Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGEMENTS
P. Braun (Dana-Farber Cancer Institute, Boston) kindly provided the gold-standard clones used in this project. S. Shtivelband helped in the initial phase of this project. S. Fields, T. Hazbun and M. Koegl made helpful comments on the manuscript. This project was supported by the Landesstiftung Baden-Württemberg (Germany), the US National Institutes of Health (R01GM79710) and the European Union (HEALTH-F3-2009-223101).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Yu-Chi Chen[1], Seesandra Venkatappa Rajagopala[1], Thorsten Stellberger[2] & Peter Uetz[1,3]

[1]J. Craig Venter Institute, Rockville, Maryland, USA. [2]Institute of Toxicology and Genetics, Karlsruhe Institute of Technology, Karlsruhe, Germany. [3]Present address: Proteros Biostructures, Martinsried, Germany.
e-mail: peter@uetz.us

1. Braun, P. *et al. Nat. Methods* **6**, 91–97 (2009).
2. Rajagopala, S.V., Hughes, K.T. & Uetz, P. *Proteomics* **9**, 5296–5302 (2009).
3. Stellberger, T. *et al. Proteome Sci*. **8**, 8 (2010).

Braun *et al.* reply: We applaud the thorough and revealing study by Chen *et al.*[1] in this issue of *Nature Methods*. The work expands our previous findings in thoroughly characterizing different yeast two-hybrid (Y2H) implementations, with respect to overall assay sensitivity, by testing each implementation against a panel of reference protein-protein interactions[2]. The standardized reference sets[3] make the data easily comparable to our previous analyses and clearly demonstrate that different Y2H assays detect different subsets of interacting pairs of proteins[2,3]. Given the proven utility of using several assay configurations, the next question is where and how to deploy them. The high-throughput capabilities of Y2H[4,5] make it an ideal primary screening assay. Having multiple versions of Y2H that detect different subsets of interactions will be of a great value to generate more comprehensive data sets, which would then need to be validated using a scheme such as the "confidence scoring" scheme that we proposed[2]. A key concept of our confidence scoring method is that any interaction detected by a given screening assay is subsequently confirmed by multiple orthogonal validation assays. The screening and validation assays must be as independent from each other as possible to eliminate the danger of systematic assay-dependent artifacts, which could make protein pairs appear as robustly interacting when they may not be. Use of a single type of assay for both screening and validation, even if implemented in different configurations, may introduce such systematic biases. It is therefore critical to obtain orthogonal validation ideally of all interacting pairs identified in an initial screen.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Pascal Braun[1,2], Murat Tasan[1,3], Michael Cusick[1,2], David E Hill[1,2] & Marc Vidal[1,2]

[1]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA.
e-mail: pascal_braun@dfci.harvard.edu or marc_vidal@dfci.harvard.edu

1. Chen, Y.-C., Rajagopala, S.V., Stellberger, T. & Uetz, P. *Nat. Methods* **7**, 733–736 (2010).
2. Braun, P. *et al. Nat. Methods* **6**, 91–97 (2009).
3. Venkatesan, K. *et al. Nat. Methods* **6**, 83–90 (2009).
4. Rual, J.F. *et al. Nature* **437**, 1173–1178 (2005).
5. Yu, H. *et al. Science* **322**, 104–110 (2008).

Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions

To the Editor: Pyrosequencing has revolutionized microbial community analysis by allowing the simultaneous assessment of hundreds of microbial communities in multiplex with sufficient depth to resolve meaningful biological patterns. These techniques have been used to study microbial processes on scales ranging from continents[1] to within an individual's body[2].

Although powerful new analysis tools such as 'visualization and analysis of microbal population structures' (VAMPS), Mothur and 'quantitative insights into microbial ecology' (QIIME)[3] greatly streamline the process of interpreting microbial community information—for example, by clustering of reads, taxonomy assignments and visualizations—substantial questions remain about the suitability of pyrosequencing for addressing questions concerning alpha diversity, the amount of diversity in each individual community and non-phylogenetic beta-diversity measures. In particular, noise introduced during pyrosequencing and PCR amplification can inflate estimates of the number of operational taxonomic units (OTUs) (chosen at the 97% identity level) in a given habitat by orders of magnitude[4]. The current state of the art is to reduce noise by clustering the flowgrams (patterns of intensities in each read) before conversion to sequences to eliminate issues owing to homopolymer read errors[5], yet this approach is exceedingly computationally expensive and beyond the reach of most individual investigators.

Inability to accurately determine which sequences are present in a sample, and hence the abundance of rare taxa, greatly inhibits our ability to infer important ecological parameters such as comprehensive rank-abundance curves. Yet the rank-abundance curve of the common taxa can reduce the computational expense of denoising. Empirical rank-abundance curves of actual microbial communities tend to be dominated by a relatively small number of abundant taxa. Consequently, performing all-on-all comparisons for clustering is exceedingly inefficient; instead, a subset of reads suffices to identify the common OTUs, which can then be iteratively removed by recruitment to an existing cluster. First, we devised a fast pre-filter, removing reads that are strict prefixes of other reads, and computed an initial sequence distribution. We then sorted the prefix clusters in descending order of abundance and used this initial distribution to cluster similar reads, comparing each additional unclustered read to the most abundant clusters because we expected the abundant clusters to yield a larger number of erroneous near-matching reads owing to their numerical dominance alone. Then we clustered only the leftover reads representing more divergent sequences (**Supplementary Methods**).

This approach retains the advantage that clusters with only one member are not discarded entirely, allowing exploration of the rare biosphere[6]. We could analyze a small dataset of 40,000 sequences in less than an hour on a single laptop computer and a full Roche 454 sequencer run with 500,000 sequences on a midsize computer cluster in one day (**Supplementary Table 1**). We can thus address questions in community ecology that were previously intractable.

Applying these new methods to human-associated body habitats[2] and several test communities (**Supplementary Table 2**), we found that denoising produced a substantial decrease in the diversity at the OTU level and in phylogenetic alpha diversity. However,