"Bringing functional genomics into focus"

E. Ledesma-Fernández[1]*, P.H. Thorpe*[3] *and R.A.M. de Bruin*[1,2]
[1]MRC Laboratory for Molecular Cell Biology, [2] The UCL Cancer Institute, University College London, London WC1E 6BT, [3]*The Francis Crick Institute, 1 Midland Road, London NW1 1AT.*

In the current issue of Cell Systems, Styles et al. develop an optimized method that combines genome-wide yeast genetics, high content microscopy and automated phenotypic analysis to identify genes in DNA damage repair. This method has the potential to be adapted to a multitude of phenotypes and cellular processes by comparing chemically or genetically perturbed cells to wild type cells with automated computational analysis.

The creation of genome-wide mutant collections in yeast and the development of automated techniques to combine and screen mutants have made yeast a pioneering organism for functional genomics – methodology that enables molecular function to be determined from genomic and proteomic data. In the current issue of *Cell Systems*, Styles et al. add automated, high-content, image-based screening to the functional genomics toolbox.

To establish the functions of a particular gene and investigate its relationship with other genes, it is important to link its role to a specific cellular process (phenotypic analysis). To date, genome-wide phenotypic analysis in yeast has been largely based on cell fitness (using colony size or growth as a readout). This has proven to be an invaluable tool to predict genetic interactions (Baryshnikova et al., 2010), and has been instrumental in guiding detailed follow up analysis to establish gene function. However, since not all phenotypic defects cause a significant defect in growth, this approach has its limitations. In addition, essential cellular processes, where mutants cause inviability are less amenable to this type of analysis.

Imaging fluorescently-tagged proteins seems to provide an answer to both of these issues. It allows a particular cellular process to be studied by tracking a fluorescent signal, without relying on cell fitness or mutations that compromise this process. For example, function can be derived from 1) how protein abundance or localization is affected under specific conditions or 2) how abundance or localization of a marker is affected by loss of any particular gene. Since the creation of the GFP collection (Huh et al., 2003), a number of studies have looked at the abundance and localisation of proteins under specific conditions genome-wide (Torres et al., 2016). In addition, some studies have combined endogenously or plasmid-based tags in genes of interest with arrays of mutants by taking advantage of 'synthetic genetic array' (SGA) or an systematic hybrid LOH method (e.g. Alvaro et al., 2007; Tkach et al., 2012). Finally, quantitative methods have been described to identify foci and quantify fluorescence (e.g. Joglekar et al., 2006; Gonzalez et al., 2012; Herbert et al., 2014).

Initially, the major bottleneck to these studies was the difficulty to capture thousands of images, however, this has largely been solved with new automated imaging platforms. Now the challenge is image analysis, converting a picture into rigorous quantitative data, which was previously undertaken laboriously by eye – as so called 'eye-throughput' approach

(Torres et al., 2016). As well as being time consuming, this type of analysis remains subjective since an investigator's ability to score a phenotype may change over the weeks required to assess a genome-wide screen. However, most automated analysis methods are designed to find predicted phenotypes such as changes in distribution or fluorescence intensity of a tagged protein of interest. Automated computational analysis able to detect any difference in fluorescence levels or localization from a wild type cell would be of great value for large-scale screens.

It was only recently that a large-scale study successfully combined SGA and image-based high-content screening with automated analysis (Chong et al., 2015). This study assessed general localisation and abundance of GFP-tagged proteins and how this changes in time and under specific perturbations in the cell. This study adapted the CellProfiler software (Carpenter et al., 2006) to broadly identify changes in fluorescence in different conditions. In the current issue of Cell Systems, Styles et al. take this approach one step further by optimizing a pipeline that combines genome-wide genetic and chemical perturbations with image-based high-content screening (Figure 1A). They apply this approach to a specific process with a readout that represents a direct outcome instead of a change in abundance or localisation of a specific protein.

Styles et al. use pattern classification through machine learning to identify mutants that affect a specific cellular process, DNA damage repair, by monitoring fluorescently tagged Rad52. In budding yeast Rad52 – a key protein in the repair of DNA double strand breaks, is distributed throughout the nucleus but forms foci at sites of active DNA damage repair (Lisby et al., 2001). They first use SGA technology to create arrays of single and double mutants containing fluorescent nuclear and cytoplasmic markers for spatial and cell cycle references and fluorescently-tagged Rad52. Then, they devised a *Support Vector Machine* (SVM) training algorithm to classify cells according to the presence or absence of Rad52-GFP foci as a readout of DNA damage.

By carrying out their imaging screen in specific mutant strains they are able to screen specifically for one of two DNA repair mechanisms, Non-Homologous End-Joining (NHEJ) or Homology-directed repair (HR). Specific mutations disrupt one of these pathways, which allow screening for genes involved in the intact repair pathway. In addition, they look at single mutants with and without treatment with a chemical that breaks DNA, phleomycin. Their combination of mutations and chemical perturbations allows for identification of genes and pathways that act under specific conditions and would be undetected under a single mutation or perturbation background.

DNA damage repair (DDR) has been extensively studied, which has provided a wealth of data on the proteins involved in this process. In this respect, DDR represents a good proof-of-principle for the current study since many of the genes involved in DDR have already been identified and so the veracity of the new technique can be compared with previous studies. The authors identify 345 mutants out of the ~5000 mutants screened - including genetic and chemical perturbations - with elevated levels of DNA damage. The overall list of genes is highly enriched for genes annotated to be involved in 'DNA repair', 'DNA replication', 'homologous recombination' and 'cohesion' and in addition, some are only found in backgrounds that are compromised for the NHEJ (enriched for abnormal telomere size) or HR pathways (enriched for DNA metabolism), which might have been missed due to genetic redundancy. In aggregate, this study represents a major development in functional genomics by providing automated discriminating analysis to large image datasets.

Rad52 has been the subject of previous functional genomics studies. Comparing the observations of Styles et al. to this previous work reveals the importance of screening conditions. Surprisingly, the overlap with a previous functional genomics study, identifying mutants that had increased numbers of Rad52 foci, was limited (Alvaro et al., 2007). At this point the reason for this is unclear, but it could be due to a multitude of technical issues and to different methodology (for example, the previous screen was performed with hybrid diploids). However, Styles et al. correlate increased Rad51 foci with a decrease in fitness in single mutants, which has been the readout for many previous screens (colony size), lending support to their conclusions about biological relevance of their image-based phenotypes.

Future application of Styles et al.'s approach to cellular processes, which when deregulated do not necessarily cause a decrease in cell fitness, is expected to be a major driving force for new discoveries (Figure 1B). Overall, the ability to discover phenotypes via automated computational analysis is significant since it allows detection of alterations in fluorescence levels or localization in an otherwise wild type cell. This essentially provides a discriminating analysis akin to a human operator to large image datasets. The SVM training algorithm developed by the authors is focused on detecting the presence of at least one or more DNA damage foci within a cell and showcases the power of this approach. Adapting this classifier to detect, count and quantify different fluorescence features throughout the cell it will open up a whole new world of discovery.

Figure 1. Automated computational image analysis of fluorescently-tagged proteins in budding yeast allows query of any particular cellular process by tracking a fluorescent signal. A) By monitoring fluorescently tagged Rad52, Styles et al. screen for proteins involved in DNA damage repair. Rad52 can be found distributed throughout the nucleus but forms foci at sites of active DNA damage repair. An increase in Rad52 foci therefore indicates DNA repair deficiency. B) Any specific cellular processes can be queried by imaging a fluorescently tagged protein that marks the particular process. HC: High Content, SGA: Synthetic Genetic Array.

References

Alvaro, D., Lisby, M. and Rothstein, R. (2007). Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. PLoS Genetics, *3*(12), 228.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J., Ou, J., San Luis, B., Bandyopadhyay, S., Hibbs, M., Hess, D., Gingras, A., Bader, G., Troyanskaya, O., Brown, G., Andrews, B., Boone, C. and Myers, C. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nature Methods, *7*(12), 1017-1024.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P., Sabatini, D.M. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol., *7*(10), R100

Chong, Y., Koh, J., Friesen, H., Duffy, S., Cox, M., Moses, A., Moffat, J., Boone, C. and Andrews, B. (2015). Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. Cell, *162*(1), 221.

Gonzalez, J.E., Lee, M., Barquinero, J.F., Valente, M., Roch-Lefevre, S., and García, O. (2012) Quantitative image analysis of γH2AX foci induced by ionizing radiation applying open source programs. Anal. Quant. Cytol. Histol., *34*(2), 66-71.

Herbert, A.D., Carr, A.M. and Hoffmann, E. (2014). FindFoci: A Focus Detection Algorithm with Automated Parameter Training That Closely Matches Human Assignments, Reduces Human Inconsistencies and Increases Speed of Analysis. PLoS ONE, *9*(12), 114749.

Huh, W., Falvo, J., Gerke, L., Carroll, A., Howson, R., Weissman, J. and O'Shea, E. (2003). Global analysis of protein localization in budding yeast. Nature, *425*(6959), 686-691.

Joglekar, A., Bouck, D., Molk, J., Bloom, K. and Salmon, E. (2006). Molecular architecture of a kinetochore-microtubule attachment site. Nature Cell Biology, *8*(6), 581-585.

Lisby, M. Rothstein, R. and Mortensen, U.H. (2001). Rad52 forms DNA repair and recombination centers during S phase. PNAS, *98*(15), 8276-8282.

Tkach, J., Yimit, A., Lee, A., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J., Ou, J., Moffat, J., Boone, C., Davis, T., Nislow, C. and Brown, G. (2012). Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nature Cell Biology, *14*(9), 966-976.

Torres, N.P., Ho, B. and Brown, G.W. (2016). High-throughput fluorescence microscopic analysis of protein abundance and localization in budding yeast. Crit. Rev. Biochem. Mol. Bio*., 51*(2), 110-119.