions include only those interactions that are completely novel and do not derive from previously present interactions through duplication. This type of gained edges allows to investigate the core idea of preferential attachment that posits that highly connected nodes are more likely to acquire new edges which would explain how hubs arise in evolution and scale-free degree distributions originate in extant networks [12,47]. Our dataset would allow to directly observe this ‘the rich get richer’ mechanism as nodes with high initial degrees are predicted to acquire more edges. We observe that large nodes that arose in the network acquired their size by edge gain (Fig 4A, 4) and in addition

by edge duplication (conserved interactions, Fig 4A, 2). However, we find that for the MADS-box networks, the initial node degree does not predict the number of interactions that will be gained (Fig 4A, 2). Rather the opposite is true, in all three evolutionary lineages the initial degree is positively correlated to the number of interactions a protein loses: large nodes tend to lose more interactions (Fig 4A, 5). The final degree is, as can be expected, not correlated to the number of lost interactions (Fig 4A, 6). Overall we observe that throughout evolution, new intermediate degree proteins emerge by gaining novel interactions at the expense of previous high degree proteins, suggesting that the MADS-box PINs rewired to gain more intermediate hubs. This clarifies how the SEPs partially lost their hub characteristics in the evolution from the ancestral to the extant networks, while other proteins gained them.

While it is generally agreed that natural selection does not need to be inferred to explain network structure and network topology of PINs could be explained by gene duplication [48,49], at some level natural selection needs to interact with network structure. To investigate whether the intermediate hubs that originated in the extant networks are selected for or whether different selective pressures occurred in hubs versus non-hubs, we estimated the selective pressure (ω = dN/dS) for each MADS-domain subfamily in general (ω_b) and more specifically for the branches following the γ triplication (ω_t). In agreement with previous studies [50], branch models showed that some subfamilies were subjected to relaxed negative selection directly after the γ triplication (S2 Table). These differences in selective pressure however, were not linked to the protein degree or its rewiring (S3 Table). Only in the Pre-PIN network we found that high degree nodes evolved more slowly ($p = 0.012$) but we did not find this correlation in the post-PIN or in the Ara- or Sol-PIN networks. Because the ω_b of a clade can be influenced by branch specific periods of accelerated protein evolution and because we couldn't calculate the ω_t of all branches between the Asterid-Rosid split and *Arabidopsis*/*Solanum*, we also used sequence similarity as a proxy for the rate of protein evolution and correlated them to the node degree. Again, no significant correlations were found (S3 Table). In general, we would conclude that for the MADS-domain PINs, hubs were not differently selected for than intermediate or non hubs. Even though previously thought otherwise [51], this is in line with more recent beliefs that hub

proteins are not evolving slower [48,52,53]. Finally, to more generally test whether proteins that rewired more strongly experienced natural selection, we linked the number of gained and lost interactions of each protein to their evolutionary rate (ω_b, when available ω_t, sequences similarity and identity) (S3 Table). Here we observe that MADS-subfamilies with a slower evolutionary rate (ω_b) lost more interactions ($p = 0.005$) in the rewiring after the γ event. However a lack of correlation between gained or lost interactions and sequence evolution or ω_b of MADS-proteins from post-PIN to Ara- / Sol-PIN, implies that there is no generally applicable link between selective pressure, protein evolution and the number of gained or lost interactions.

Previously, it has been observed that many hub proteins are essential for the organism [53] and are more often redundant than less connected nodes [54]. We investigated this idea by looking at the correlation between node degree and the severeness of the phenotypes caused by mutants of these genes in *Arabidopsis*. Loss of isolated nodes seems to result in stronger phenotypes than loss of hub proteins (Fig. 4C), supporting the idea that hubs are important communication nodes between proteins with essential functions.

Fig 4. Analysis of MADS-box PINs edge dynamics during the evolutionary time scale. (A) Scatter plots indicating node degree versus conserved interactions (*1 and 2*), gained interactions (*3 and 4*) and lost interactions (*5 and 6*) as determined between Pre-PIN and Post-PIN (*top*), between Post-PIN and Ara-PIN (*middle*) and between Post-PIN and Sol-PIN (*bottom*). Correlations are represented by the Pearson correlation coefficients of the relationships and the associated p value. (B) Example of preferential attachment as a result of WGD. k is the degree of the proteins before (left) and after (right) a WGD. (C) One-way ANOVA between node degree and mutant severity in *Arabidopsis thaliana*.

Avoidance of mis-interaction facilitates the evolution of modular characteristics

The observation that the γ triplication resulted in rapid rewiring yet without an apparent link with selective protein evolution, prompted us to investigate whether the emergence of specific topological organisations characteristic to modular networks is influenced by alternative selective pressures. In a

crowded cellular environment, gaining and maintaining beneficial interactions are not the only selective pressures driving the evolution of PINs. Avoiding specific deleterious interactions likely concomitantly contributes to the evolved PIN [55]. Indeed, protein mis-interaction avoidance influences protein evolutionary rate, *in vivo* protein concentration, as well as the number of distinct proteins in a single cell [55–58]. In order to understand whether mis-interactions could have played a role in the evolution of the MADS-box network, we developed an evolutionary algorithm in which fitness of an individual is defined by the interactions between proteins in a network. As such, the formation of the protein interaction network in an individual correlates to the reproductive success of that individual. We made use of bZIP coiled-coil parameters to determine the pairwise biophysical interaction between different proteins [59] as no specific biophysical data are available for MADS-domain proteins and their interactions are governed, like bZIPs, by the formation of coiled-coils [60]. Previously, Johnson & Hummer already predicted using a physical model that minimization of nonspecific interactions could act as an evolutionary pressure leading to scale-free topological organisation [57]. However, the predictability of evolutionary outcomes is not merely determined by biophysical constraints but also by mutational contingency and population dynamics. The sequence in which substitutions occur or not occur can also have a significant influence on the accessibility of evolutionary paths.

We simulated evolution towards 500 randomly generated target topologies consisting of 12 proteins connected by 24 interactions with and without the presence of disadvantageous mis-interactions. The starting genotype was identical for each individual and each evolutionary run. The size and density of these topologies were selected because of computational limitations. After 200,000 iterations the final population birth rate relative to the initial birth rate was analyzed in function of specific network measurements by linear regression (Fig 5A-C). As reproductive success or fitness in this evolutionary model is solely determined by the birth rate (in absence of other factors; such as predation, disease, competition, etc.) the fitness of an individual is given by its birth rate. Our results show a highly significant correlation between the average clustering coefficient and relative birth rate when mis-interactions are present (Fig 5A). As each simulation represents an independent evolutionary

trajectory, a high amount of variance is to be expected which is reflected by noise around linear regressions and indicated by the generally low R^2 values. This high clustering coefficient suggests evolution favored modular networks, in which few nodes contain a large proportion of edges referred to as hubs while most nodes have a low proportion of edges [61]. Indeed, topologies with a high final relative birth rate also have a higher maximum degree, indicating the presence of such hubs (Fig 5B). The maximum or highest degree of a topology is the largest degree in that topology. Furthermore, the average node connectivity, i.e. the number of nodes that on average need to be removed to disconnect the graph, is negatively correlated with relative birth rate (Fig 5C). Topologies with low node connectivity are likely to contain hub-like proteins as such topologies are easily disconnected through targeted node removal. Without mis-interactions, the average clustering coefficient and average node connectivity are also significantly correlated with relative birth rate, albeit to a much lesser extent than with mis-interaction. To test whether these results are independent from genotype, simulations and subsequent analyses were performed using nine alternative starting genotypes. Results of these simulations, not presented in Fig 5A, are summarized in S4 Table and show qualitatively similar results.

To better understand the underlying evolutionary mechanism behind the link between network properties characteristic to modular networks and the presence of mis-interactions, we determined the number of nonsynonymous substitutions (dN) of each protein and plotted this against their degree (Fig 5D). Our results indicate that when mis-interaction influences birth rate, the evolutionary rate of proteins follows a U-shaped curve with increasing degree. More specifically, proteins with either a low or high degree, i.e. hub and non-hub proteins, exhibit a higher evolutionary rate compared to proteins with an intermediary degree. The low evolutionary rate of proteins with an intermediary degree might be explained due to the conflict that exists at the interaction interface of these proteins to increase their interaction strength of specific advantageous interactions while also minimizing mis-interactions. In contrast, low and high degree proteins, on average, either minimize mis-interactions or maximize beneficial interactions respectively. As modular networks consist of both proteins with low and high degree, these observations explain why topologies with properties characteristic to modular

networks achieve a higher relative birth rate when avoidance of mis-interaction is an important selective pressure (Fig 5A-C). Without mis-interaction as a selective pressure, dN increases linearly with degree (Fig 5D). This can explain why average clustering coefficient as well as average node connectivity is also correlated with relative birth rate in absence of mis-interactions, as the evolution of networks consisting of high degree protein are favored (Fig 5A and C). However, this trend is not as pronounced as with mis-interaction because low degree proteins exhibit a low evolutionary rate.

Fig 5. **The role of avoidance of mis-interactions in the evolution of protein interaction networks.** (A) Relation between network properties and logarithm of relative birth rate (final birth rate of the population / initial birth rate of the population) of simulations of 500 different topologies (same starting populations) with and without mis-interaction. R^2 and p values indicate the coefficient of determination and significance level respectively of linear regression analyses. Lines indicate the best-fit linear regression line. (B) Relation between the number of nonsynonymous substitutions (dN) and number of edges (Degree, k) and the effect of presence or absence of mis-interaction. X-axes indicate the number of beneficial interaction (k, black +) as well as the number of deleterious interactions (red -). Box plots represent the median and interquartile range of simulations of 500 different topologies.

Discussion

The study of networks has become an integral component of molecular biology as illustrated by the numerous complex datasets that have been generated [9,62–64]. These studies have shown that protein interaction network structural properties among distantly related species are similar [1,2], suggesting that there are generic background mechanisms driving the evolution of protein interaction networks. In comparison to previous network evolution studies, our study contributes the clarity provided by direct observation of evolutionary rewiring events by resurrecting ancestral proteins and their associated networks. Although we focused on a protein subnetwork, our study includes the largest number of ancestral sequences used in a single study and this permitted us to accurately characterise evolutionarily related networks using a single consistent method. By selecting a case from plants,

which characteristically have undergone several whole genome duplications in their evolution, we observed the dynamics of protein interactions before and after concerted gene duplication events allowing for the evaluation of gene duplication in the evolution of protein interactions and their networks.

An important observation is that growth of the MADS-PIN appears to be constrained because the size of the network acquires an apparent maximum and network density is constant despite multiple additional rounds of whole genome duplication. This suggests that a mechanism operates that restricts the network size and density. A plausible explanation would be that structural properties of the proteins themselves limits the possible interactions of MADS-domain proteins so that the constraint would be biophysical imposing reorganisation following a concerted duplication event. The γ whole genome triplication indeed triggered the network to rapidly rewire which occurred mainly through addition of new protein interactions or neofunctionalization. We also observed a process of which the extent has previously not been realised to our knowledge, in which proteins gain interactions with two recent paralogs simultaneously, probably because their sequences are highly similar. We propose to call this neo-redundancy. The extent to which this phenomenon is present in the evolution of other PINs remains to be studied but - if widespread - could significantly affect the results of comparative interactome studies because they may have underestimated the number of ancestral interactions when inferring these from extant interactions using parsimony [65].

While several mechanism that drive the evolution of protein interaction networks have been proposed, until now they remain plausible explanations which have not been directly tested. We found that in the case of MADS-domain proteins, the gene balance theory could not explain the loss or retention of nodes. This is surprising as MADS-domain proteins could be expected to follow this proposed process given that they typically assemble into higher order complexes and are preferentially retained after whole genome duplications [18,33]. We did observe preferential attachment of conserved interactions through gene duplication, however we did not find evidence for preferential attachment of new interactions to existing hubs. By contrast, hubs preferentially lose interactions in our data. This is in

agreement with the rewiring and evolution of new hubs we observe. Once a protein gains new interactions it seems to have acquired the capacity to become a hub through preferential attachment by gene duplication.

Our strategy allowed us to directly follow the evolution of modular organisation which revealed that all four networks can be described as small world networks that evolved from having one prominent hub towards having several intermediate hubs. Simulations revealed that all networks are hierarchically modular and that the increase of nodes through the γ triplication was required to retain modularity under strong edge dynamics. Specifically, we identified the dominance of edge addition as an important driver in the retention of modularity. Interestingly, we found that the protein's ability to have few or many interactions is not differently selected for. Rather, using an evolutionary algorithm, we show that the avoidance of mis-interactions can lead to the evolution of networks consisting of hubs, ie. modular networks. This suggests that the observation that PINs are often characterised as being modular can occur due to biophysical limitations in a crowded cellular environment. As such modularity in protein interaction networks can be represented as an evolutionary spandrel to avoid mis-interaction.

Methods

Reconstruction of ancestral MADs-box proteins

Sequence alignments. Initial nucleotide alignments of the MADS-box subfamilies *AP1*, *AP3*, *PI*, *AG*, *AGL2/3/4* and *SOC1* were obtained from [66] and [16]. These data matrices were supplemented with sequences obtained from Genbank (http://www.ncbi.nlm.nih.gov/), OneKP (https://sites.google.com/a/ualberta.ca/onekp/home) [67], Phytozome [68], or from the *Gunnera maniacata* and *Pachysandra terminalis* RNAseq dataset from [16]. *STK*, *SEP3*, *SVP/AGL24* alignments were generated *de novo* from sequences obtained from aforementioned databases. The final data matrices contained between 70 (*STK*) and 222 (*AGL2/3/4*) sequences representing all major

angiosperm clades with an emphasis on orders that branched off around the γ triplication. Sequences were initially aligned with MUSCLE [69] and manually curated in McClade 4.08 [70]. Accession numbers of all genes used for ancestral reconstruction are listed in S1 Appendix.

Phylogenetic reconstruction. Maximum Likelihood phylogenies of MADS-box subfamilies *AP1*, *AP3*, *PI*, *AG*, *SEP1/2/4* and *SOC1* were retrieved from [66] and [16]. *STK*, *SEP3*, *SVP/AGL24* ML phylogenies were constructed using PhyML 3.0 as implemented in Geneious 5.4 or by RAxML [71–73] using the GTR substitution model. Even though the resulting SVP-phylogeny highly insinuated that SVP, AGL24 and StMADS11 are sister-clades originating at the γ triplication, we used synteny implemented in PLAZA 3.0 [74] to further confirm their origin at the triplication (S7 Fig). In order to infer ancestral proteins, a tree representing the evolutionary history of the different taxa is needed. Since the acquired ML gene trees do not always follow the consensus angiosperm phylogeny, all phylogenies were manually improved up until the order level to match with the angiosperm phylogeny described in [74,75]. Branch lengths were estimated on these manually curated trees using RAxML 7.0.4 [76] with the JTT+G or JTT+I+G models of protein evolution, as determined by ProtTest 2.4 [77].

Ancestral sequence reconstruction. The indel history of the sequence alignments was manually reconstructed. All insertions that occurred after the gamma-event were deleted from the matrix. Next, the nucleotide sequence alignments were translated to proteins. The optimized gene trees with branch lengths, the protein alignments and best-fit model of evolution were then used for maximum likelihood marginal reconstruction implemented in PAML4.4 [78]. Ancestral sequences were estimated at the last node before the γ triplication (after the divergence of Buxales and before the divergence of Gunnerales) and at the Asterid-Rosid split. Phylogenetic trees used for ancestral reconstruction indicating ancestral nodes at which ancestral proteins were reconstructed are shown in S1 Fig. Finally, the obtained ancestral protein sequences were converted to nucleotide sequences, codon optimized for yeast and Arabidopsis and synthesized by Genscript USA. All ancestral sequences are shown in S2 Appendix.

Cloning of *Arabidopsis thaliana*, *Solanum lycopersicum* and ancestral MADS-box genes

The full-length coding sequence of *Arabidopsis thaliana* MADS-box genes were used from the *Arabidopsis* Information Resource (TAIR) in order to design primers for gene amplification. The following 17 genes were selected: *AP1, CAL, FUL, AP3, PI, AG, SHP1, SHP2, STK, SEP1, SEP2, SEP4, SEP3, SVP, AGL24, AGL42* and *SOC1*. Tissue samples from *Arabidopsis thaliana* were frozen in liquid nitrogen and stored at -80°C. RNA was isolated from these samples using the TRIzol® method following the manufacturer's instructions. The purity and concentration of RNA samples were determined using NanoDrop spectrophotometer. Synthesis of cDNA from RNA was carried out using the GoScript reverse transcription system (Promega). *Solanum lycopersicum* MADS-box genes were amplified from yeast two-hybrid constructs from [31]. The following 21 cDNAs were subcloned into the pGADT7 (pAD) and pGBKT7 (pBD) vectors (Clontech Laboratories, Inc.): *MC, TM4, SLMBP7, SLMBP20, TAP3, TM6, LePI, TPI, TAG1, TAGL1, SLMBP3, TAGL11, TM29, RIN, LeMADS1, SLMBP21, TM5, JOINTLESS, SLMBP24, SLMBP18 and TM3*). Ancestral genes were amplified from pUC57 constructs containing the ancestral genes obtained from GenScript and cloned into pGADT7 (pAD) and pGBKT7 (pBD) vectors. Due to unknown reasons the subcloning of *TM6* into the pBD vector was not achieved. Miniprep was carried out using PureYield™ Plasmid Miniprep System (Promega). All miniprep plasmid samples were sent for sequencing to confirm in frame insertion of the correct gene in the expression vectors (LGC Genomics GmbH).

High-throughput yeast two-hybrid method

Recombinant pAD and pBD vectors containing ancestral or extant MADS-box genes were co-transformed into the Y187 yeast strain. To determine possible auto-activation, recombinant pBD constructs were co-transformed with an empty pAD vector. Co-transformation of empty pBD and pAD vectors was used as a negative control to measure background signal. Yeast transformation was performed as described in [79]. Following transformation, double transformants were selected on SD/-Leu -Trp plates. To analyze protein-protein interaction, β-galactosidase activity was detected by use of *ortho*-Nitrophenyl-*β*-galactoside (ONPG) as a substrate [80]. After 4 days on selection plates, 3-5 independent co-transformants were pooled into co-transformant groups as we expect no significant biological variation between co-transformants given their identical genotype. These co-transformant

pools were grown overnight in 2 mL SD□medium at 30 ˚C with shaking at 230 rpm. The following day, 100 μL YPD medium was transferred to each well in a 96□well 200 μL micro-plate and for each combination 25 μL of the overnight culture was added into three different wells to perform the β-galactosidase assay in triplicate. This allowed us to accurately take into account variation during the assay. The cells were grown at 650 rpm at 30 ˚C and harvested by centrifugation (5 min at 1000 x g) when OD600 reached 0.5-0.8. Cell pellets were resuspended in 150 μL Z buffer and shaken at 700 rpm at 30 ˚C for 15 min to ensure sufficient homogenization of the cell pellets. Subsequently 100 μL of the resuspended cell culture was transferred to a 2.2 mL 96□well plate (MegaBlock, Sarstedt). Cells were broken by 3 cycles of freezing in liquid nitrogen and thawing in a 42 ˚C waterbath. Afterwards, 160 μL 4 mg/mL ONPG in Z buffer and 700 μL *β*□mercaptoethanol in Z buffer (1:370, *v/v*) were added to each well. The MegaBlock was then incubated at 30 ˚C for 6 hours. Following incubation, 96 μL from each well was transferred to a 200 μL 96□well plate. To stop the reaction, 40 μL of 1 M Na_2CO_3 was added to each well. Finally, absorbance at 420 nm was measured. The amount of *β*-galactosidase which hydrolyzes 1 μmol of ONPG to *ortho*□Nitrophenol and D□galactose per minute per cell is defined as 1 unit. Therefore, *β*-galactosidase activity (Miller units) was calculated using the following formula: Miller units = (1000 x *A*420) / (t x V x OD600) with *A*420 = absorbance at 420 nm, OD600 = Optical density at 600 nm, t = 360 minutes and V = 0. 1 mL.

Identification of positive protein-protein interactions

To determine whether two proteins interacted, Miller units of each combination were compared with the background activity of the negative control by Student's t-test (one-tailed). If the combination was significantly higher than the control (p-value<0.05), these combinations were considered as possible interactions. As auto-activation can lead to false positives, we also determined for each possible interacting combination the presence of auto-activation by comparing them with their respective auto-activation samples, again using the Student's t-test (one-tailed). If no auto-activation was detected, the combination was assigned as truly interacting. For those combinations for which auto-activation was also detected, Miller units were compared to their auto-activation samples by Student's t-test (one-tailed). If the Miller units of the combinations was significantly higher than their autoactivation, the

combination was assigned as a true interaction. If they were not significantly different, these combinations were considered as false positives due to autoactivation.

Network parameters Inferred PIN

All network parameters were calculated in Cytoscape 3.2.1 [81] or NetworkX [82].

Network analyses using parameterized simulations

All computations, network analysis and generation of plots were carried out using the R programming language (version3.1.0) and igraph 0.7.1 [83].

Construction of initialization networks. We initialized our simulations with the following different network types. Thereafter, we applied the network dynamics as measured from the Y2H Pre-PIN, Post-PIN, Ara-PIN and Sol-PIN. (a) Randomized network: An Erdős–Rényi network (1960) [84] was generated using erdos.renyi.game function of the igraph R package. We used a G(n,p) undirected graph which had $n=1000$ nodes and the probability that an edge was present in the graph was $p=0.01$. The resultant node connectivity in the network followed a Poisson distribution. (b) Scale-free network: To simulate a scale-free network, we implemented the Barabási–Albert model of preferential attachment (1999) [12] using R script. We started with a single node and no edges in the first step. Then we added one node at a time in each of the subsequent steps. The probability of node attachment was dependent on the degree; k of the previous node. Finally, this network grew to a size of 1000 nodes. We achieved this by generating a matrix of data elements created by random permutation of all the elements of a vector; x. A vector of weights that can be explained as the node degree; k was used to obtain the elements of vector; x which was being sampled. This scale-free network was initialized to possess a degree distribution that followed the power law $P_k \sim k^{-\gamma}$ with the exponent in the range $2 \leq \gamma \leq 3$. (c) Scale-free modular network: The Pre-PIN network of 7 connected nodes was upscaled to a node size of 1000 and 1005 edges using pure preferential attachment.

Simulation process. The various events of WGDs and network dynamics were simulated on an evolutionary time scale and the empirical probabilities measured from the actual Y2H networks were

applied. The networks were triplicated or duplicated to hypothetically re-create the WGD events as a result of which the whole set of nodes started interacting with their partners and their partners paralogs. For node deletion, random nodes were sampled and eliminated. With random node deletion several links associated with those nodes were also deleted. The removal of nodes was always placed after WGD events because several studies have confirmed that after WGDs, the nodes that are not functional will be quickly silenced or lost from the genome [9]. The process of application of edge dynamics was randomized. A sample space of 'n' simulation steps was created for for each of the time periods using the calculated age estimates. Each time-step reflected bins of 0.001 Mya assuming that 1000 edges might have been lost or gained per Mya. There is no reference to plants however, it has been reported in eukaryotes that the interaction turnover is approximately 100 to 1000 gains and losses of interactions per Mya per proteome [85,86]. We took the upper limit. Edges were added or deleted by deciding on a random Bernoulli draw with the fixed edge probabilities. Based on the Bernoulli trials, a new interaction was created by appending an edge to the existing network matrix. The edges to be added were pre-determined by randomly sampling any two nodes at a time from the pool of remaining nodes in the network and linking them. To eliminate an interaction, we randomly chose one existing interaction at a time and deleted it.

Statistical analysis. Ten simulations were carried out and the mean, variance, standard deviation was calculated. At each stage of application of network dynamics, we determined (a) Degree distribution (P_k) and (b) Clustering coefficients; (C_k) of the nodes in the network. To determine the scale-freeness of the network, we plotted the (P_k) of the nodes against the node degrees in the logarithmic scale. To determine the hierarchy of the network, we plotted the (C_k) of the nodes having (k) connections as a function of (k) in the logarithmic scale. Linear regression was used to model the relationships between the above variables at each step of the simulation process. The quality of the linear fit of the model was estimated using R-squared estimate of goodness of fit.

Detecting selection pressures

Selection pressures (ω = dN/dS, the ratio of nonsynonymous over synonymous substitutions) were estimated using the PAML4.4 software package [78] and the phylogenetic trees and (nucleotide)

alignments used for ancestral reconstruction. Branch lengths were re-estimated using the nucleotide alignments. To test for differences in selection pressure on the branches between two nodes compared to the selection pressures on rest of the tree (background branches), we compared the 'one-ratio' branch model (M0) to a 'two-ratio' branch model (M2) in which we selected all branches between the γ triplication (or the branching of the Buxales when the gene did not duplicate) and the Rosid-Asterid split as a foreground [87]. Nested likelihood ratio tests (LRT = 2*(lnL alternative model – lnL null model)) were performed between branch models M0 and M2. P-values were obtained using the χ^2 distribution with a 0.05 significance at a critical value of 3.84 for 1 degree of freedom.

Determination of interaction changes and correlation analyses

To examine correlations between network rewiring (as in number of changes in interactions) and selective pressure or sequence identity/similarity, we compared the interactions of each individual protein to the interactions of its ancestor. Since there are always four possible fates that a (mis-)interaction can undergo, we used the following definitions to describe the network rewiring. Changes in interactions can be caused by gains or losses of interactions. A gained interaction is here defined as an interaction between two proteins, while their ancestors did not interact. E.g. anceuAG interacts with ancFBP22, while it's pre- γ ancestor ancAG did not interact with ancSOC1 (Fig 1). A lost interaction is defined as a lack of interaction between two proteins, while both ancestors did interact. E.g. anceuAG does not interact with ancSVP nor with ancAGL24, while their ancestors ancAG and ancSVP/AGL24 did. In this case we counted two lost interactions, because anceuAG lost the ability to interact with both ancSVP and ancAGL24. Conservation in interactions can apply to conserved interactions, or to a conserved mis-interactions. A conserved interaction is defined as an interaction between two proteins, when both ancestors already interacted. E.g. anceuAG interacts with ancSEP3 just like their ancestors ancAG and ancSEP3 already did. Finally, a conserved mis-interaction is defined as a lack of interaction between two proteins, when both ancestors already did not interact. E.g. anceuAG does not interact with anceuAP3 or with ancTM6, while their pre- γ ancestors ancAG and ancAP3 already did not. This accounts for two conserved mis-interactions.

After quantifying the changes in interactions, Pearson correlations were used to link them to the selective pressure over the whole gene-tree (background = ω_b), the dN/dS over the branches between the branching off Buxales en the Rosid-Asterid split (foreground = ω_f) and to the sequence identity and similarity between the proteins and their direct ancestor. Significance was determined at the $p < 0.05$ level. Sequence similarity and sequence identity were determined at the protein level using EMBOSS Needle Pairwise sequence alignment with default parameters. Mutant severity of the different proteins was based on the phenotypical effect of knockout or knockdown their genes in *Arabidopsis*. Differences in proteins' mean degree between weak, medium and strong phenotypes was determined by a one-way ANOVA. Significance was measured by a Tukey's HSD test.

Evolutionary model

Our model evolves a population of simple organisms, which each consist of a specified number of genes of 81 nucleotides in length. These genes encode for 29 amino acid long polypeptides that interact by forming coiled-coils with an interaction score determined using the base-optimized weights method developed by Fong et al. (2004) [59]. Note that folding into α-helices is not explicitly modelled and assumed to automatically occur. This method uses electrostatic rules and coupling energy measurements to predict a coiled-coil interaction between two protein sequences that form α-helices. The probability of an organism's survival is proportional to the extent the interactions its proteins form corresponds to an a priori defined target network. The network score, Γ, of an individual is the product of the interaction score of all beneficial interactions divided by the product of the interaction score of all non-functional mis-interactions. The birth rate, β, of an individual in a population of N individuals is given by the function: $\beta_{ind}=\beta_0(\Gamma_{ind} / \Gamma_{pop})$ where $\beta_0=0.1$ is the basal growth rate and Γ_{pop} ,the average network score, that gives the average of the network score for all individuals in the population. As reproductive success or fitness in this model is solely determined by the birth rate (in absence of other factors; such as predation, disease, competition, ...) the fitness of an individual is given by its birth rate. β_{ind}allows the evolution of increasingly higher interaction scores of beneficial interactions and lower interaction scores of deleterious mis-interactions, ie. mutations that

increase interaction scores of beneficial interactions and/or decrease interaction scores of deleterious interactions result in a higher birth rate and thus these mutations are selected for. For neutral mutations $\Gamma_{ind} = \Gamma_{pop}$ and thereby β_{ind} equals β_0 resulting in no selective pressure. To prevent exponential population growth, individuals are removed from the populations with probability equal to their death rate $\delta_{ind} = \delta_0$ N/K where δ_0=0.1 is the basal death rate, N is the population size and K the carrying capacity ensuring a constant population size around N=K. In all simulation the starting population size equals K=1000. The model was written in python.

To calculate the dN and dS of proteins at the end of a simulation, we used the counting method, which constitutes of simply counting the number of nonsynonymous and synonymous substitutions and sites between the starting sequence and the final sequence of a specific protein from the most dominant genotype. The ratio of the number of nonsynonymous substitutions to the number of nonsynonymous sites equals dN. dS equals the ratio of the number of synonymous substitutions to the number of synonymous sites.

Acknowledgements

The authors whish to thank Anja Vandeperre for technical support and Dr. Chuck Leseberg and Professor Long Mao for providing the tomato constructs. We would like to thank the oneKP platform for the availability of unpublished sequences. More specific, we would like to thank C. de Pamphilis, D. Soltis, J. C. Pires, M. Chase, M. Deyholos, N. Stewart, R. Baucom, R. Sage, S. Cannon, T. Kutchan and Tracy McLellan who provided the sequences to oneKP which we used in analyses.

References

1. Li D, Li J, Ouyang S, Wang J, Wu S, Wan P, et al. Protein interaction networks of Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster: large-scale organization and robustness. Proteomics. 2006;6: 456–461.

2. Wuchty S, Peter U. Protein-protein Interaction Networks of E. coli and S. cerevisiae are similar. Sci Rep. 2014;4: 7187.

3. Zhang Y, Yixiang Z, Peng G, Joshua Y. Plant Protein-Protein Interaction Network and Interactome. Curr Genomics. 2010;11: 40–46.

4. Braun P, Pascal B, Sébastien A, Van Leene J, De Jaeger G, Claire L. Plant Protein Interactomes. Annu Rev Plant Biol. 2013;64: 161–187.

5. Rao VS, Srinivasa Rao V, Srinivas K, Sujini GN, Sunand Kumar GN. Protein-Protein Interaction Detection: Methods and Analysis. Int J Proteomics. 2014;2014: 1–12.

6. Futschik ME, Chaurasia G, Herzel H. Comparison of human protein protein interaction maps. Bioinformatics. 2007;23: 605–611.

7. Barshir R, Ruth B, Omer S, Smoly IY, Esti Y-L. Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases. PLoS Comput Biol. 2014;10: e1003632.

8. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AMJJ, et al. A structure-based benchmark for protein-protein binding affinity. Protein Sci. 2011;20: 482–491.

9. Giot L. A Protein Interaction Map of Drosophila melanogaster. Science. 2003;302: 1727–1736.

10. van Noort V, Berend S, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. 2004;5: 280–284.

11. Barabási A-L, Albert-László B, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5: 101–113.

12. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999;286: 509–512.

13. Debodt S, S D, S M, Y V. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 2005;20: 591–597.

14. Adams KL, Wendel JF. Polyploidy and genome evolution in plants. Curr Opin Plant Biol. 2005;8: 135–141.

15. Wagner A. The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. Mol Biol Evol. 2001;18: 1283–1292.

16. Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, et al. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. Mol Biol Evol. 2012;29: 3793–3806.

17. Willis K, McElwain J. The Evolution of Plants. Oxford University Press; 2013.

18. Smaczniak C, Immink RGH, Angenent GC, Kaufmann K. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. Development. 2012;139: 3081–3098.

19. Honma T, Goto K. Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature. 2001;409: 525–529.

20. Theißen G, Günter T, Heinz S. Plant biology: Floral quartets. Nature. 2001;409: 469–471.

21. Yook S-H, Oltvai ZN, Barabási A-L. Functional and topological characterization of protein interaction networks. Proteomics. 2004;4: 928–942.

22. Watts DJ, Strogatz SH. Collective dynamics of “small-world” networks. Nature. 1998;393: 440–442.

23. Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings of the Royal Society B: Biological Sciences. 2001;268: 1803–1810.

24. Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, et al. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol. 2012;10: e1001446.

25. McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. Cell. 2014;159: 58–68.

26. Anderson DP, Whitney DS, Hanson-Smith V, Woznica A, Campodonico-Burnett W, Volkman BF, et al. Evolution of an ancient protein function involved in organized multicellularity in animals. Elife. 2016;5. doi:10.7554/eLife.10147

27. Eick GN, Colucci JK, Harms MJ, Ortlund EA, Thornton JW. Evolution of minimal specificity and promiscuity in steroid hormone receptors. PLoS Genet. 2012;8: e1003072.

28. Yang Y, Jack T. Defining subdomains of the K domain important for protein-protein interactions of plant MADS proteins. Plant Mol Biol. 2004;55: 45–59.

29. Fan HY, Hu Y, Tudor M, Ma H. Specific interactions between the K domains of AG and AGLs, members of the MADS domain family of DNA binding proteins. Plant J. 1997;12: 999–1010.

30. de Folter S. Comprehensive Interaction Map of the Arabidopsis MADS Box Transcription Factors. THE PLANT CELL ONLINE. 2005;17: 1424–1433.

31. Leseberg CH, Eissler CL, Wang X, Johns MA, Duvall MR, Mao L. Interaction study of MADS-domain proteins in tomato. J Exp Bot. 2008;59: 2253–2265.

32. Chen Y-C, Yu-Chi C, Rajagopala SV, Thorsten S, Peter U. Exhaustive benchmarking of the yeast two-hybrid system. Nat Methods. 2010;7: 667–668.

33. Veron AS, Kaufmann K, Bornberg-Bauer E. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. Mol Biol Evol. 2007;24: 670–678.

34. Ruelens P, de Maagd RA, Proost S, Theißen G, Geuten K, Kaufmann K. FLOWERING LOCUS C in monocots and the tandem origin of angiosperm-specific MADS-box genes. Nat Commun. 2013;4: 2280.

35. Veitia RA, Potier MC. Gene dosage imbalances: action, reaction, and models. Trends Biochem Sci. 2015;40: 309–317.

36. Guo H, Lee T-H, Wang X, Paterson AH. Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. Plant Physiol. 2013;162: 769–778.

37. Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 2003;422: 433–438.

38. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485: 635–641.

39. Ohno S, Susumu O. Evolution by Gene Duplication. 1970.

40. Pastor-Satorras R, Smith E, Solé RV. Evolving protein interaction networks through gene duplication. J Theor Biol. 2003;222: 199–210.

41. Vázquez A, Alexei V, Alessandro F, Amos M, Alessandro V. Modeling of Protein Interaction Networks. Complexus. 2002;1: 38–44.

42. De Smet R, Van de Peer Y. Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr Opin Plant Biol. 2012;15: 168–176.

43. He X, Zhang J. Gene complexity and gene duplicability. Curr Biol. 2005;15: 1016–1021.

44. Albert R, Réka A, Albert-László B. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74: 47–97.

45. Koonin EV, Wolf Y, Karev G. Power Laws, Scale-Free Networks and Genome Biology. Springer Science & Business Media; 2007.

46. Ghoshal G, Chi L, Barabási A-L. Uncovering the role of elementary processes in network evolution. Sci Rep. 2013;3: 2920.

47. Eisenberg E, Eli E, Levanon EY. Preferential Attachment in the Protein Network Evolution. Phys Rev Lett. 2003;91. doi:10.1103/physrevlett.91.138701

48. Wang Z, Zhang J. In search of the biological significance of modular structures in protein networks. PLoS Comput Biol. 2007;3: e107.

49. Dwight Kuo P, Wolfgang B, André L. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. Biosystems. 2006;85: 177–200.

50. Shan H, Zahn L, Guindon S, Wall PK, Kong H, Ma H, et al. Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. Mol Biol Evol. 2009;26: 2229–2244.

51. Fraser HB. Modularity and evolutionary constraint on proteins. Nat Genet. 2005;37: 351–352.

52. Jordan IK, Wolf YI, Koonin EV. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol. 2003;3: 1.

53. Batada NN, Hurst LD, Tyers M. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2006;2: e88.

54. Kafri R, Dahan O, Levy J, Pilpel Y. Preferential protection of protein interaction network hubs in yeast: Evolved functionality of genetic redundancy. Proceedings of the National Academy of Sciences. 2008;105: 1243–1248.

55. Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci U S A. 2012;109: 20461–20466.

56. Zhang J, Jingshan Z, Sergei M, Shakhnovich EI. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. Mol Syst Biol. 2008;4. doi:10.1038/msb.2008.48

57. Johnson ME, Gerhard H. Nonspecific Binding Limits the Number of Proteins in a Cell and Shapes their Interaction Networks. Biophys J. 2011;100: 32a.

58. Heo M, Maslov S, Shakhnovich E. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci U S A. 2011;108: 4258–4263.

59. Fong JH, Keating AE, Singh M. Predicting specificity in bZIP coiled-coil protein interactions. Genome Biol. 2004;5: R11.

60. Puranik S, Acajjaoui S, Conn S, Costa L, Conn V, Vial A, et al. Structural basis for the oligomerization of the MADS domain transcription factor SEPALLATA3 in Arabidopsis. Plant Cell. 2014;26: 3603–3615.

61. Ravasz E. Hierarchical Organization of Modularity in Metabolic Networks. Science. 2002;297: 1551–1555.

62. Jasny BR. Mapping the Yeast Protein Interaction Network. Science Signaling. 2008;1: ec226–ec226.

63. Butland G, Gareth B, Peregrín-Alvarez JM, Joyce L, Wehong Y, Xiaochun Y, et al. Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature. 2005;433: 531–537.

64. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, et al. An in Vivo Map of the Yeast Protein Interactome. Science. 2008;320: 1465–1470.

65. Reinke AW, Baek J, Ashenberg O, Keating AE. Networks of bZIP protein-protein interactions diversified over a billion years of evolution. Science. 2013;340: 730–734.

66. Viaene T, Vekemans D, Irish VF, Geeraerts A, Huysmans S, Janssens S, et al. Pistillata--duplications as a mode for floral diversification in (Basal) asterids. Mol Biol Evol. 2009;26: 2627–2645.

67. Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, et al. Data access for the 1,000 Plants (1KP) project. Gigascience. 2014;3: 17.

68. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40: D1178–86.

69. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

70. Phylogenetics MacClade 4: Analysis of Phylogeny and Character Evolution, Version 4.06 David R. Maddison Wayne P. Maddison. Am Biol Teach. 2004;66: 511–512.

71. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28: 1647–1649.

72. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52: 696–704.

73. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol. 2008;57: 758–771.

74. Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inze D, Mueller-Roeber B, et al. PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. 2014;43: D974–D981.

75. Moore MJ, Nasr H, Gitzendanner MA, Bruenn RA, Matthew C, Alexia V, et al. Phylogenetic Analysis of the Plastid Inverted Repeat for 244 Species: Insights into Deeper-Level Angiosperm Relationships from a Long, Slowly Evolving Sequence Region. Int J Plant Sci. 2011;172: 541–558.

76. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22: 2688–2690.

77. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21: 2104–2105.

78. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24: 1586–1591.

79. Gietz RD, Woods RA. Yeast transformation by the LiAc/SS Carrier DNA/PEG method. Methods Mol Biol. 2006;313: 107–120.

80. Miller JH. Experiments in molecular genetics. Cold Spring Harbor Laboratory Pr; 1972.

81. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27: 431–432.

82. Schult DA, Swart P. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science. permalink.lanl.gov; 2008; Available: http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-08-05495

83. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. necsi.edu; 2006; Available: http://www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf

84. Erd6s P, Rényi A. On the evolution of random graphs. Publ Math Inst Hungar Acad Sci. Citeseer; 1960; Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.348.530&rep=rep1&type=pdf

85. Wagner A. How the global structure of protein interaction networks evolves. Proc Biol Sci. rspb.royalsocietypublishing.org; 2003;270: 457–466.

86. Beltrao P, Serrano L. Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. PLoS Comput Biol. 2005;1: e26.

87. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998;15: 568–573.

88. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449: 463–467.

Supporting information

S1 Fig. Simplified phylogenetic trees of the MADS-box gene subfamilies used for ASR and selective pressure estimation. AP1- **(A)**, AP3- **(B)**, PI- **(C)**, AG- **(D)**, STK- **(E)**, SEP1/2/4 **(F)**, SEP3 **(G)**, SVP/AGL24 **(H)**, SOC1- **(I)** gene lineages. Arrows denote the nodes used for ASR before the γ triplication (blue) and at the Rosid/Asterid split (green). Red arrows denote the nodes for which no reliable ancestral genes could be reconstructed.

S2 Fig. Ancestral sequences and their accuracy. Dots represent the posterior probabilities (y-axis) for each amino acid site (x-axis) for all reconstructed genes. Arrows denote ambiguous sites defined as sites for which the most likely amino acid has a posterior probability lower than 0.80 and that have an alternative amino acid state with a posterior probability higher than 0.2. Green arrows represent ambiguous sites of which the most likely amino acids are biochemically similar (S and T; V, L and I; D and E; N and Q; K and R were considered similar).

S3 Fig. Comparison of the PPIs obtained in this study with the determined PPIs from previous published literature. (A-B) Colored blocks indicate PPI between two proteins. PPIs that have been detected in one direction are not differentiated from reciprocal ones. Black represents PPIs present in both studies. Yellow refers to PPIs detected only by previous studies [30,31]. Red indicates the PPIs only found in this study. **(A)** Ara-PIN compared to Folter *et al.* 2005 [30,31]. **(B)** Sol-PIN compared to

Leseberg *et al.* 2008 [31]. **(C)** Comparison between the *Arabidopsis* or *Solanum* network from this study to those reconstructed in de Folter et al (2005) or Leseberg et al (2008) respectively [30,31].

S4 Fig. Gene dosage balance and duplication-divergence models (A) Test of gene dosage balance hypothesis from Pre- to Post-PIN. According to the gene dosage balance hypothesis the proteins (A and B) involved in the nine PPIs in Pre-PIN Y2H should be retained in similar balance after the γ triplication. The graph shows the dosage of the gene copies of all proteins involved in the nine PPIs in Pre-PIN. The number of interacting proteins that were retained in balance after γ are marked in orange. **(B)** Schematic overview of the duplication-divergence possibilities as observed in the Y2H PINs. R, Redundancy; S, subfunctionalization; N, neofunctionalization; N-R, neo-redundancy. **(C)** Summary of subfunctionalization, redundancy and neofunctionalization of all *γ* triplicates when we compared their interactions from the Pre-PIN to the Post-PIN. Genes that didn't duplicate or had no interactions in Pre-PIN are excluded.

S5 Fig. *C(k)* and *P(k)* distribution for tracing of elementary process in network evolution from Pre-PIN, via Post-PIN until extant Ara-PIN and Sol-PIN by simulation models (A) The flowchart of simulation process based on the actual ancestral and extant MADS-box PIN parameters for two models, namely the initial upscaled Pre-PIN model and the initial Erdős–Rényi (ER) random Pre-PIN model. Red arrows indicate critical steps: 1. Actual Pre-PIN with 7 nodes (i) was upscaled into simulated Pre-PIN with 1000 nodes (ii). This step is only achieved for the initial upscaled Pre-PIN model; 2. The network size of upscaled Pre-PIN (ii) or new builded initial ER random Pre-PIN (ii) were three fold upscaled into simulated Pre-PIN with 3000 nodes (iii) by γ triplication; 3. (iii) evolved to simulated Post-PIN (iv) through 37% node deletion followed by 27% edge deletion and 73% edge addition; 4. The network size of (iv) was two fold upscaled into simulated Ara-PIN (v) by *β* WGD; 5. (v) evolved to simulated Ara-PIN (vi) through 52% node deletion followed by 47.2% edge deletion and 22.7% edge addition; 6. The network size of (vi) was subsequently two fold upscaled into simulated Ara-PIN (vii) by *α* WGD; 7. After 23% node deletion followed by 20.3% edge deletion and 9.8% edge addition, (vii) evolved to final simulated Ara-PIN (viii); 8. The network size of (iv) was

three fold upscaled into simulated Sol-PIN (ix) by T WGT; 9. (ix) evolved to final simulated Sol-PIN (x) through 59% node deletion followed by 58% edge deletion and 42% edge addition. **(B)** *C(k)* distribution of each simulated network from the initial upscaled Pre-PIN model (left) and from the initial ER random Pre-PIN model (right) corresponding to simulation process discussed in (A).

S6 Fig. Additional simulations for understanding the consequence of initialization network size and topology, the role of γ triplication and the dominance of random edge addition. Roman numerals in brackets are representing each simulated network according to the flowchart of simulation process in Supplemental Figure 5A. **(A)** Schematic explanation of individual node i's $C(k)_i$ value increasing through random edge addition dynamics by network evolving from simulated Pre-PIN (iii) to simulated Post-PIN (iv). **(B)** Degree distribution *P(K)* for the initial ER random Pre-PIN model. **(C), (E)** and **(G)** Exponent β distribution for the initial upscaled Pre-PIN model without γ triplication. The simulation process began with upscaled Pre-PIN with 1000 nodes (ii) from actual Pre-PIN (i) where Pre-PIN (ii) was not subsequently upscaled to 3000 nodes to Post-PIN (iii) by γ triplication, but the network size kept at 1000 nodes for further simulation process. **(D), (F)** and **(H)** Exponent β distribution for the initial ER random Pre-PIN model. For both models, the average value of exponent β distribution with the standard deviation (plots in black, yellow and red and error bars in gray) of total 10 replicated simulations are shown following the simulation process. The *C(k)* distribution of simulated networks (iv), (viii) and (x) are demonstrated as the additional plots in inset. Blue arrows indicate positions of several important simulated networks following the simulation steps corresponding to the simulation process from **(A)**. **(I)** and **(J)** The dominance of random edge addition was demonstrated by the average value of exponent β distribution and the standard deviation of total 10 replicated simulations from Pre-PIN (ii) to Post-PIN (iv). The additional plots in inset show the *C(k)* distribution of Post-PIN (iv) for random edge deletions preceded random edge additions (left) and random edge additions prior to random edge deletions (right). **(K)** Exponent β distribution in function of edge dynamic ratios (edge addition frequency (%) to edge deletion frequency (%), from 0:100 to 100:0). Blue arrows indicate exponent β values by several main ratios. **(L)** *C(k)* distribution for each simulated network by the annotated ratios from **(K)**.

S7 Fig. Phylogenetic relationships of SVP, AGL24 and StMADS11. (A) Maximum likelihood phylogeny of the StMADS11-subfamily with bootstrap support values. Three subclades of *SVP*, *AGL24* and *StMADS11* can be distinguished. Since all three contain only core eudicots (including Gunnerales for *SVP* and *AGL24*, highlighted in orange) and Buxales and other basal eudicots don't fall within any clade, these results suggest these clades originated from the γ triplication. **(B)** Synteny of *Vitis vinifera SVP-like* genes further supports the idea that *SVP*, *AGL24* and *STMADS11* are sister-clades. (left) *AGL24* and two out of three *StMADS11* genes are located on chr18 and chr3 (blue) which are known to contain syntenic blocks [88]. Because SVP is positioned on an contig unmapped to a chromosome, we looked at some of it's colinear genes (marked in orange (*O-FUCOSYLTRANSFERASE (OFUT)*), green ((MECHANOSENSITIVE ION CHANNEL PROTEIN 3 (*MscS)*) and magenta (Chaperone DnaJ-domain protein)) which were also found in the proximity of *AGL24*. **(C)** and **(D)** ML-trees of *OFUT* and *MscS* show that both genes duplicated at the γ event, indicating that *SVP* and *AGL24* are indeed sister genes. Species and accession numbers are shown in both trees.

S1 Table. Table of calculated Miller Units to determine PPIs. Derived PPIs by F- and T-tests

S2 Table. LRT and parameter estimations of different Branch models

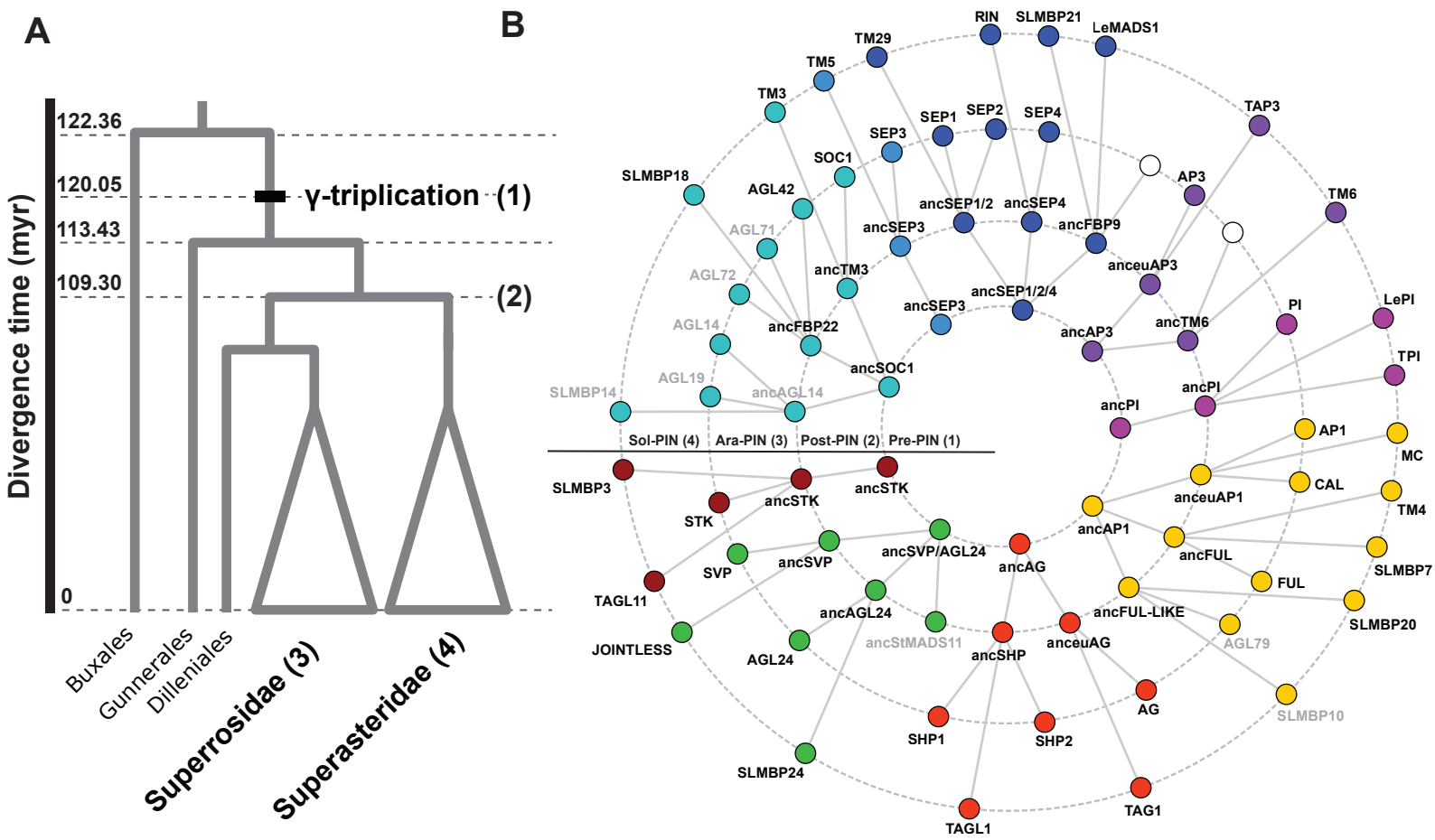
LRT and parameter estimations of different branch models. M0 estimates one ω over all branches. M2 allows ω to vary among selected branches (foreground = ω_f) compared to the rest of the branches (background = ω_b). p-values < 0.05 indicate a significant difference between the M0 and M2 model and thereby implies that there was a difference in selective pressure on the foreground branches compared to the rest of the gene tree. Simplified phylogenetic trees used in these analyses are shown in S1 Fig.

S3 Table. Pearson correlations between edge rewiring (number of gained, lost or conserved interactions) and rate of protein evolution (sequence similarity and identity, selection pressure of foreground branches and selective pressure of the background).

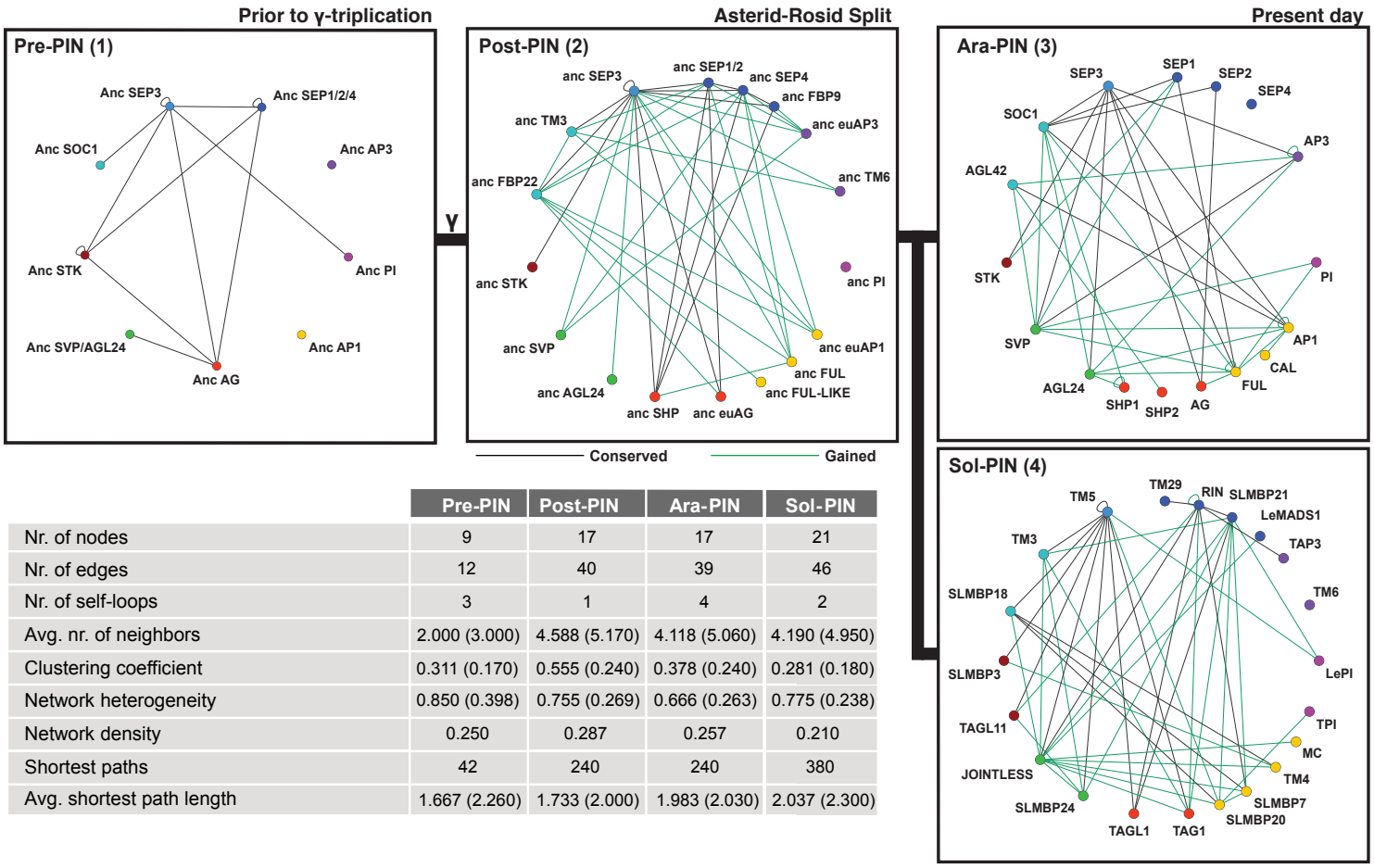
S4 Table. Evolutionary simulation with alternative starting genotypes.

S1 Appendix. Accession numbers of nucleotide sequences used for ancestral reconstruction.

S2 Appendix. Reconstructed ancestral protein sequences in FASTA format


A
B
Divergence time (myr)
122.36
120.05
113.43
109.30
0
γ-triplication
(1)
(2)
Buxales
Gunnerales
Dilleniales
Superrosidae (3)
Superasteridae (4)
RIN
SLMBP21
LeMADS1
TM29
TM5
TM3
TAP3
SEP1
SEP2
SEP4
SEP3
SOC1
SLMBP18
AGL42
AP3
TM6
ancSEP1/2
ancSEP4
AGL71
ancSEP3
ancFBP9
anceuAP3
AGL72
ancTM3
ancSEP1/2/4
ancSEP3
PI
LePI
AGL14
ancFBP22
ancAP3
ancTM6
TPI
ancSOC1
AGL19
ancAGL14
SLMBP14
ancPI
ancPI
Sol-PIN (4)
Ara-PIN (3)
Post-PIN (2)
Pre-PIN (1)
AP1
MC
SLMBP3
ancSTK
ancSTK
anceuAP1
CAL
STK
TM4
ancAP1
ancSVP/AGL24
ancAG
ancFUL
SLMBP7
SVP
ancSVP
FUL
TAGL11
ancAGL24
ancFUL-LIKE
SLMBP20
JOINTLESS
ancStMADS11
anceuAG
AGL79
AGL24
ancSHP
AG
SLMBP10
SHP1
SHP2
SLMBP24
TAG1
TAGL1

A



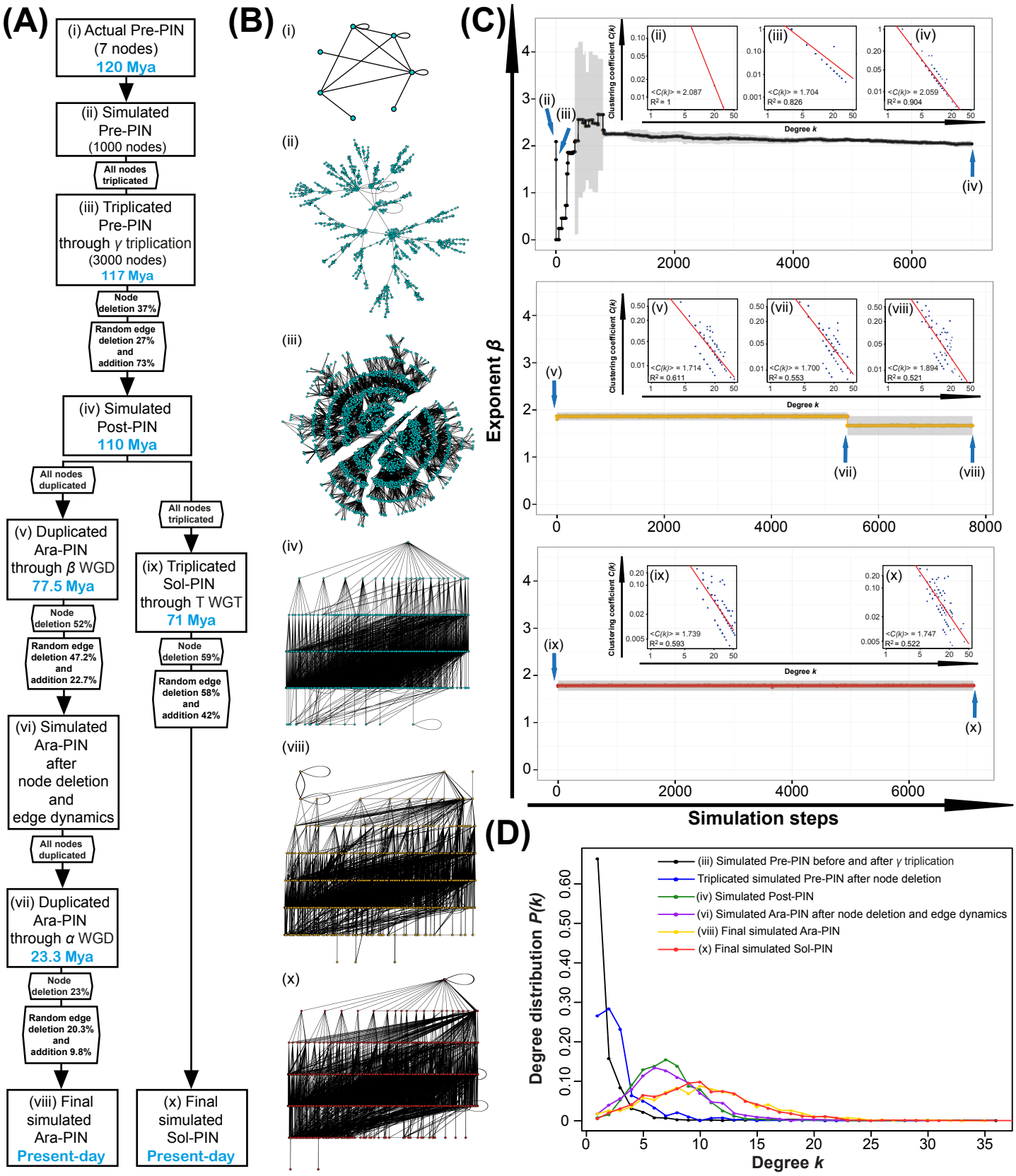
B

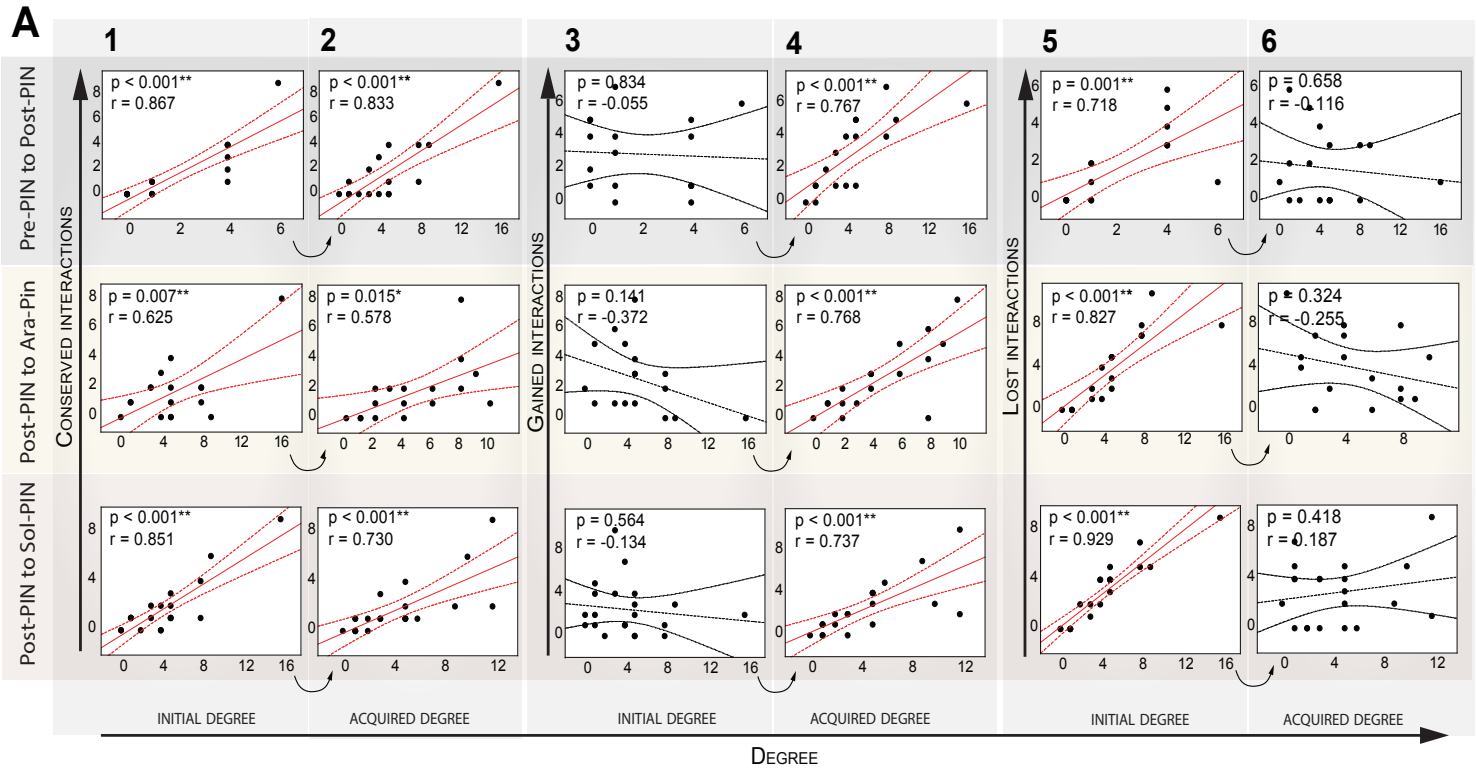
	Pre-PIN	Post-PIN	Ara-PIN	Sol-PIN
Nr. of nodes	9	17	17	21
Nr. of edges	12	40	39	46
Nr. of self-loops	3	1	4	2
Avg. nr. of neighbors	2.000 (3.000)	4.588 (5.170)	4.118 (5.060)	4.190 (4.950)
Clustering coefficient	0.311 (0.170)	0.555 (0.240)	0.378 (0.240)	0.281 (0.180)
Network heterogeneity	0.850 (0.398)	0.755 (0.269)	0.666 (0.263)	0.775 (0.238)
Network density	0.250	0.287	0.257	0.210
Shortest paths	42	240	240	380
Avg. shortest path length	1.667 (2.260)	1.733 (2.000)	1.983 (2.030)	2.037 (2.300)

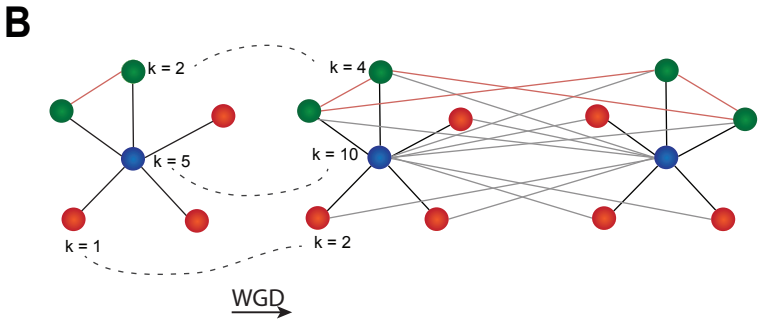
C

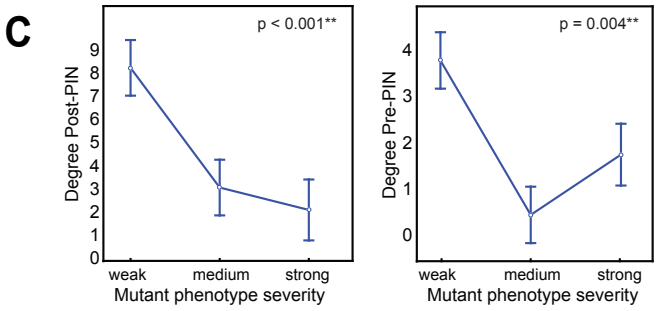
	Divergence time (myr)	Conserved	Gained	Lost	Gain rate[a]	Loss rate[a]
Pre-PIN to Post-PIN	10.75	15	25	17	1.5E-2	1.0E-2
Post-PIN to Ara-PIN	109.3	15	24	35	1.4E-3	2.1E-3
Post-PIN to Sol-PIN	109.3	19	27	31	1,1E-3	1.2E-3

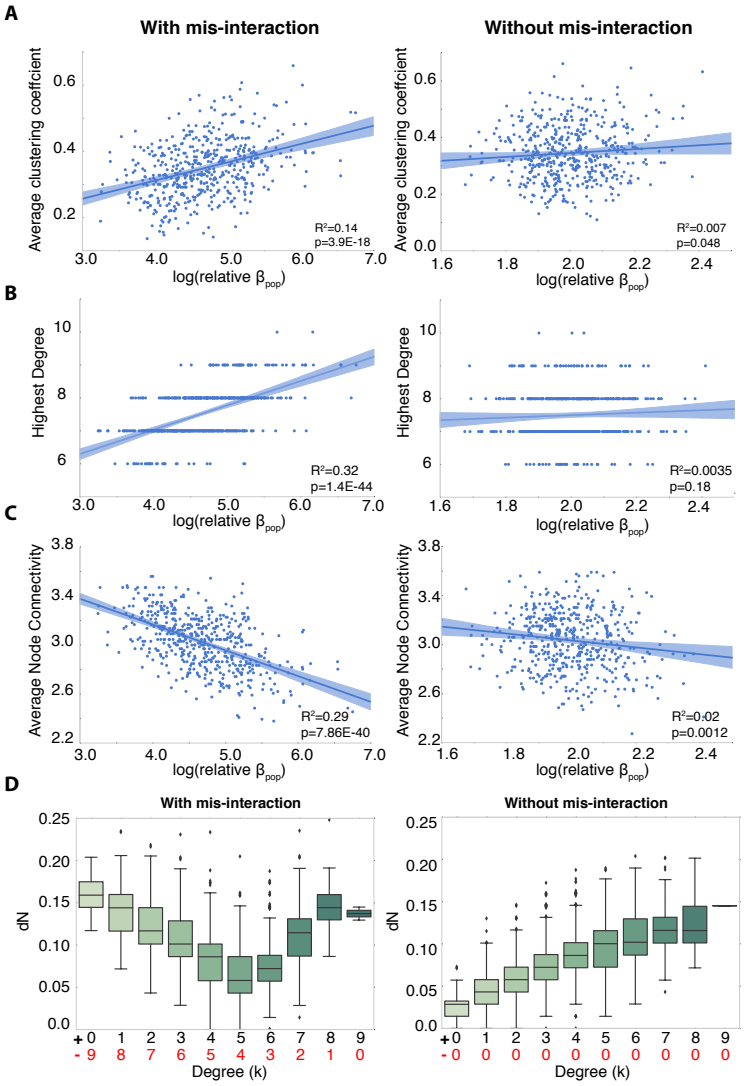
[a]Gain or loss rewiring rate is measured as gain or loss per edge per myr


(A)
(i) Actual Pre-PIN (7 nodes) 120 Mya
(ii) Simulated Pre-PIN (1000 nodes)
All nodes triplicated
(iii) Triplicated Pre-PIN through γ triplication (3000 nodes) 117 Mya
Node deletion 37%
Random edge deletion 27% and addition 73%
(iv) Simulated Post-PIN 110 Mya
All nodes duplicated
All nodes triplicated
(v) Duplicated Ara-PIN through β WGD 77.5 Mya
(ix) Triplicated Sol-PIN through T WGT 71 Mya
Node deletion 52%
Random edge deletion 47.2% and addition 22.7%
Node deletion 59%
Random edge deletion 58% and addition 42%
(vi) Simulated Ara-PIN after node deletion and edge dynamics
All nodes duplicated
(vii) Duplicated Ara-PIN through α WGD 23.3 Mya
Node deletion 23%
Random edge deletion 20.3% and addition 9.8%
(viii) Final simulated Ara-PIN Present-day
(x) Final simulated Sol-PIN Present-day
(B)
(i)
(ii)
(iii)
(iv)
(viii)
(x)
(C)
Exponent β
Clustering coefficient C(k)
Degree k
<C(k)> = 2.087 R² = 1
<C(k)> = 1.704 R² = 0.826
<C(k)> = 2.059 R² = 0.904
<C(k)> = 1.714 R² = 0.611
<C(k)> = 1.700 R² = 0.553
<C(k)> = 1.894 R² = 0.521
<C(k)> = 1.739 R² = 0.593
<C(k)> = 1.747 R² = 0.522
Simulation steps
(D)
(iii) Simulated Pre-PIN before and after γ triplication
Triplicated simulated Pre-PIN after node deletion
(iv) Simulated Post-PIN
(vi) Simulated Ara-PIN after node deletion and edge dynamics
(viii) Final simulated Ara-PIN
(x) Final simulated Sol-PIN
Degree distribution P(k)
Degree k


A
1
2
3
4
5
6
Pre-PIN to Post-PIN
Post-PIN to Ara-Pin
Post-PIN to Sol-PIN
Conserved interactions
Gained interactions
Lost interactions
p < 0.001**
r = 0.867
p < 0.001**
r = 0.833
p = 0.834
r = -0.055
p < 0.001**
r = 0.767
p = 0.001**
r = 0.718
p = 0.658
r = -0.116
p = 0.007**
r = 0.625
p = 0.015*
r = 0.578
p = 0.141
r = -0.372
p < 0.001**
r = 0.768
p < 0.001**
r = 0.827
p = 0.324
r = -0.255
p < 0.001**
r = 0.851
p < 0.001**
r = 0.730
p = 0.564
r = -0.134
p < 0.001**
r = 0.737
p < 0.001**
r = 0.929
p = 0.418
r = -0.187
INITIAL DEGREE
ACQUIRED DEGREE
INITIAL DEGREE
ACQUIRED DEGREE
INITIAL DEGREE
ACQUIRED DEGREE
Degree


B
k = 2
k = 5
k = 1
k = 4
k = 10
k = 2
WGD


C
p < 0.001**
Degree Post-PIN
weak
medium
strong
Mutant phenotype severity
p = 0.004**
Degree Pre-PIN
weak
medium
strong
Mutant phenotype severity


A
With mis-interaction
Without mis-interaction
Average clustering coefficient
log(relative β_pop)
R²=0.14
p=3.9E-18
R²=0.007
p=0.048
B
Highest Degree
R²=0.32
p=1.4E-44
R²=0.0035
p=0.18
C
Average Node Connectivity
R²=0.29
p=7.86E-40
R²=0.02
p=0.0012
D
With mis-interaction
Without mis-interaction
dN
Degree (k)