# Chapter 9 Odd Solutions

**1.** The tag is fairly large and could disrupt the interactions being looked for. One could reasonably expect that if a tag at one end disrupts the interactions, then using a tag at the other end might be successful.

**3.** Both papers tagged all identified ORFs. Tags were added only at the 3′ end of the gene.

Gavin et al. purified 1993 TAP-fusion proteins from 6466 ORFs, a hit rate of only 30%. This is not a bad hit rate, but one would expect to have missed a significant number of complexes, particularly from membrane proteins (although membrane proteins were targeted, and 340 out of 628 target proteins were successfully purified). In the complexes, they identified 2760 proteins, or about 60% of the total number of proteins expected from their growth conditions (exponential growth). These proteins come from all functional classes and all subcellular locations, indicating a relatively good coverage. As one would expect, coverage is greatest for the most abundant proteins. Interestingly, 139 purifications were performed in duplicate, from which 69% of proteins were common to both purifications; or, conversely, 31% of proteins were only found in one purification and not in the other. There is clearly therefore a large amount of variability in the data (or, as the paper expresses it, this variability provides an approximation of false-positive/false-negative rates). A comparison with known databases of complexes in yeast, particularly MIPS, revealed that 73% of known complexes had been identified.

The results were analyzed in various ways to check for completeness, including detailed statistical analysis. For example, several well-known complexes were retrieved several times, giving some indication of the level of coverage. Coverage and accuracy were derived mainly by comparison with known complexes, from which it was concluded that coverage and accuracy were, respectively, 83% and 78% under ideal conditions, or maybe about 70% over the complexes reported.

Krogan et al. purified a similar number of proteins (2357 of them), and identified 72% of the proteins predicted in the proteome. Proteins from different subcellular locations were covered fairly well, with more than 70% from most compartments (including membrane proteins) but only 61% and 59% for endoplasmic reticulum and vacuolar proteins, respectively. A statistical analysis was performed, again using MIPS as the standard for comparison. It was concluded that the coverage (here called precision) was just over 70% (Figure 3b in the Krogan et al. paper), but the homogeneity (roughly equivalent to accuracy) was only 38%.

As described in Chapter 9, the two analyses were combined [SR Collins et al., *Mol. Cell. Proteomics* 6:439–450, 2007, the corresponding author being Krogan] using a Bayesian statistics method, which according to various methods described in the paper provides greater accuracy than either method alone. The combined database contains only about 50% of the interactions reported from the two previous studies, again indicating the significant level of error in the

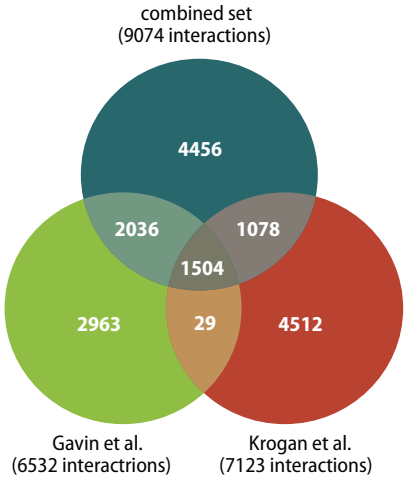


**FIGURE 10**
A Venn diagram showing the number of protein/protein interactions identified in yeast, in two independent genome-wide TAP-tag screens and in a combined screen.

earlier datasets; it would be reasonable to expect a reduced but still significant error in the combined dataset. The coverage and accuracy were estimated not only by comparison with known complexes (a difficult comparison because these complexes were used to derive the method in the first place) but also by comparing subcellular localization, gene ontology annotation, and mRNA coexpression, on the grounds that proteins in the same complex should for example all be located in the same subcellular compartment. Roughly 50% of proteins within a complex were annotated as coming from the same compartment. This does not sound like a very high number, but it is difficult to judge how far this may have been due to deficiencies in the localization data rather than false positives in the complexes. The Venn diagram shown in **Figure 10** (redrawn from Figure 2A in the Collins et al. paper) is of interest, because it shows that the combined dataset produced 4456 'new' high-confidence interactions that were not present in either of the two original datasets, despite being based on the same data. Thus, the combined database is certainly an improvement, but it is also not error free. The combined dataset is estimated to have a coverage of about 80% "of interactions accessible to the TAP approach under the conditions used," which as we have seen probably includes only the higher-affinity complexes. The authors note that it remains difficult to estimate the proportion of false positives.

**5.** False positive interactions are interactions identified between proteins that do not in fact interact. False negatives are interactions not identified between proteins that **do** interact. False positives are usually of more concern because they give rise to false conclusions about how things work, whereas false negatives lead to no conclusions being drawn at all. Having no idea is usually better than having a wrong one. One can find numerous examples of a tentative conclusion being drawn in a paper based on what turns out later to be a false positive, which is subsequently referenced by others as being a result, and becomes entrenched in the literature as a 'fact,' not least because most of us struggle to go back to the original research each time to check exactly what was really said.

**7.** The polymerase core of RNA polymerase II consists of 12 subunits, called RPB1–RPB12 (see Table 9.2). Of these, RPB5, RPB6, and RPB8 are found in all three RNA polymerases, and two others are almost identical. In addition, Pol I and Pol III share two subunits. Pol I has 8–14 subunits, all of which are homologous with subunits in either Pol II or Pol III or both. In addition, the polymerase is associated with two main transcription factors, UBF (which binds to DNA) and SL1 (which is recruited subsequent to the binding of UBF), along with TATA-binding protein (TBP). This is the same TBP as that found in TFIID in Pol II. Pol III has 17 subunits and also uses TBP in initiating tRNA transcription.

**9.** Work cited in Chapter 9 suggests that the best way to identify such complexes would use *in vivo* labeling, because any attempt at purification is likely to disassemble the complex. Thus, for example, one might hope to label different components of a complex with differently fluorescing tags and perform fluorescence resonance energy transfer (FRET) experiments (Section 11.2.4), looking for differences in FRET efficiency before and after stimulation by the appropriate cellular need. One of the fluorescent labels could alternatively be

attached to the membrane, for example by producing an appropriate inositol phosphatide (which could be designed to target particular types of membrane). The other way that has been used successfully to identify complexes is channeling, by feeding a suitably isotope-labeled precursor and measuring dilution of the label in possible products.

**N1.** You would expect 184 complexes containing 4 proteins. A protein complex composed of 63 proteins has a probability of less than 0.5. This does not of course mean that it could not exist, because the correlation is only approximate.

**N3.** Assume that the on-rate is diffusion-controlled; that is, $10^9\,M^{-1}\,s^{-1}$. If $K_d$ is $10^{-6}\,M$, then the off-rate is $K_d \times k_{on}$; that is, $10^3\,s^{-1}$. This is the same as the constant $k$ in the hint, so that for example at a time $t = 1/k$ (1 ms), the concentration of complex will have decreased by a factor of $e^{-1}$ or 0.4. If my limit for observation is 10%, then I need $A/A_0 = 10\%$ or 0.1, or $-kt = \ln(0.1)$, implying $t = 2.3\,ms$. Thus, my washing step needs to take less than 2.3 ms, clearly an impossibly fast wash, and hence it is reasonable that weak complexes are lost. The arguments on macromolecular crowding in Chapter 4 suggest that the affinity should be much stronger if the washing used a crowding agent, for example a high concentration of dextran.