21

Large-Scale Data Management for High Content Screening

Leon S. Garfinkel

Summary

High content screening (HCS) plays an important role in target selection in primary and secondary screening, but further developments in informatics and data management are needed for strategic implementation of HCS in the drug discovery process. An organization charter for the Research Informatics and Infrastructure Organization is described and is consists of four basic parts: Partner, Build Trust, Champion, and Core vs Noncore. The successful evolution of the charter over the last 5 yr is mapped using high-throughput screening and HCS data as an example. A future view of large-scale data management for the drug discovery process will incorporate all scientific information into multiple parameter type runs for many aspects of the science. This information will subsequently be aligned into a subset that an individual can digest and more easily choose the next appropriate steps.

Key Words: Bioinformatics; data integration; information technology; large-scale cell-based assays; platform independent.

1. Introduction

The importance of high content screening (HCS) in the drug discovery process is target selection in primary and secondary screening but it is only in infancy from the standpoint of informatics and data management. As the imaging equipment in the laboratories becomes more sophisticated; produces better images, more graphics with higher resolution, and allows for faster collection of data, the infrastructure specialists working in the background will not only have a hard time keeping up in some organizations, but might at times ask the scientists to stop while they catch up. Where does one want your HCS informatics organization to be? Ideally, one wishes to be a step ahead of the drug discovery process and taking a strategic view of this area as opposed to being in a purely operational tactical solution oriented mode, which will need daily management. How does one get to this point, what are some of the options, methods, and proven answers that will get one ahead of the curve? The author will try and address as many of these issues as possible and convey the real hands on information to make it happen for your organization.

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 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 advise. 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

From: *Methods in Molecular Biology, vol. 356:*
High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery
Edited by: D. L. Taylor, J. R. Haskins, and K. Giuliano © Humana Press, Inc., Totowa, NJ

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.

First let us examine a brief bit of history and background that will first put in place the foundation for computing in a Scientific Research environment as opposed to general computing in business. In 2000, out of all the departments in the Roche Nutley Research organization, (chemistry, biology, four therapeutic areas, and screening), there were two items that stood out which needed immediate attention. First, there was no single primary user of computer systems and, second, all of the departments combined had accumulated almost two terabytes of storage. In 2000, we were aware of the inefficiencies of trying to maintain multi-vendor computer systems and we anticipated a huge upswing in the amount of data storage space needed by the servers to handle the increases coming from the laboratory equipment. Although most of the information contained here might be specific to our organizational history, the author will certainly include examples and relevant comments from peer organizations in which the same situations might have occurred being better or worse. Some of the typical environmental situations that were contained within the labs:

- Computers were stored under laboratory work benches.
- Add-on disk drives were stacked up on cardboard boxes.
- Network connections went from some slow speed LAN to internal dial-ups.
- Servers were stored in closets in which space was available.
- Backups were done sporadically or not at all.
- There were published incidences in the industry in which tapes or other valuable/confidential backup computer documents were tossed into whatever available drawer space there was or even tossed out into dumpsters by accident.
- A mixture of Macintosh and PC type computers were in use, to the delight of scientists.
- Highly specialized, nonstandard computers that required extensive hands-on maintenance were in use throughout the facility.
- Large CRT monitors took-up valuable laboratory counter space.
- Scientists were doing science as well as Informatics work.

The first step in altering the situation was the realization that the research department had very specialized computing needs, which were very different from every other division of the company. Sales and marketing, finance, human resources, and even to some degree manufacturing departments were able to make use of off-the-shelf computer software and hardware. Scientific research on the other hand was not able to use of off-the-shelf items in most cases; there were requirements for special hardware to connect to instruments, special software for collecting, manipulating, synthesizing and managing the data coming from those instruments. Once this realization took place, an infrastructure and informatics organization was formed inside of the research department, reporting to Research to handle all of these issues.

2. Organizational Structure

There are four basic parts to the charter for the research informatics and infrastructure organization, they are: partner, build trust, champion, and core vs noncore.

2.1. Partner

Partner with the scientists to allow IT and IM professionals to take over responsibility for informatics tasks, thus, proving to the scientists that the informatics staff could successfully handle all of the informatics duties for the scientists. This would not only allow everyone's expertise to shine but would also return a large amount of time to the scientists, whom in the past were making extensive use of their time for informatics work. The side-by-side work would allow the informatics organization to learn about the workflows of the scientists, so the computer processes could be placed inline with the workflow and not be forced to add extra steps to the workflow.

2.2. Build Trust

Build the necessary level of trust with research management and scientists to ensure that the desired partnership "bears fruit."

- Do the installations of computer equipment for the scientists on time.
- Write the necessary utilities and programs to meet the needs of the scientists.
- Be diligent in purchases because whatever money is spent on IT cannot be spent for science.
- Require perfection from the informatics staff.
- Have the scientists validate the informatics work and sign-off that it meets the specified requirements.

2.3. Champion

Champion the research department's needs with corporate IT and IM and take full responsibility as the intermediary between these parts of the organization. Own the responsibilities of being in the research organization, yet being an IT professional.

- That is, system maintenance on databases can only be done twice a year rather than four times a year as the rest of the organization does, because of the nature of programs in the research department and the fact that programs might be executing for weeks at a time.

2.4. Core vs Noncore

Defining Core vs Noncore activities for the Research Informatics area. Does it make sense for Research to own, operate, and manage a data center? It makes sense because Corporate IT, already has the market on this, can provide the service for the RO also. It allows Research to take advantage of economies of scale. When an organization decides to share commodity services, there are security polices, practices, and standards available that are good enough for the entire company? Do they work for the research department also? In 99% of the cases the answer is "yes." We can make use of these governances; however, where it is not appropriate, research IT must have the latitude to break the rules.

2.4.1. Core Activities

- What type of desktop and servers are required for scientific computing? (brand, type, speed, and memory)
- How do the computer systems interface with the necessary data collection instrumentation and "talk" to the network and servers at the same time?
- How is the data stored, managed, and protected for short-, medium-, and long-term use?
- Are there government regulatory requirements that need to be met?
- Is the required software available off-the-shelf or must it be written in-house? This decision is collaboration between IT and scientists based on the defined requirements.
- Can allowances and accommodations be made for external collaborations and programs shared among scientists? Will we in the corporate world make the necessary adjustments to the IT systems so the scientists can freely collaborate with their peers in the academic world, whose computer systems might not be up to the standard we have set for security and virus protection? Can we create a safety buffered zone outside of our firewalls to allow this external data exchange?

Answering the earlier questions tells one exactly which business you want to be in and what type of work you want to be doing. This obviously needs the cooperation of the entire organization and all aspects of it. It re-enforces that there are experts in different areas and although scientists are highly skilled and learned, they do not have the specific professional knowledge that people in the computer industry have about making the correct business decisions related to IT. In today's day, in which just about everyone has a computer at work and at home, we tend to think of this not as a profession but as a hobby. If we bring our hobby to work with us, we might even save some dollars by not calling on the professionals. Wrong attitude and thought process, going down this path gets everyone in trouble. There is a world of expertise beyond the small domain of home computing that gives professionals that edge to do strategic business planning

as well as tactical operations for the benefit of the organization. It is end-to-end thinking rather than dealing with a single isolated situation.

2.4.2. Noncore Activities

- Physical data center along with facilities management (electricity, air conditioning, fire protection, physical security).
- Networking.
- Security.
- Backup and recovery.
- Standards for PCs and peripherals, servers, desktop applications, middleware applications, web standards.

2.5. Organizational Summary

The accomplishments of RO/IT partnering actually amazed many people on both sides of the organization. The Informatics people never thought it would be possible to gain the acceptance, respect and ultimate responsibility to take computing away from the scientists. On the other hand, the scientists did not believe that anyone could do as good a job as they had been doing, no matter what level of expertise people had. Additionally, this opportunity ultimately allowed a number of scientists, who no longer wanted to work in labs to take their expertise and become computer experts and start new careers. There are a number of people in the organization that were PhD chemists and are now UNIX system administrators, Visual Basic programmers, project managers and workflow experts. Below are a number of outcomes that this effort was able to "bear fruit on":

- Standardized desktop systems (hardware and software) were in place throughout Research including the same application suite for all (i.e., Microsoft Windows XP, Microsoft Office, antivirus software, and Web access and portals).
    - o All users defined and categorized to ensure proper equipment (super users are equipped with very high-end pc or workstations, normal users, users traveling/working from home are equipped with laptops or notebooks).
- All servers are located in the corporate data center.
- Backups are done on a scheduled "off hours" basis to minimize work disruption.
- Data storage is part of the enterprise storage program used by the entire corporation, but with the understanding that the research department is the single largest user and might have special requirements.
- Most laboratories have only the minimum amount of computer equipment located on the actual premises now (flat panel monitors, 100 megabyte or gigabyte network connections).
- Partnership between research informatics and corporate informatics with a new understanding of requirements and demands in both parts of the organization.
- Scientific hardware/software vendors are now partnering with computer hardware/software vendors as a result of us sitting in the middle and demanding the best of all possible solutions. This type of partnering is something we in the IT field have been doing for years. The author was told by his colleagues on the science side that this partnering is something new and they are not used to working with Vendors in this type of fashion. The bottom line here is we challenged all of our "business partners" not vendors to come work with us on what for us was a unique set of initial requirements—a solution for data archiving of HCS imaged data. However, it turned out once the solution was tested and put in place that many other organizations had the same need for this solution.

The accomplishments described above took approx 3 yr to put in place. The infrastructure (hardware—desktops, servers) staff was six people and Informatics (software—packages, code writing) staff was approx 11 people. A basic operating expense budget for this type of area was approx $5–8 million annually and a capital budget of approx $2–5 million annually. This was to support approx 250–300 scientists and all of their related laboratory computer equipment and data collection. This work set the foundation for all of the integration of laboratory equipment and computer systems to follow. It allowed almost a cookie-cutter approach to future solutions that would be required.

3. Implementation Decisions

The author would also like to include information regarding some of the considerations that we decided not to implement and why. Along with this some insight concerning why some of these decisions where right for this organization but might not be right for every organization or how you decide on which bits and pieces to proceed with.

- Turn all servers over to corporate IT for system administration—lack of specialized knowledge as previously discussed. There is scientific instrumentation at the end of these systems and the applications running on these computers are closely coupled to these instruments.
- Security/virus protection needs required some very special considerations—the potential to have an effect on the entire enterprise must be considered and, therefore, requires special handling.
- Rejection of standards for PC and monitors resulting from specialized requirements, such as high-resolution monitors/flat panels so scientists are able to view images.
- Setup email servers and manage these resources—the suggestion that science has special requirements when it comes to this type of application might or might not be justifiable true. It is, however, a no win situation. Do not get involved in this headache, leave it to corporate IT to do this for the entire enterprise.

There were two schools of active thought that were required to accomplish this entire mission. The first being a tactical/operational, for example, "the author has a situation now that must be addressed, to ensure operational integrity, systems continuing to run as well as a nondisruptive approach." The second being a strategic view, for example, "Whatever the author does must be as generic as possible, must be portable within the United States and globally, must be cost effective in the long-term and must address needs both current and looking into the future."

The tactical exercises had to address the deficiencies that existed in the immediate environment, to allow the scientists to continue their research, whereas the informatics organization put the necessary components in place to allow for the master plan. We added short-term cheap disk drives, while implementing an enterprise-wide storage subsystem that would allow us to grow by terabytes a year. We then added memory to local PC or servers, while installing servers in the data center, in which where SAN attached to the necessary storage, they had full networking capabilities and a backup silo for backup and recovery. The strategic view was a vision of what Informatics needed to do, to enable a true partnering relationship with Research as well as put in cost effective solutions. This part of the solution is what really took the 3 yr to accomplish. The IT teams were busy justifying costs for new servers and peripherals to management, showing the long-term payback of the solution and then implementing the solutions to show the scientific gains.

4. Specific Example: High Throughput and HCS

High-throughput screening (HTS) and HCS was the first major request received by Research Informatics. This was a request to attend vendor meetings, get an understanding of the product, and meet the vendor technical staff both from a science perspective as well as an informatics perspective before purchasing the laboratory equipment. In addition, it was an opportunity for informatics to really "break bread" with the scientists and partner for the success of the project. The goal was that the project would not be broken into separate components (i.e., a research and informatics portion); it would cover the project as one project including costs, resources and timelines from start to finish.

Some of the more interesting requirements of this first project were:

- The primary device collected imaging data instead of collecting statistics.
- The data needed to be reviewed only by the HTS area.
- The possibility that the data/visualization needed to be available across sites locally and globally.
- The data retention depended on specific studies.

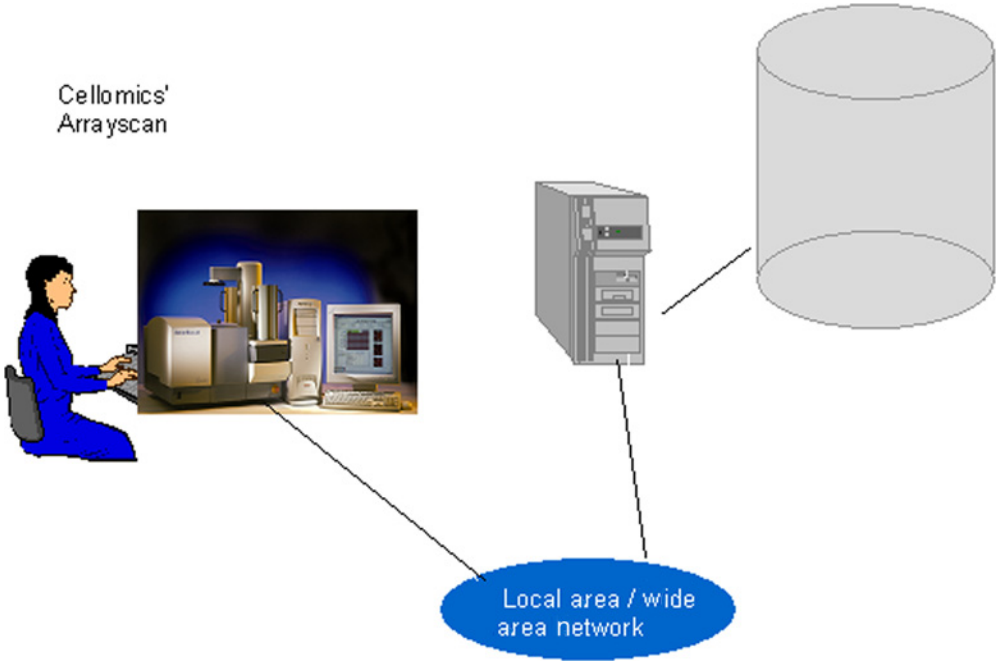

Fig. 1. Single site single instrument configuration. (Please *see* the companion CD for the color version of this figure.)

- The default microliter plate density was 384 and not 96.
- Cellomics, Inc. (Pittsburgh, PA) had written a proprietary database with their software.
- The assumption was that both the scientists and the vendor of this equipment felt that they would be able to store 1 terabyte of data in the laboratory in a storage array located along side a server under the workbench.
- Biological applications that were covered in the initial rollout:
  - o Cytoplasm–nucleus translocation.
  - o Cell viability.
  - o Multiparameter apoptosis 1.
  - o Neurite outgrowth.

Figure 1 shows single physical location, a single user, a single instrument, a single server, and single set of files. As described earlier, this was our first "baby steps" into taking clearly defined user and vendor requirements. Ultimately, this brought the integrated requirements into existence and made a production system out of the specifications. As time continued technology improved, requirements changed and within the next 2 yr we had a new set of requirements:

- The data needed to be reviewed not only by the HTS area but by therapeutic areas as well.
- The data/visualization needed to be available across sites (locally and globally).
- The data retention depended on specific studies and review of materials obtained in primary as well as secondary screening campaigns.
- There was a planned migration from 384-well plates to 1536-well plates.
- The vendor had rewritten their database to now use Oracle of Redmond, CA.
- The vendor had rewritten its software product, partly based on the recommendations and suggestions of the Roche Scientific and Informatics communities.
- The data growth was anticipated to be as much as 2–4 TB annually.
- Biological applications that were covered at the 2 yr point:

- Cytoplasm–nucleus translocation.
- Cell viability.
- Multiparameter apoptosis 1.
- Neurite outgrowth.
- Cell motility.
- Cell cycle.
- Cell spreading.
- Compartmental analysis.
- Cytoplasm–cell membrane translocation.
- GPCR application.
- General screening application.
- Micronucleus.
- Mitotic index.
- Molecular translocation.
- Multiparameter cytotoxicity 1.
- Extended neurite outgrowth.
- Receptor internalization.
- Target activation.
- Cell health profiling.
- Morphology explorer.

At this point, a more difficult part came into play. After many discussions with Cellomics, Inc. personnel (the maker of the Arrayscan device) and then a number of conversations with our primary storage vendor EMC (Hopkinton, MA) we challenged the two vendors to team up and come up with a solution that would address:

- The continuing growth requirements of collecting imaging data.
- Containing costs of data storage that anticipated growing in an exponential fashion.
- The total amount of data kept (because it is only required for a limited amount of time).
- Ensure whatever technology is used that there is no difference in access time.
- Creation of software components within the Cellomics software to do the necessary archiving and image migration in some automated fashion.

The solution that we ultimately brokered and came up with consisted of the addition of a single piece of new equipment: a device from EMC called the Centera. Cellomics, Inc. then wrote a specific exit for their programs, which allowed for the archiving of the data as the scientific community saw fit. The solution allowed for the data to reside and be accessed off of the primary tier one-storage platform as it was gathered and came off of the Arrayscan. Then, as determined by the scientists, it would seamlessly migrate to the Centera, which is a tier two storage platform. The Centera was slightly slower, but much cheaper than alternative technologies. Also it used a different type of technology for ensuring availability of the data. A scientist accessing a piece of data would not know in which it was coming from as the system behind-the-scenes obtained and returned the data in the same relative time.

Some of the immediately identifiable and evident benefits from the work the entire team had done were:

- The users were able to archive data-based on their specific business applications while adhering to the IT policies.
- The security of the images was insured by the methods or storage.
- Backup/recovery solutions were put in place.
- Short-term operational goals were met and long-term storage goals were met.
- The timelines for screening and rescreening as well as review of the data and the ability of data sharing were met.
- The solution offered scalability.

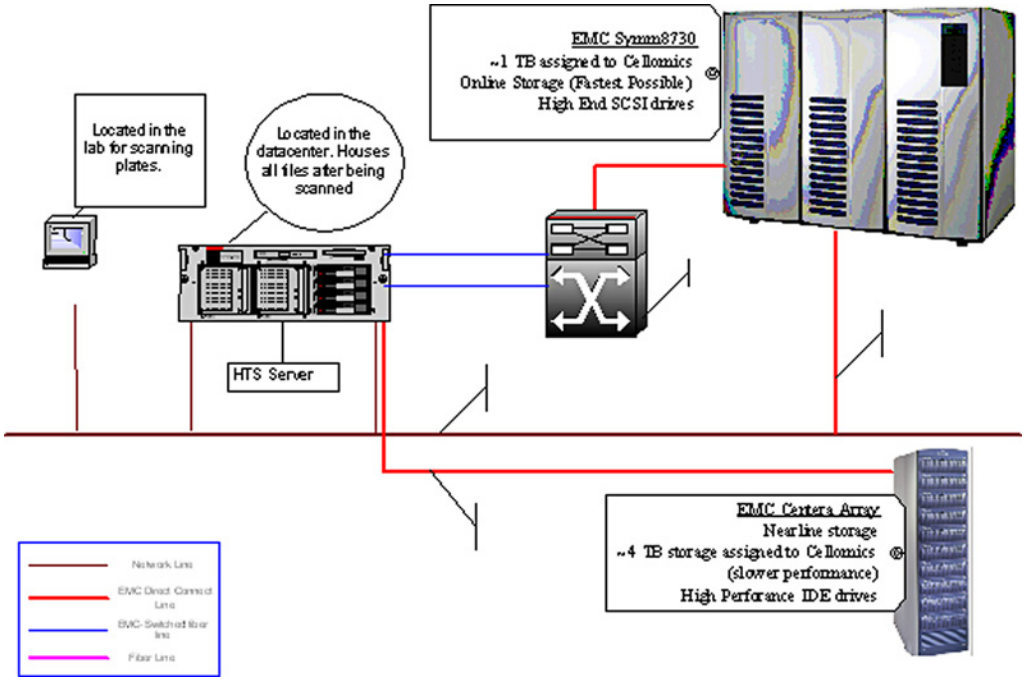

Fig. 2. Currently existing infrastructure configuration. (Please *see* the companion CD for the color version of this figure.)

- This solution also enables entire lifecycle management of the data as it moves through the stages of the scientific process.
- All regulatory issues for data collection and retention were being met.

Figure 2 was implemented 2 yr after the original configuration and incorporated the new set of requirements. It consisted of a single physical location, multiple users, a single instrument, a single server, a database and integrated data management solution included. The solution allowed for the data to reside and be accessed, off of the primary tier one storage platform, as it was gathered from the ArrayScan. As determined by the scientists, the data then would seamlessly migrate to the Centera, which is a tier two storage platform.

Some additional considerations that were part of the decision process surrounded the issues of database technology. Although there were corporate standards in place for different size databases or databases that had intended numbers of users transaction types, we had some very specific scientific HCS needs as well. In addition to Oracle for the Cellomics, Inc. data we, used IDBS ActivityBase (Gilford, UK) to store experimental data from the instrumentation. The visualization tool used for viewing the data is Spotfire (Summerville, MA). In addition, some organizations have any number of very specific "home written" tools and utilities that enhance the scientific experience.

Figure 3 is a schematic of our Cellomics, Inc. HCS data movement through the system. Our database overview and some of the software tools described are show in the figure.

Figure 4 shows a very high level view of the entire organization and how systems and data are organized. We have attempted to ensure that, as a global organization, no matter where one physically is, that person should be able to access their required data. This has been put in place over 3–4 yr and is a multi-million dollar network with very sophisticated security and technology in place to allow only authorized users to their specific data.

Figure 5 is the cellomics data handling of all multiparameter bioapplications.

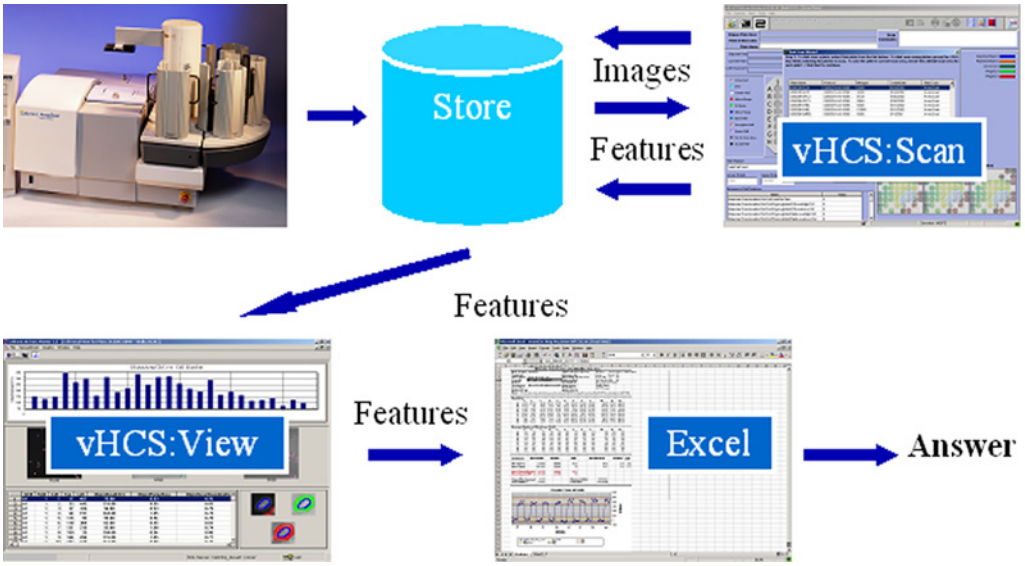

Fig. 3. Current features of Cellomics database management system. (Please *see* the companion CD for the color version of this figure.)

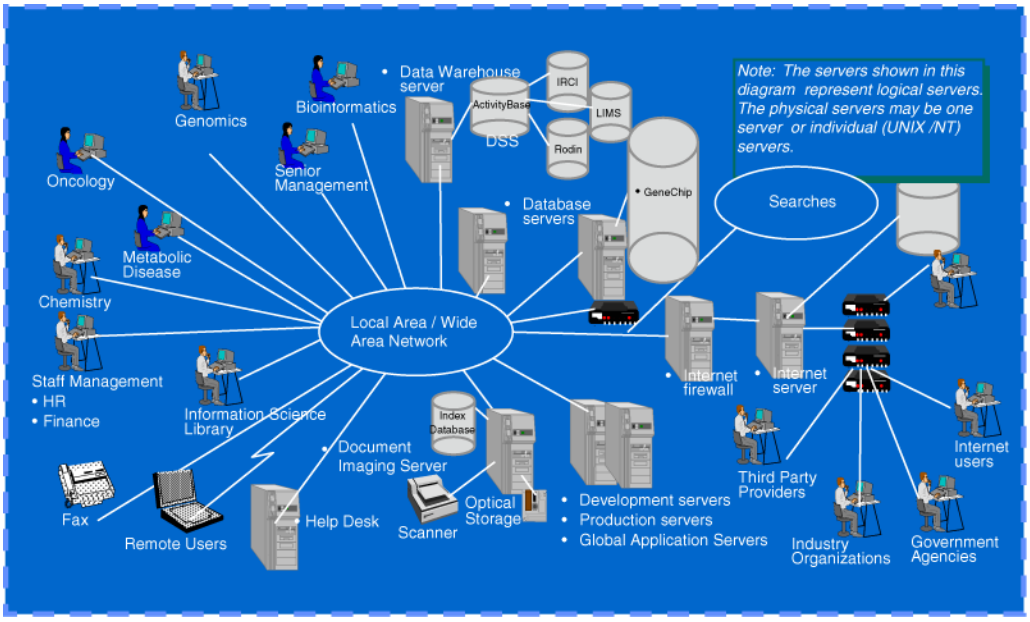

Fig. 4. High level logical view of the information technology architecture. (Please *see* the companion CD for the color version of this figure.)

The solutions depicted here were engineered with some of the following concerns in mind:

- As little daily operator intervention as possible.
- Cost effective in the short term as well as the long-term.
- Scalability to ensure that the solution can grow as the science changes with minimal intervention.

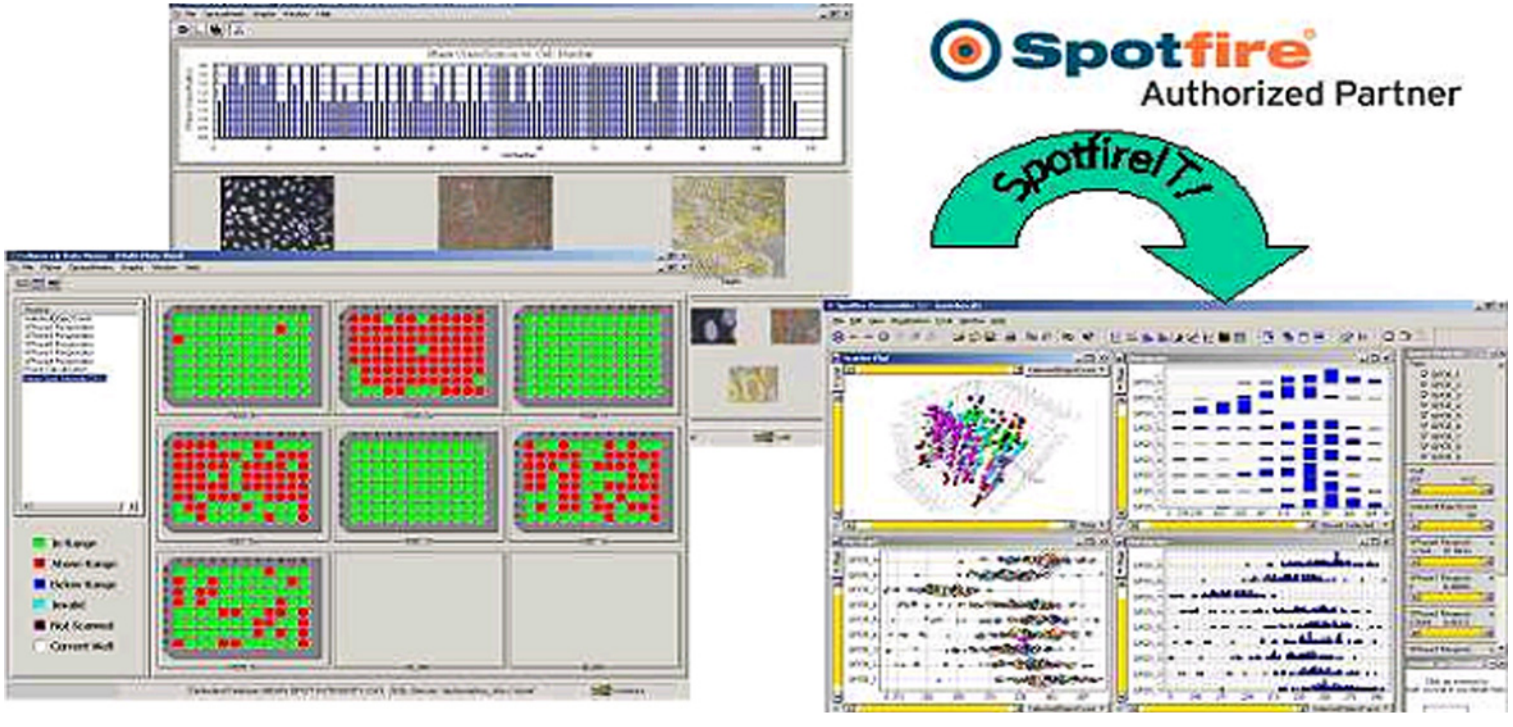

Fig. 5. Cellomics Data Handling of All Multiparameter Bioapplications. (Please *see* the companion CD for the color version of this figure.)

- Flexibility to ensure that additional scientific instrumentation could be integrated with little or no involvement from the Informatics area. This solution has allowed almost any device to be brought into the organization for testing and or permanent placement with minimal intervention by the Informatics organization. All of the information below is from projects already undertaken and proven within the environment.

5. Summary and Future Consideration

Where are we today, about 5 yr later? Scientific research is now consuming approx 12–16 terabytes of data. We seem to be adding around 4 TB each year. The single largest user, group amassing data is the HTS area. This is not only because of their increasing the number of wells on a plate, but as you can see by the instruments below the quality of the data is becoming much more robust. For the images being stored, studied and passed on to other areas, the data sets are getting larger and larger. One additional piece of information that can easily facilitate the entire process of implementing good Informatics practices and having a proactive organization to ensure the proper strategic planning for growth and technology is a user advisory committee or steering committee. This is a low cost option and should have full management support to ensure the success.

The mission of this steering committee is to:

- Assure alignment of Informatics activities with local site and global priorities.
- Recommend prioritizations of activities to senior management.

The members of this group:

- Have a strategic view of the interactions between scientists and Informatics.
- Represent the overall needs and concerns of their constituents as well as the site.
- Membership should include—biology, chemistry, HTS, genetics, genomics, bioinformatics, therapeutic areas (3), pharmacology, safety, animal resources, informatics, finance, senior management.
- Be able to make decisions on operational issues and not for the sole benefit of individuals.
- Improve communication regarding requirements and deliverables.
- Pro-actively provide management with information that will have an effect on deliverables and on current projects.
- Provide transparency around resource allocation as well as costs.
- Dedicate the time to the meeting and bring information back to your department as well as to the next committee meeting. Once a month was sufficient for our organization but this can be adjusted based on topics and needs.

Attempting to look into the future, we hope to be able to see that the technology used in science will not only continue to improve, from the standpoint of what the scientists are doing on a single biological or chemical level, but will incorporate all of the information into multiple parameter type runs for many aspects of the science. In other words, while the author performs screening, can a scientist combine the biological analysis and chemical analysis and also do some mapping? include genomics information and genetics. The ultimate goal would be to align all of this information into a subset of the information that one individual can digest and relate patent information and competitor information. Armed with such information, one can more easily choose the next appropriate steps.

Acknowledgments

Special thanks to Dan Weiss, Keith Groco, Ann Hoffman, and Ralph Garippa, who knew the value of collaboration and realized that this was the team that would make the projects a success. Cellomics and EMC for not being mere vendors, but partnering with Roche to come up with a unique solution to our problem. Irene, Rebecca, and Mark for allowing me to spend all the necessary time at work and work at home when possible to ensure that these projects were a success.