Pooling in systems biology becomes smart

Nicolas Thierry-Mieg

A simple 'smart-pooling' screening strategy for large-scale systems biology experiments promises to provide considerable improvement in experimental efficiency, while simultaneously allowing improved accuracy and coverage.

A variety of systems biology projects aim to identify low-frequency events. Such projects typically face three issues: reducing the number of experiments, recognizing false positives and avoiding false negatives. In this issue, Huang and collegues[1] propose a new pooling strategy capable of simultaneously increasing efficiency, accuracy and coverage. They validate the technique in three experimental contexts: protein chips, yeast two-hybrid assay and drug resistance screening. Such an approach should be of widespread interest and could lead to substantial improvements in the overall quality of these types of data.

Knowledge of the sequence of whole genomes, transcriptomes or proteomes has given scientists access to a much larger playground, where one can in principle interrogate all genes or their products at once. Large-scale experiments trying to identify rare positive molecules in a particular yes-or-no assay are being performed. Efficiency is essential, but the noise inherent to high-throughput biological assays is also a major concern: both false positive and false negative observations are to be expected, and reproducibility is far from granted. For example, in one of the first large-scale interactome mapping efforts[2], the authors screened their two-hybrid arrays with 192 baits in duplicate: only 20% of the interactions were found twice.

In practice, a frequently used protocol involves a two-step naive pooling procedure: probes are first tested in pools, and then retested individually when the pool is positive. Compared to the individual testing method, this strategy increases the efficiency, but false negatives remain a problem that can still only be addressed by repeating the entire experiment.

Jin *et al.*[1] test and validate an alternative 'smart-pooling' approach, which the authors call PI-deconvolution. The strategy consists in assaying well-chosen pools of probes, such that each probe is present in several pools (that is, the pools are redundant as shown in an example of a simplistic smart-pooling approach; **Fig. 1**). Pools are designed so that the positive probes can usually be identified from the pattern of positive pools, and when this is not the case, only a few candidates need to be retested. In addition, the pools' redundancy means that each probe is tested several times: this provides a potential increase in both sensitivity and specificity. The authors illustrate the usefulness and versatility of smart-pooling by applying it to three different assays.

In a first experiment using their PI-deconvolution approach, yeast proteome microarrays were individually screened with 15 baits, providing a reference network of bidirectional protein-protein interactions among these 15 proteins. The baits were then combined into 8 pools of 8 proteins with a fourfold redundancy. By screening the pools, they identified every reference interaction while using only 8 arrays rather than 15.

Relying on a similar design (8 pools of 8 proteins, fourfold redundancy), the researchers screened a yeast genome-wide two-hybrid array with 16 baits, including 13 that had been formerly screened individually in duplicate. They identified many new interactions, although surprisingly 42% of the reproducible single-bait hits were not recovered: the authors attribute this to intrinsic yeast-two-hybrid variability. Nevertheless, the smart-pooling strategy identified roughly as many reproducible interactions as the duplicated single-bait method, but required only 8 screens instead of 32.

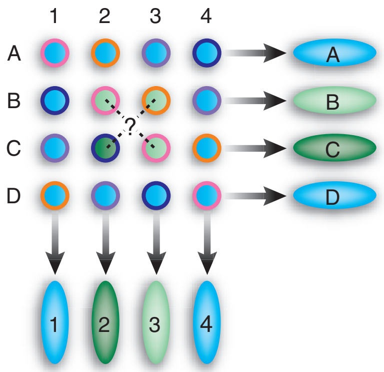

Figure 1 | Example of a simpler design to illustrate the 'smart-pooling' concept. Sixteen probes are arrayed on an imaginary grid (positions A1–D4) and mixed in 8 combinations or 'pools' (one pool per row, A–D; and one pool per column, 1–4), each containing 4 probes. If the pools are tested against a bait in ideal noiseless conditions, and pools C and 2 are positive (green), then C2 is the only positive probe. But if pools B and 3 are also positive (lighter green), the two solutions (B2 and C3) or (B3 and C2) cannot be distinguished. This can be resolved by adding four additional pools, built along one of the grid's diagonals as indicated by the colors circling the probes. The continuity of the diagonals can be visualized by rolling the figure around a cylinder. If the pink diagonal pool is positive, B2 and C3 is the solution, whereas if both the orange-circled and blue-circled diagonal pools are positive, the solution is B3 and C2.

Nicolas Thierry-Mieg is at the Software-Systems-Networks laboratory, Centre National de la Recherche Scientifique (CNRS), Grenoble, France.
e-mail: nicolas.thierry-mieg@imag.fr

Finally, the authors applied the strategy to the identification of drug-resistant mutants from the *Saccharomyces cerevisiae* deletion collection. A total of 128 strains were smart-pooled into 14 pools of 64 with a sevenfold redundancy, and assayed against two drugs. Both previously known resistant strains were unambiguously identified despite the ninefold increase in efficiency.

Although the smart-pooling design used by the authors is intuitive and highly efficient, it can be optimized further. It can be described as the *n*-dimensional hypercube extension of the grid design illustrated in **Figure 1**, restricted to side length 2 and without the diagonal pools. Two potential weaknesses are that the pool sets of some pairs of probes are too similar and others are complementary. Consequently, decoding is ambiguous in the presence of noise or of multiple positives. Another limitation is that choosing the pools' size entirely determines the design: for example, pools of 8 probes necessarily have a fourfold redundancy. This is problematic in the two-hybrid experiment for instance, in which small pools are required for sensitivity but the high error rates would call for more redundancy.

Alternative designs[3–5], which have been described and await experimental validation, may overcome these limitations. For example, a recently proposed design[3] can be precisely adapted to the characteristics of any experiment (for example, pool size and redundancy are set independently). Of course, these more mathematical constructions are less intuitive, and decoding must be done *in silico*, but they scale up extremely well.

Naturally, the biological characteristics drive the choice of the smart-pooling design. The expected fraction of true positive probes and the error rates should be carefully estimated beforehand, and one must also determine the maximal pool size that can be used without excessive degradation of the assay's sensitivity: in many applications, this essentially determines the achievable efficiency. Based on this information, an optimized design—powerful enough yet not wasteful—can be selected and the method applied on a large scale.

Smart-pooling can reduce the number of experiments and yet considerably increase sensitivity and specificity, because the redundancy allows the investigator to identify and correct both false positive and false negative observations. This strategy's feasibility and versatility in real-world applications is now demonstrated, using three different high-throughput technologies. It opens new perspectives for systems biology experiments seeking to detect low-frequency events in the presence of noise: substantial improvements in the quality and breadth of such datasets are now within close reach.

1. Jin, F. *et al. Nat. Methods* **3**, 183–189 (2006).
2. Uetz, P. *et al. Nature* **403**, 623–627 (2000).
3. Thierry-Mieg, N. *BMC Bioinformatics* **7**, 28 (2006).
4. Balding, D., Bruno, W., Knill, E. & Torney, D. In *Genetic mapping and DNA sequencing*. (Speed, T. & Waterman, M.S., eds.) 133–154 (Springer, New York,1996).
5. Ngo, H. & Du, D.Z. *DIMACS Ser. Discrete Math. Theoret. Comput. Sci*. **55**, 171–182 (2000).

Retroviral TCR gene transduction: 2A for two

Rémy Bosselut

A recently developed multigene viral expression system is put to work to generate mice carrying a single T-cell receptor (TCR) specificity. Complementing the transgenic-mice technique, this method offers new practical options to researchers studying T-cell development.

Receptors and ligands that mediate T cell–antigen recognition are remarkable because of their staggering diversity. Thus, a defining breakthrough in T-cell immunology was the generation of transgenic mice whose T cells all carry the same antigen receptor[1], thereby fixing the 'receptor side' of the interaction. An elegant trick reported in this issue of *Nature Methods* should facilitate the generation of monospecific T-cell populations by retroviral transduction of hematopoietic stem cells[2].

Most T cells use a TCR comprising two antigen-specific chains (α and β) to recognize antigenic peptides bound to classical major histocompatibility complex (MHC) molecules[3]. Neither the α nor the β chain is germline encoded; rather, each is produced from an open reading frame generated by random rearrangement of TCRα and TCRβ loci during T-cell development in the thymus[4] (a process called V(D)J recombination). This process results in extensive TCR diversity, so that each developing T cell (thymocyte) carries a distinct TCR specificity.

Because of the extreme allelic diversity of MHC genes, most TCRαβ specificities generated in any given individual are of inappropriate avidity for self peptide–self MHC complexes[5] (**Fig. 1**). Low-avidity receptors are useless and fail to rescue thymocytes from death by neglect, whereas receptors with high avidity for self are potentially harmful and trigger active thymocyte deletion (negative selection). In the end, only the small subset of thymocytes carrying intermediate-avidity TCRs survive and differentiate into mature T cells (a process called positive selection), which normally react against foreign peptide–self MHC complexes.

Key to the study of T-cell development and responses has been the development of mice in which all (or most) T cells express the same TCRαβ specificity, mediating either positive or negative selection and responding to a defined antigen[1]. Note that expression of such recombinant TCR chains prevents or substantially impairs endogenous TCR gene rearrangement[4]. Such mice have so far been generated using conventional transgenic procedures, by introducing into the mouse genome recombinant TCRα and TCRβ minigenes controlled by their own regulatory elements. More recently[6], retroviral

Rémy Bosselut is at the Laboratory of Immune Cell Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
e-mail: remy@helix.nih.gov