Supplementary Note 1: Methods for Conant and Wagner, "Convergent evolution of gene circuits," *Nature Genetics*

All references and figures refer to the main text.

Circuit types. In *E. coli*, we examined feed-forward loops and "bi-fans" (see **Fig. 1b**), which we identified from the network data released by Shen-Orr and collaborators[7]. Before analysis, we removed any regulatory elements representing an operon rather than a single gene product from the *E. coli* data. A third kind of *E. coli* circuit, the dense-overlapping regulon described by Shen-Orr and collaborators[7], does not have uniform topology and is thus not suitable for our approach.

In *Saccharomyces cerevisiae* we considered six classes of circuits, five of which were described by Lee and collaborators[5]. These are autoregulation loops (one regulator influencing its own transcription), multi-component loops (a chain of regulators forming a closed regulatory loop), single input motifs (a regulator with multiple target genes), feed-forward loops as defined by Shen-Orr and collaborators[7], multi-input motifs, and regulatory chains (see **Figure 1b** for the last three types of circuits). We identified a sixth circuit type in yeast, the bi-fan, as defined by Milo and collaborators[6]. Of these six circuit types, autoregulation and single input motifs were inappropriate for our analysis as they contain only a single regulatory gene. The multi-component loop was also unsuitable because there are only 3 circuits of this type, containing only three distinct genes. Analysis of multi-input motifs and regulatory chains is complicated because these circuits contain variable numbers of regulatory genes. We analyzed each different size of these circuit types separately (a total of 16 analyses).

In analyzing all these circuits, we considered only the regulatory genes in the circuits and not their downstream targets (including target genes would result in even fewer circuits showing common ancestry). We identified duplicated genes using gapped BLAST[9] at a threshold value of $E_{crit} \leq 10^{-5}$. (Varying E_{crit} from 10^{-3} to 10^{-11} did not change our results.) We are well aware that there are more sophisticated methods to identify duplicate genes. For example, one can call two genes duplicates only if their alignable regions exceed a certain length, if there is a high ratio of mismatches to gaps, and if a minimum percentage of nucleotides match (*e. g.* Conant and Wagner, 2002, *Nucleic Acids Research,* **30**:3378). Any such criterion serves to eliminate false positive duplicates. However, we did not pursue a more stringent approach because false positive duplicates disfavor our hypothesis of independent circuit origin. That is, by using a liberal assay for identifying gene duplicates, the number of duplicate circuits may appear larger than it is. That our hypothesis of independent circuit evolution holds up in spite of this methodical bias against it speaks in its favor.

When evaluating instances of gene circuit duplication it is important to distinguish circuit duplication from simple gene duplication. We thus evaluated the probability that two circuits appear to be duplicates of each other merely because they both happen to contain duplicated gene pairs. For any circuit which showed potential for common circuit origin (*i.e.* $A>0$), we created a distribution of 1000 randomized circuit graphs, each formed by substituting randomly chosen genes for each gene in the original graph. These new random graphs contain the same number of circuits and genes as the original graph, except that two circuits (nodes) are connected if the randomly chosen genes in them are paralogs of each other (as defined by a BLAST-threshold of $E \leq E_{crit}$).

Naively, one might think that the random genes should be drawn from the genome as a whole. However, because only 112 regulator genes could potentially occur in these circuits under the experimental design of Lee and collaborators[5], the appropriate pool of genes to draw from contains only these regulatory genes.

Our analysis of yeast circuits rests on genome-scale chromatin precipitation experiments that use a statistical error threshold (P_e) to identify true regulatory interactions[5]. When varying this error threshold we found that the number of duplicated feed-forward circuits was different from what would be expected by chance alone, albeit at marginal significance (P=0.03, see main text). Further analysis revealed that this result was due to the presence of a single large component containing between 3 and 5 circuits but including only 5 different genes (*ABF1, MBP1, MOT3, SWI4*, and *SWI6*). The presence of this component can be explained solely through the presence of the duplicated *SWI4* gene. The remaining 43 of 48 feed-forward circuits do not share a common ancestor.

We also assessed whether genes that are duplicates of each other are more likely to occur in the same type of circuit. There are 112 transcriptional regulators in yeast on which our circuit information is based, but not every regulator occurs in each circuit type. We define P_{motif} as the probability that a randomly-chosen regulator will occur in a particular circuit type (for example a bi-fan, see **Table 1**). We then calculated $\mathbf{P}_{motif|duplicate}$, the probability of a regulator occurring in that circuit type, given that at least one of its duplicates does. If many gene circuits originated through gene duplication, members of one gene family will be more likely to co-occur in a circuit type than genes at large ($\mathbf{P}_{motif} < \mathbf{P}_{motif|duplicate}$). To evaluate the statistical significance of observed values

for $\mathbf{P}_{\text{motif|duplicate}}$, we tested the hypothesis $\mathbf{P}_{\text{motif}} = \mathbf{P}_{\text{motif|duplicate}}$ with an (exact) one-sided binomial test (**Table 1**) for all circuit types where $\mathbf{P}_{\text{motif}} < \mathbf{P}_{\text{motif|duplicate}}$.

In a second, complementary, analysis, we asked whether instances of any one circuit type, such as the bi-fan, might have evolved through duplication of ancestral regulators within that circuit. In that case, certain circuits may be over-represented within a given gene family. However, a statistical analysis similar to that outlined above reveals no such evidence. Duplicate genes seem to be scattered randomly across the different circuits (results not shown).

We finally note that the instances of convergent evolution we identified differ in one important respect from other examples of convergent evolution: the same circuits have evolved multiple times within the same organism. Our approach can also be applied to detect convergent evolution of transcriptional regulation circuitry in different genomes, as well as convergent evolution of gene circuits not exclusively involving transcriptional regulation.