# 1 What to Do with All the Data?

High-content screening can easily generate more than one Terabyte in primary images and metadata per run, that have to be stored and organized, which means an appropriate laboratory information management system (LIMS) has to be established. The LIMS must be able to collect, collate and integrate the data stream to allow at least searching and rapid evaluation of the data. After image acquisition and data transfer, image analysis will be run to extract the metadata. Further evaluation includes testing for process errors. Heat maps along with pattern recognition algorithms help to identify artefacts such as edge-effects, uneven pipetting, or simply to exclude images that are not in focus. All plates should be checked so that the selected positive and negative controls exhibit values in a pre-defined range. Further, data may be normalized against controls before further statistical analysis is run to identify putative hits. Known proteins of the pathway being screened should score, and are a good internal control for the accuracy of the assay and workflow. Hits have to be verified by going back to the original images. Further, results have to be compared between independent runs. After this, an appropriate hit verification strategy has to be applied as discussed above. Target gene expression should be confirmed, for example, by running a microarray analysis of gene expression for the given cell line. Finally, data will be compared to other internal and external data sources. Cluster analysis will assist in identifying networks and correlations.

A critical aspect of high content screening is the informatics and data management solution that the user needs to implement to process and store the images. Typically multiple images are collected per microplate well at different magnifications and processed with pre-optimised algorithms (these are the software routines that analyse images, recognize patterns and extract measurements relevant to the biological application, enabling the automated quantitative comparison and ranking of compound effects) to derive numerical data on multiple parameters. This allows for the quantification of detailed cellular measurements that underlie the phenotype observed. From an image analysis perspective the following should not be overlooked when reviewing vendor offerings: the breadth of biology covered; how the software is delivered, does it run quickly, or open a script; is analysis done on-the-fly or offline; have the algorithms been fully validated with biology; the ease of exporting image files to other software packages; and access to new algorithms, is the user dependent on the supplier or is it relatively easy to develop your own or adapt existing algorithms?

The key theme and piece of information repeated throughout this chapter is "partnering". Scientific research and informatics must work together for the mutual benefit of screening like the drug discovery process. To really be part of the winning team in any organization, all areas must bring their collective expertise together and make the extra effort to understand one another and defer where there is lack of knowledge to those on the team with the experience and expertise or to seek external advises. It is necessary to start off by setting the stage concerning where laboratory computing, which includes the data management (we will discuss a bit later in the chapter), has progressed in order to gain the necessary understanding of where it currently is and where we anticipate it will be going in the HCS area in the future.

A goal of this chapter is to provide an overview of the key aspects of informatics tools and technologies needed for HCS, including characteristics of HCS data; data models/structures for storing HCS data; HCS informatics system architectures, data management approaches, hardware and network considerations, visualization, data mining technologies, and integrating HCS data with other data and systems.