Teaching materials

A brief introduction to DNA micro-arrays

Rasmus Wernersson

Center for Biological Sequence Analysis
Institute for Systems Biology, Building 208
Technical University of Denmark
DK-2800 Lyngby, Denmark

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Preface

This brief introduction to DNA micro-arrays is meant as a short primer for teaching purposes. It's written based on my own lab experiences conducting research on the Yeast cell cycle, and this has a focus on the *experimental side of the matter*. It's based on various texts I'm written earlier, and condensed into this document for the use the **27040: Introduction to Systems Biology** (Bachelor level) in the spring of 2013.

Notice that this document is **not** meant to cover DNA micro-arrays in all details (this would require an entire book), but rather to be a primer on the subject, and to support the lecture slides.

The PDF is formatted for duplex (two-sided) printing/viewing.

Rasmus Wernersson
March 2013

Contents

Chapter 1

A brief introduction to DNA micro-arrays

The DNA micro-array technology allows the measurement of abundance of a large number of transcripts[1] in parallel. The micro-array technology can be seen as an extension of the well-known Northern blotting technique (Maniatis et al., 1982), but with a vastly improved potential for measurement quantity and quality. Whereas Northern Blotting relies on separation of the RNA fragments on a gel and subsequent hybridization with a radio-labeled probe targeting the sequence in question, the DNA micro-array technology uses the approach of fixing a large number of sequence specific probes on a surface, in a grid-like fashion, and letting labeled target molecules hybridize to the immobilized probes prior to detection.

The labeling used in DNA micro-arrays are fluorescent dyes which can be excited by a light-source (laser or lamp) and caused to emit light at a specific wave-length. Two commonly used dyes are Cy3 and Cy5 (see Figure 1.1 for the molecular structures and emission spectra). Following hybridization on the DNA micro-array, non-bound target molecules are washed away, and the array is placed in a scanner which will record the amount of light emitted from the bound target molecules. By mapping the intensity of light emitted from each individual position on the probe-grid, it is then possible to estimate the *relative* intensities of the transcripts targeted by each probe.

1.1 Main types of DNA micro-arrays

Spotted arrays: The first DNA micro-arrays used full length cDNA molecules (amplified using PCR) as probes. The probes where *spotted* onto glass slides (roughly like microscope slides, see Figure 1.4 [panel A]) with an activated surface capable of binding the DNA (Schena et al., 1995). The spotting process is the process of placing small droplets containing the cDNA probes in a organized grid on the micro-array slide, see Figure 1.2. For automating the spotting and increasing the number of probes that can be spotted, spotting robots have been developed. Spotted arrays can equally well be used with

[1] Mostly the DNA mirco-array technology is used for expression profiling, but the micro-array technology can equally well be used with DNA as the staring point, for example when using genotyping arrays.

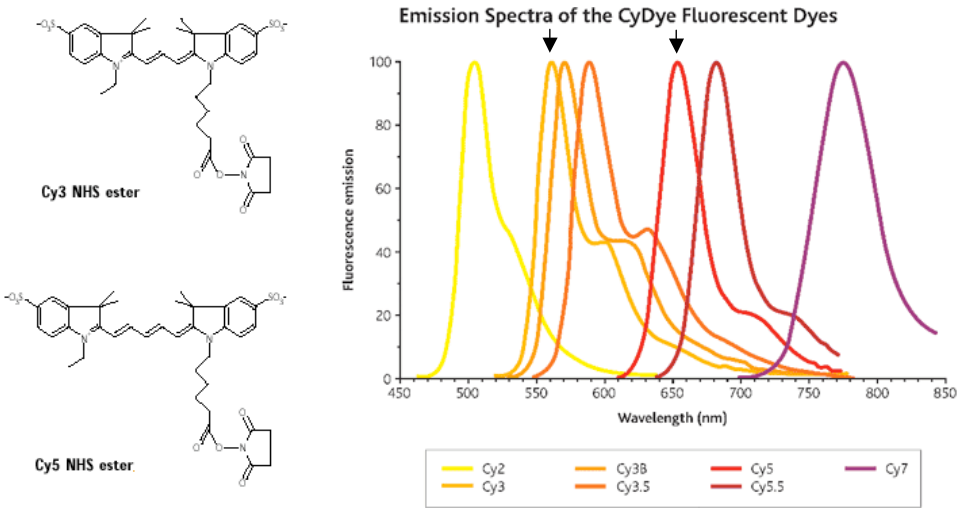

Figure 1.1: **Molecular structure and Emission Spectrum of Cy3 and Cy5.** Cy3 and Cy5 are both members of the "CyDye" family of fluorescent dyes and have an almost identical structure, only differing in the number of carbons in the linkage region. As shown in the emission spectrum graphs, Cy3 and Cy5 (indicated by the arrows) are well suited to be used in combination, since the overlap is small. [Source of images: the manufacturers (Amersham BioSciences) website.]

synthetic oligonucleotide probes.

In situ **synthesized arrays:** A completely different approach to the construction of DNA micro-arrays is to synthesize the probes directly on the surface of the array. The approach was initially commercialized by the company Affymetrix (California, USA) under the name "GeneChip" (the approach is demonstrated in Lockhart et al. (1996)). The Affymetrix method of synthesis relies on light-deprotection of the growing DNA oligonucleotide and is shown in details in Figure 1.3 - (see also, Fodor et al. (1991)). The idea is here to build the oligonucleotide one base at the time. Starting out with an empty activated silicon surface (a wafer - like in the production of microprocessors), the synthesis occurs during a series of coupling steps: in each step the four nucleotides are presented to the entire surface one at the time, and will be coupled to the growing oligonucleotides in a tightly controlled manner. The individual positions on the array are targeted for coupling by a light-based de-protection of the oligonucleotides and the use of a series of lithographic masks to shield the rest of the array. Since the coupling-efficiency puts a limit to how long the oligonucelotide can be, chips manufactured with this technique use short probes - usually around 25bp. To compensate for the short probe-length a number of probes (20 for most Affymetrix arrays) target different positions in the target transcript. This type of array offers great reproducibility, but the types of arrays is limited to what the Affymetrix company has decided to produce. Figure 1.4 [panel B] shows an actual Affymetrix GeneChip®.

In recent years a few alternative methods for *in situ* synthesis of micro-arrays that are more flexible have emerged. Using a micro-mirror device in stead of static lithographic masks, the synthesis process can be reprogrammed easily - this approach has been commercialized by NimbleGen (Iceland) (Singh-Gasson et al., 1999; Albert et al., 2003; Stengele et al.,

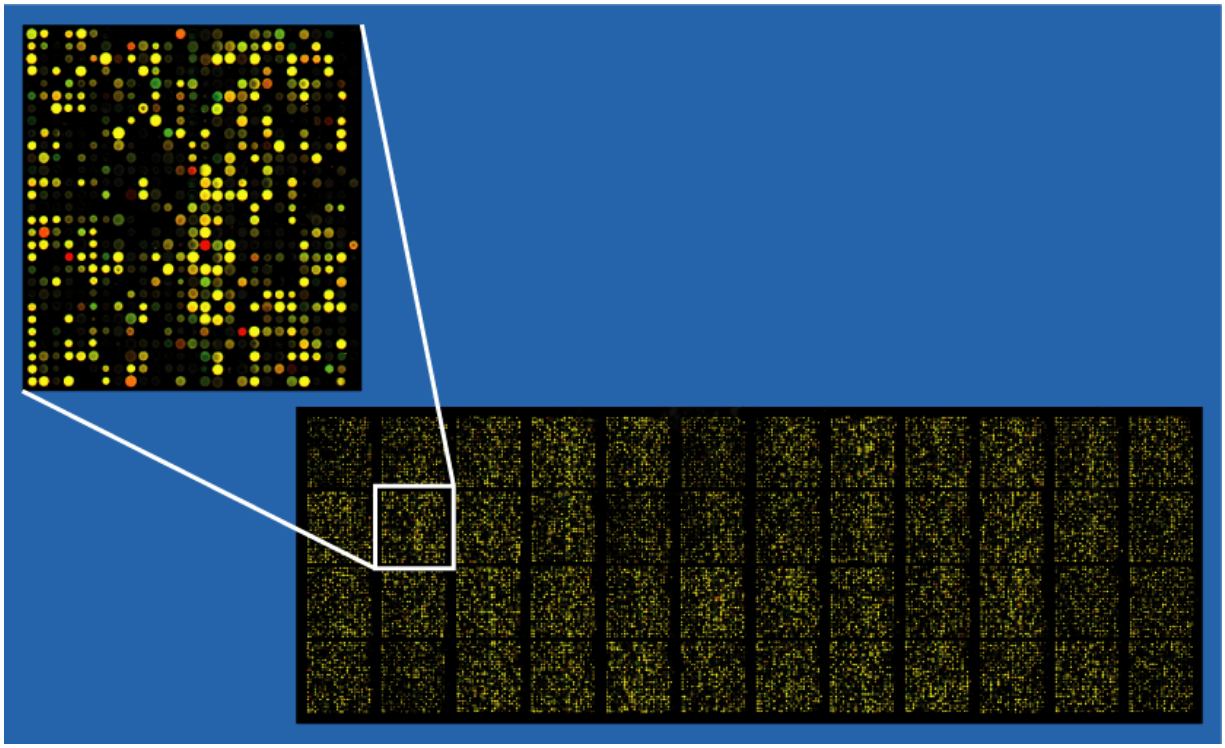

Figure 1.2: **Spotted microarray.** This image shows a pseudo-color image of the combined Cy3/Cy5 channels of a spotted oligo array with 37.500 probes. A sections of the array is shown magnified to highlight the individual spots. Observe how the spots are arranged in a matrix, allowing the signal for each spot to be mapped back to the probe in question. Image source: `www.wikipedia.org`

.

2005) and FEBIT (Germany) (Güimil et al., 2003; Baum et al., 2003). The more flexible techniques makes it feasible to produce custom arrays even for smaller studies. (In 2005 Affymetrix acquired the NimbleGen company and offers the NimbleGen custom arrays under the new name "NimbleExpress"). Custom probe design software suites such as the "*OligoWiz*" application (Wernersson et al., 2007), were developed to exploit the flexibilities of such new technologies.

1.2 Two condition hybridizations

One of the main uses of DNA micro-arrays is to compare the expression of two different conditions (typically "treated" vs "non-treated" or "mutant" vs "wildtype"). Figure 1.5 demonstrates the main steps of this type of hybridization. The idea is here to label the transcripts from one condition with one fluorescent dye (Cy3) and the transcripts from the other condition with a different fluorescent dye (Cy5). Following hybridization the array is scanned at the wave-lengths specific for the individual dyes, thus producing two images which each capture the gene expression under one of the conditions. Following normalization for the total signal in each image (this compensates for differences in amount of RNA used and technical issues), the two images are treated as two color channels (usually red and green) and combined to a single image. Spots of intermediate

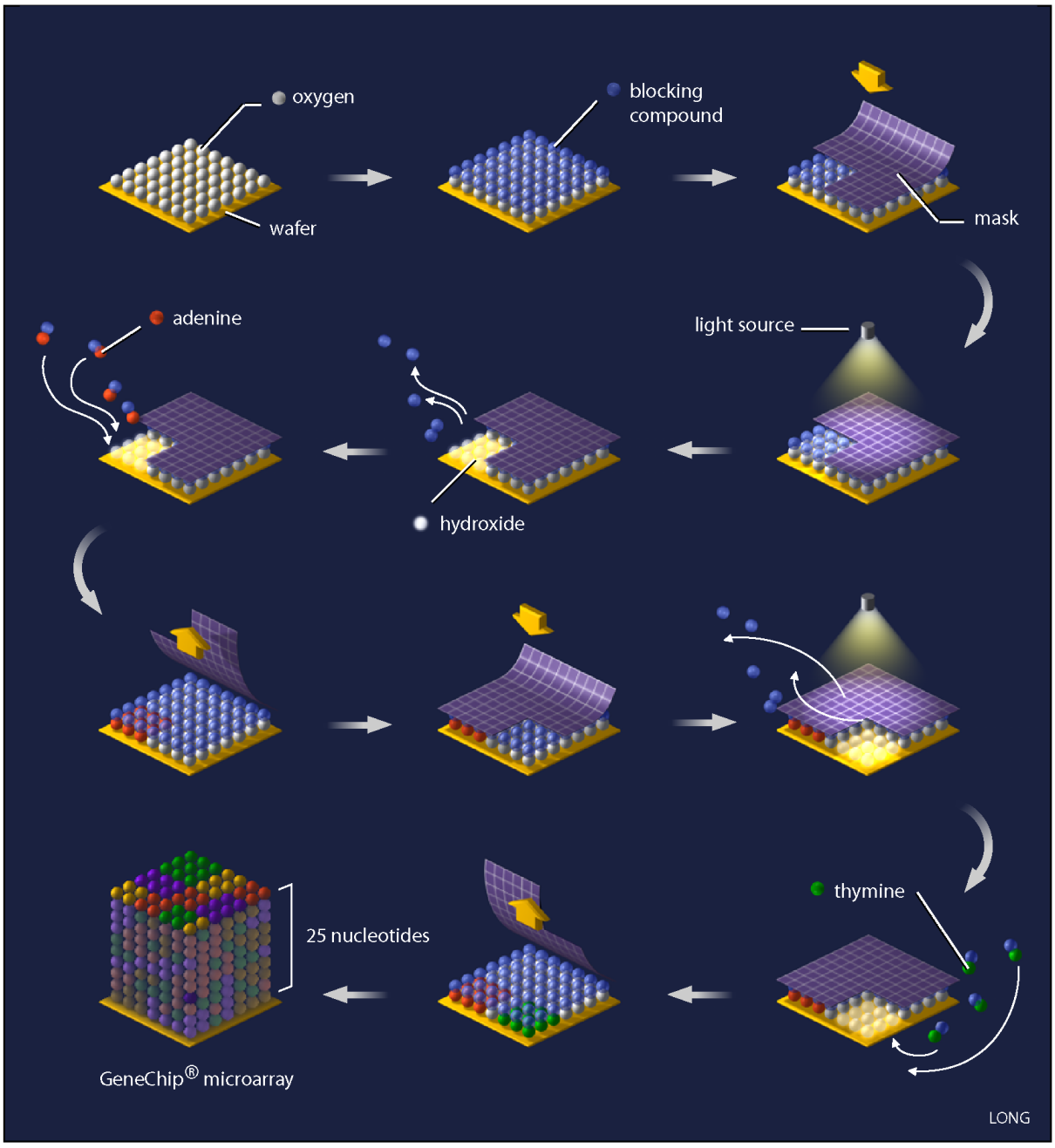

Figure 1.3: **Photolithographic *in-situ* synthesis: Building an Affymetrix GeneChip.** These figures show the steps involved in the production of an Affymetrix GeneChip. Notice how a set of masks are used to direct which area of the array surface that will be targeted during each round of nucleotide coupling. Image source: The Science Creative Quarterly, `http://scq.ubc.ca/`, artist: *Jiang Long*.

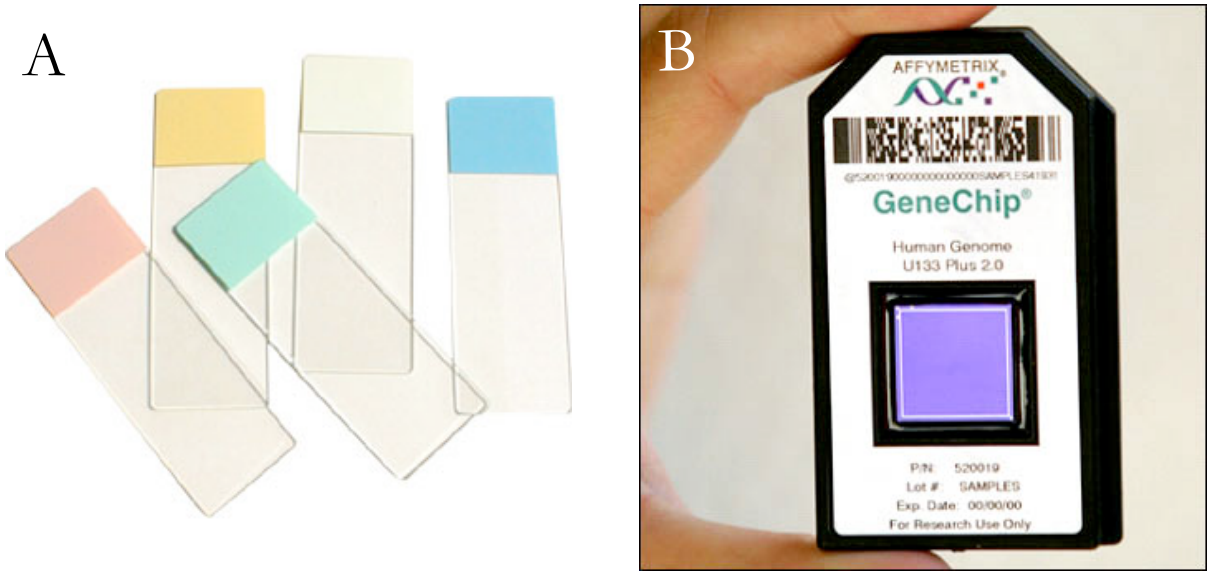

Figure 1.4: **The physical appearance of spotted arrays** *vs* **GeneChips. Panel A:** Glass slides used for spotted arrays. The slides shown here is from Bioslide Technologies - `http://www.bioslide.com/`. **Panel B:** Affymetrix GeneChip®- all affymetrix GeneChips have the same form-factor, the example shown here is the human genome array "U133". Image source: The Affymetrix website, `http://www.affymetrix.com/`

color (yellow in this case) represents transcripts where the expression is of equal magnitude in the two conditions, and red and green as cases of over-expression in one of the conditions.

The two-condition hybridization is especially useful when working with spotted arrays, since the problems with chip-to-chip variation (e.g. small differences in spot size) can be sidestepped here: it is the *relative* expression of a given gene between the two conditions that is compared.

1.3 Single condition hybridization

Following the introduction (and maturation) of *in-situ* synthesized arrays, the problems with chip-to-chip variation due to differences in the manufacturing of the arrays have been greatly reduced. This has allowed for a simpler approach for the comparison of different conditions, by allowing the individual conditions to be probed with individual arrays (see Figure 1.6 for an example).

When working with time-series data such as the cell cycle experiment described in (de Lichtenberg et al., 2005b) single-conditions hybridizations will be the most suited, since the biological problem at hand does not involve comparison of wildtype *vs* mutant, but rather a variation of expression as a function of time. For this purpose it is crucial to be able to compare all the timepoints – for example using computational methods for detecting genes periodically expressed during the cell cycle (de Lichtenberg et al., 2005a).

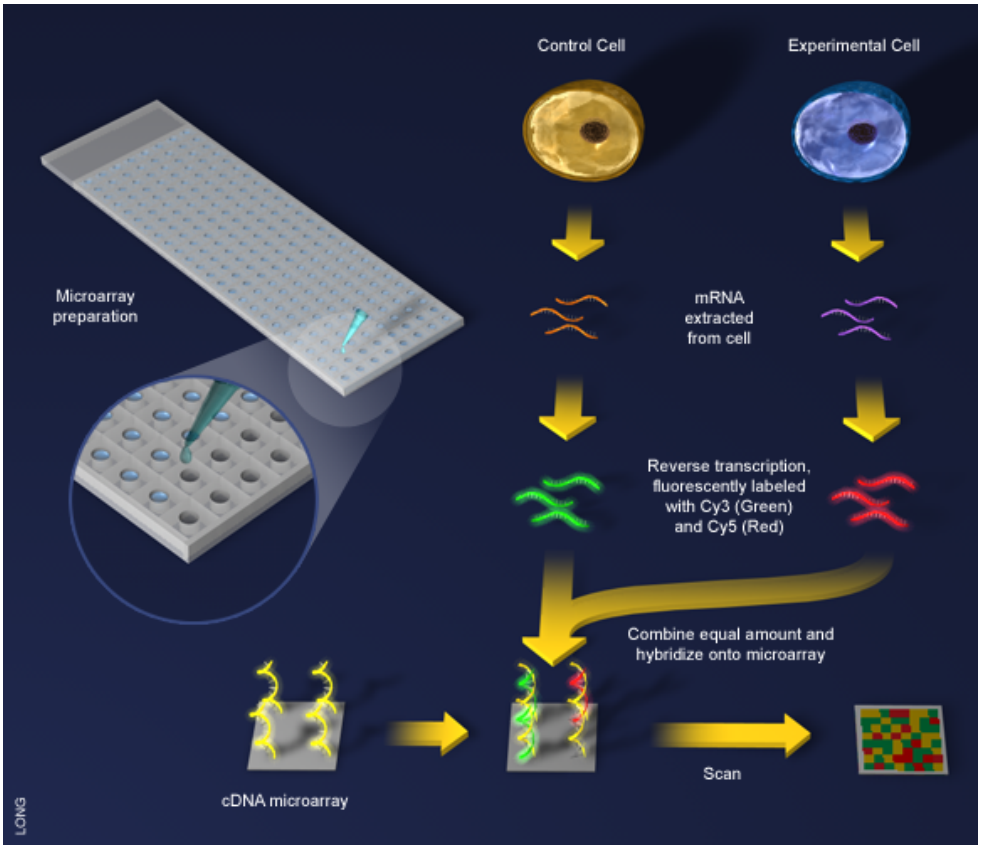

Figure 1.5: **Example of two condition hybridization.** This figure shows the steps involved in a two-condition hybridization using a spotted cDNA micro array. Image source: The Science Creative Quarterly, `http://scq.ubc.ca/`, artist: *Jiang Long*.

1.4 The array data analysis pipeline

The use of DNA micro-arrays can be divided into a number of sub-steps as shown in Figure 1.7, which will be reviewed briefly here.

1.4.1 Experimental design

One of the single most important factors in a successful array experiment is the basic experimental design: Will the experiment in question actually be able the answer the biological question asked, and will the data be sufficient for the statistical analysis planned. For experiments comparing different conditions (e.g. wildtype *vs* mutant) biological replicates will improve the robustness of the analysis, since stochastic variation is not likely to be repeated. When working with time-series data, such as cell cycle data, the individual timepoints will to some degree function as replicates, depending on the number and spacing of the timepoints. If no pilot experiments have been conducted, it may be a good idea to sample more timepoints than strictly needed (e.g. every 10 minutes instead of every 20 minutes), to have the opportunity of adding in more arrays later, if it turns out to improve the data analysis significantly.

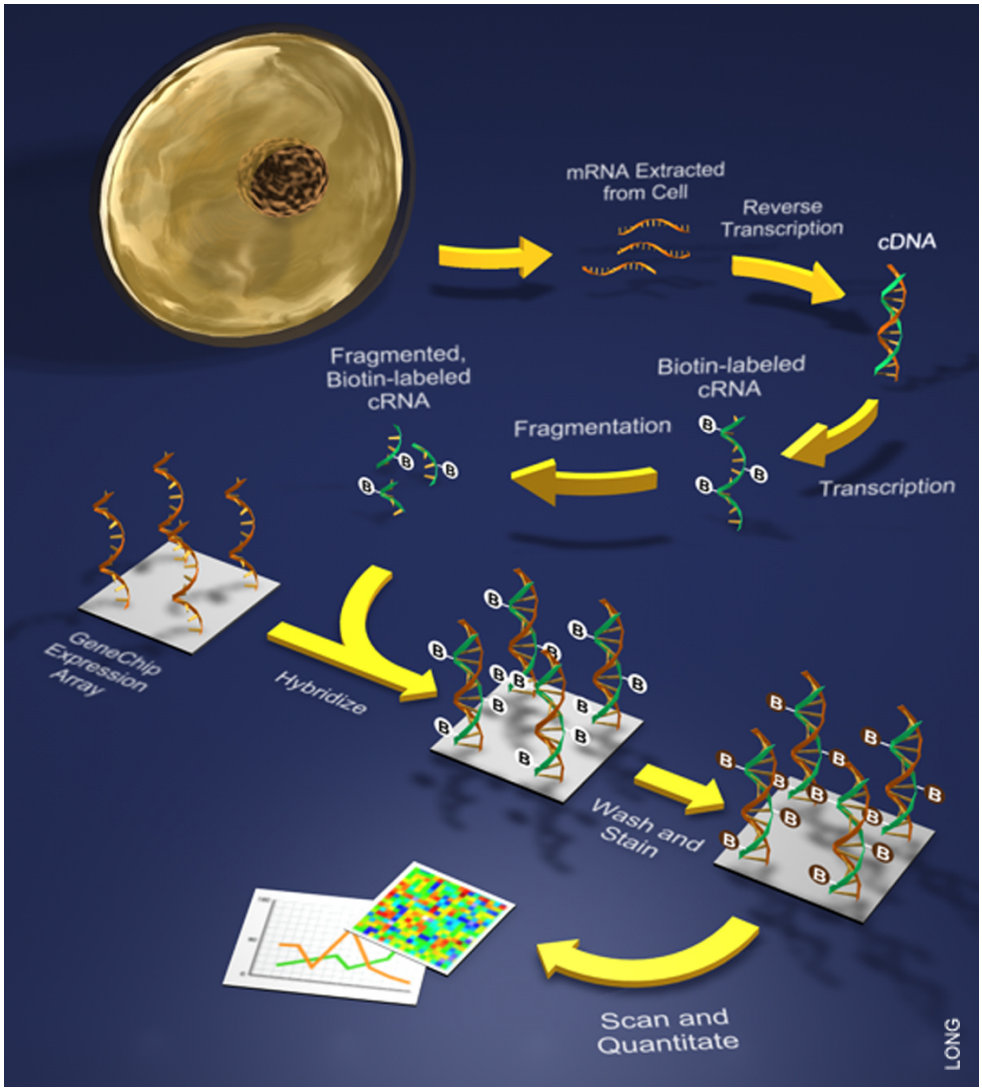

Figure 1.6: **Example of single condition hybridization.** This figures shows a single-condition hybridization using an Affymetrix GeneChip array. Image source: The Science Creative Quarterly, `http://scq.ubc.ca/`, artist: *Jiang Long*.

1.4.2 Array/Probe design

For experiments where the goal is to measure the gene expression of a well known model organism (e.g. Yeast), the cheapest and easiest solution will most often be to use a standard array, like the Affymetrix Yeast Genome 2.0 array, or a commercially available set of probes for spotting. If no suitable standard solution is available or if the purpose of the study is to investigate a special condition (like detection of splicing patterns), a custom array must be designed. For an in-depth review of probe design for DNA micro-arrays, please refer to Wernersson et al. (2007).

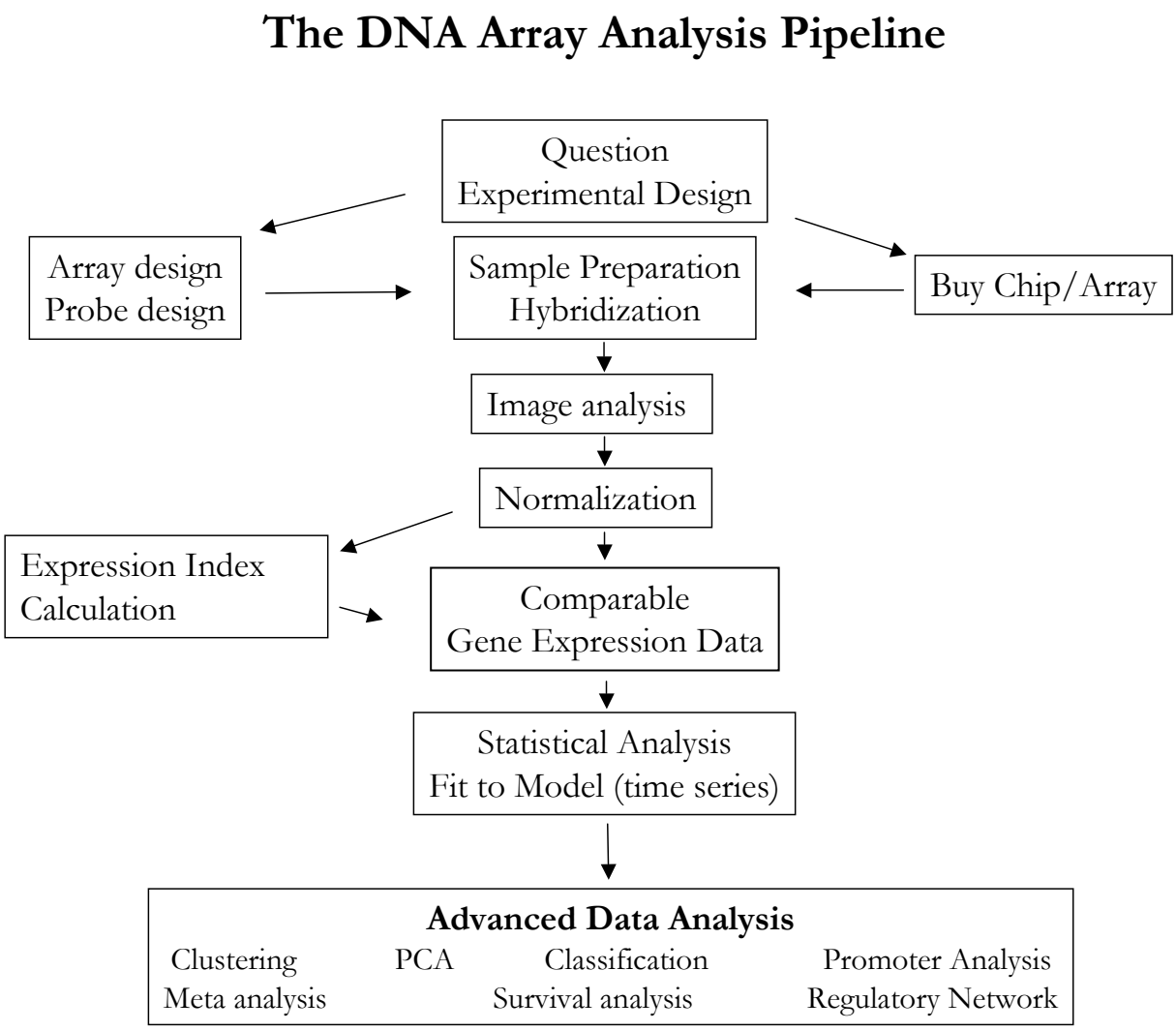

Figure 1.7: **Outline of the entire array analysis pipeline.** See the main text for a detailed explanation. Figure by Henrik Bjørn Nielsen and Agnieszka S. Juncker.

1.4.3 Sample preparation / hybridization

In order to ensure that the mRNA levels measured actually reflect real biology, it is critical to sample the cells / tissue in question *quickly* and to fix the mRNA. For example a way to do this when working with a Yeast culture could be to mix the sample with ice immediately after sampling, followed by pelleting the cells quickly in a cooling centrifuge and freezing the cell pellet in liquid nitrogen ($-196°C$). The idea is here to slow down both transcription and degradation of RNA prior to the relatively long (2-3 minutes) centrifugation step which cannot be avoided. Once the samples have been fixed, they can be kept at $-80°C$ until extraction of total RNA can be performed.

The next step is to extract the total RNA from the samples. This involves both lysing the cells by enzymatical or mechanical means (e.g. grinding the cells at $-196°C$ using liquid nitrogen) and the extraction and purification of the RNA. This could for example be done using a standard protocol such as the "hot phenol" protocol, but also a number of

commercial kits are available. Many of these kits do away with the most toxic chemicals and are often to be preferred. Most importantly, the quality of the total RNA must be assessed after the purification. The purity can be assessed by standard methods such as measuring the A260/A280 ratio on a spectrophotometer. In order to assess if the RNA has been (partly) degraded, the RNA can checked on an electrophoresis gel or analyzed using capillary-based equipment, such as a BioAnalyzer. In both cases the ribosomal RNA bands should be clearly present – a smear indicates degraded RNA.

Following the extraction the RNA is (optionally) amplified and finally labeled. The amplification/labeling protocol varies depending on the platform, but are often based upon the Eberwine protocol (Gelder et al., 1990). Both Affymetrix and Ambion offer commercial kits for the entire process for the Affymetix platform. Notice that it is important to be aware of the *strandness* of both the probes and the targets: both the Eberwine protocol and the standard Affymetrix protocol will revert the strandness of the probe (mRNA = sense → (DNA intermediates) → labeled aRNA/cRNA = anti-sense). In this case the probes must also be sense[1], in order to detect the targets.

Finally the labeled target molecules are fragmented and hybridized to the array. A number of different commercial incubators exist, but the principle is the same: to incubate the array at a fixed temperature, usually over night or up to 16 hours. Following the hybridization the arrays or carefully washed, in order to remove non-bound targets, and the arrays are scanned at a wavelength appropriate for the dye being used. This is the last "wet-lab" step, which will produce the data for the downstream processing.

1.4.4 Image analysis

The image analysis is the process of converting the digital image from the scanning to expression values assigned to the individual probes. The first step is *gridding*, which is the process of determining the location of each individual "feature" on the array (e.g. one spot on spotted arrays). This can especially be a problem for spotted arrays, due to irregularities in the spotting, and the gridding procedure may have to be aided by hand. Furthermore the area that will be used to determine the intensity of each spot must be defined. For Affymetrix / NimbeExpress arrays the gridding step is almost always handled automatically. Secondly, the signal intensity in the area around each feature ("spot") is converted to a single expression value.

1.4.5 Normalization

Array vs. array normalization

Before it is meaningful to compare the signal from two (or more) arrays, it is necessary to make the data comparable. This means removing (by mathematical means) differences in the signal that are due to technical issues (for example slight differences in hybridization

[1] Affymetrix sometimes list this condition as "anti-sense detecting" which only adds to the confusion.

conditions) while retaining the differences that reflect real biological differences. Figure 1.8 visualizes the effect of normalization on array vs. array comparison. Observe that the data-points far off the diagonal are likely to correspond to real biological variation between the two conditions.

A substantial effort has been put into developing good methods for array vs. array normalization during the recent years. Well established algorithms currently in use by the scientific community includes qspline (Workman et al., 2002), LOWESS (Cleveland, 1979), Quantiles (Bolstad et al., 2003) and RMA (Irizarry et al., 2003).

Probe affinity normalization

A different approach to normalization is to make the individual probes comparable *within the array*. The issue is that even probes targeting the same transcript, which in theory should pick up exactly the same signal, express a wide variation of signal intensities. In order to fully make the probes comparable it is important to correct for differences in the *sequence specific* annealing affinity to the target.

Being able to compare the signal from the individual probes is especially important when working with splice-detecting arrays, since differential splicing will cause the signal for *part* of a given transcript to be modified (e.g. the signal from a single exon may completely disappear). Also, when designing probes for a splice-detecing array the placement of the probes will often be restricted to a specific sub-part of the transcript (e.g. an Exon/Intron junction), and there will be very limited possibility to move "bad" probes to a different position. As an example of an approach of solving this problem using a thermodynamic model of annealing, please refer to Bruun et al. (2007).

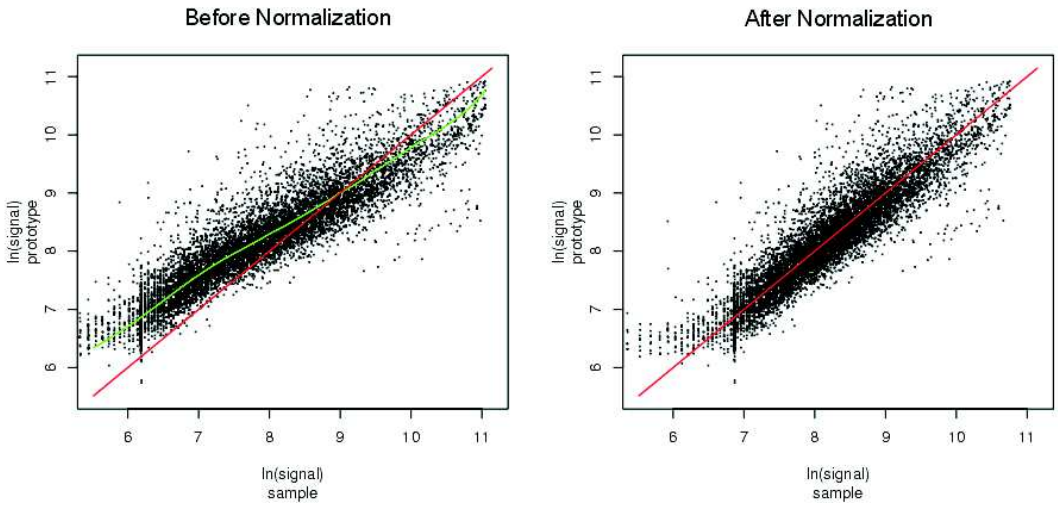

Figure 1.8: **Effect of normalization on array to array comparison.** The plots show how the expression values for each individual transcript compare to each other on two arrays, before and after qspline normalization. Figure by Carsten Friis.

1.4.6 Expression index calculation

If the array uses more than one probe per transcript (e.g. Affymetrix GeneChips), all the individual values are usually collapsed into a single value representing the level of expression for the transcript as a whole (the Gene Expression index). A number of algorithms has been developed for this purpose, from simply calculating the median to more advanced methods (dChip (Li and Wong, 2001), RMA (Irizarry et al., 2003), MAS5 (Hubbell et al., 2002)) which are robust towards outlier values.

1.4.7 Statistical analysis

Depending on the nature of the experiment different kinds of statistical analysis can be applied in order to find the genes that are differentially expressed in-between the experimental conditions investigated (e.g. "Cancer" vs. "normal" tissue). Among the most widely used statistical tools are the T-Test and ANOVA, which can be used to identify the individual genes that are significantly differentially expressed between the sample groups.

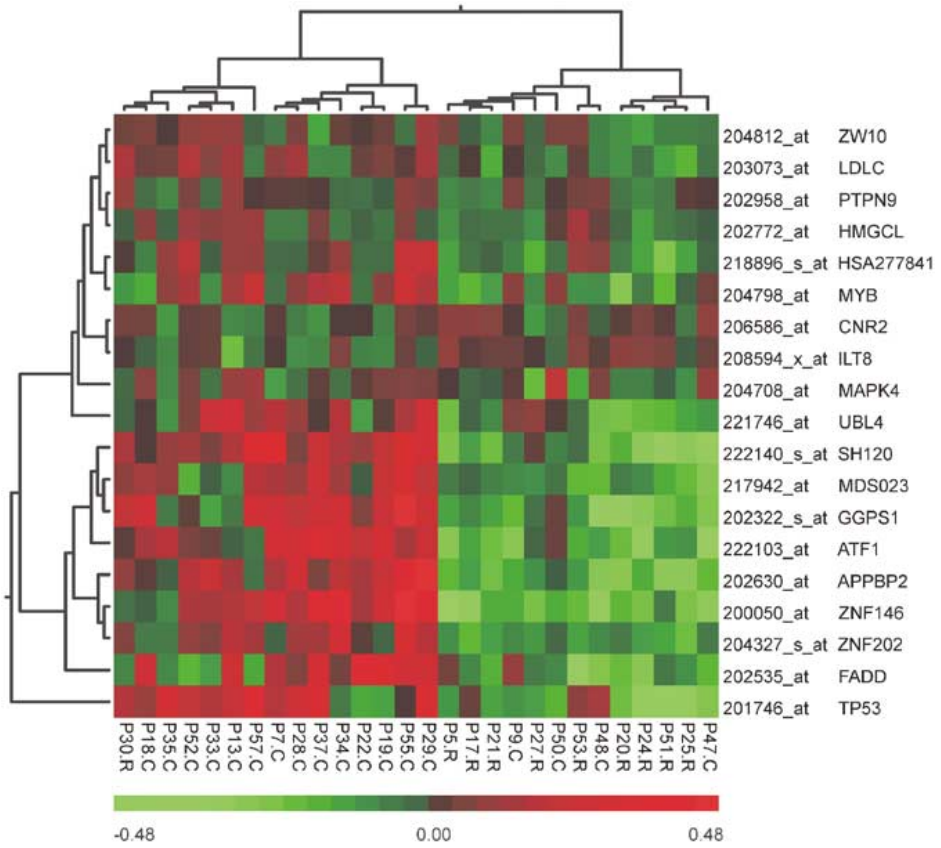

Figure 1.9: **Example of 2D hierarchical clustering.** This example shows a heatmap where leukemia patients are clustered according to their gene expression profile for the 28 most significant genes. Figure by Hanni Willenbrock and Agnieszka S. Juncker.

If a general expression pattern (that transfers between studies) can be found that significantly distinguishes between the conditions, this can potentially be used diagnostically. For this classification approaches have been widely used, and feature selection is an important issue, where for example PCA analysis has proven to be useful. Once a list of genes distinguishing between groups of interest has been generated, the expression profiles for these genes can be visualized. Figure 1.9 shows an example of this - here a number of leukemia patients which belong to two diagnosis groups ("severe" and "mild") are analyzed together (from Willenbrock et al. (2004)).

Due to the large number of genes and the fact that many genes are part of pathways or regulatory modules, more advanced methods that aim at elucidating the modularity within the data have been developed. One of the early approaches to finding modules from expression data has been devised by Segal et al. (2003). Other examples are the widely used divisive clustering methods such as PAM. A further approach is analyzing the Gene Ontology (Ashburner et al., 2000) categories of the gene groups. Such an analysis could reveal that certain GO-terms (e.g. a biological pathway or a sub-cellular localization) is over-represented in the list of differentially expressed genes or modules/clusters, compared to the background distribution.

A different type of analysis is the analysis of time-series data. Here a number of arrays representing different time-points throughout an experiment are compared. A good example of this approach is analysis of the cell cycle data reviewed in Gauthier et al. (2008) - see Figure 1.10 for a graphical overview. It should be noticed that in this case it is the *gene expression profile* (see Figure 1.10) throughout the experiment that is of importance rather than an analysis of what is up or down regulated.

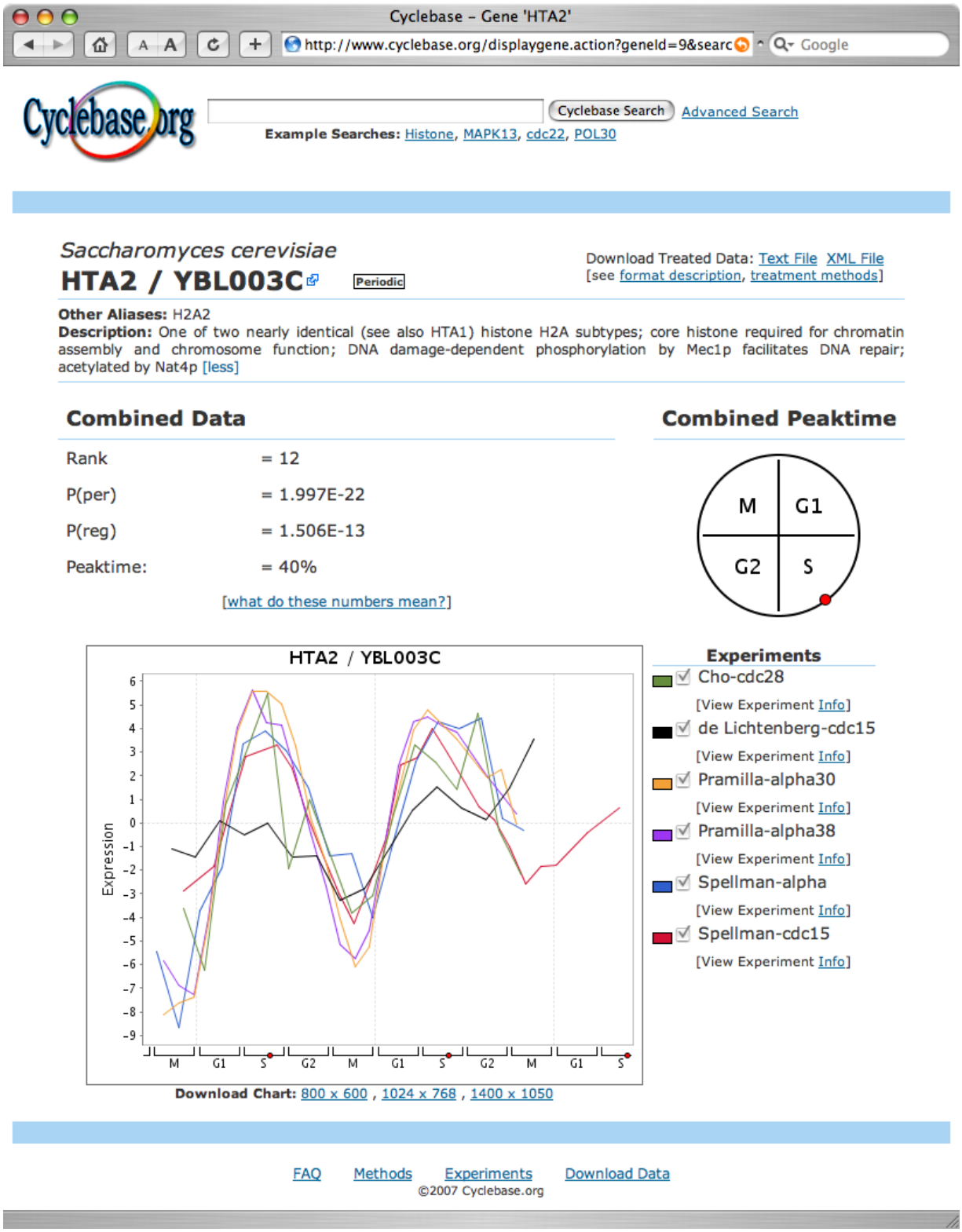

Figure 1.10: **Visualizing time-series data.** Screen shot from the `cyclebase.org` webserver (**Paper VII**) showing expression profiles for the Yeast gene "HTA2" for six independent cell cycle experiments.

Bibliography

Albert, T. J., Norton, J., Ott, M., Richmond, T., Nuwaysir, K., Nuwaysir, E. F., Stengele, K.-P., and Green, R. D. (2003). Light-directed 5'–¿3' synthesis of complex oligonucleotide microarrays. *Nucleic Acids Res*, 31(7):e35.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29.

Baum, M., Bielau, S., Rittner, N., Schmid, K., Eggelbusch, K., Dahms, M., Schlauersbach, A., Tahedl, H., Beier, M., Güimil, R., Scheffler, M., Hermann, C., Funk, J.-M., Wixmerten, A., Rebscher, H., Hönig, M., Andreae, C., Büchner, D., Moschel, E., Glathe, A., Jäger, E., Thom, M., Greil, A., Bestvater, F., Obermeier, F., Burgmaier, J., Thome, K., Weichert, S., Hein, S., Binnewies, T., Foitzik, V., Müller, M., Stähler, C. F., and Stähler, P. F. (2003). Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res*, 31(23):e151.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Bruun, G. M., Wernersson, R., Juncker, A. S., Willenbrock, H., and Nielsen, H. B. (2007). Improving comparability between microarray probe signals by thermodynamic intensity correction. *Nucleic Acids Res*, 35(7):e48.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74:829–836.

de Lichtenberg, U., Jensen, L. J., Fausbøll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005a). Comparison of computational methods for the identification of cell cycle regulated genes. *Bioinformatics*, 21(7):1164–1171. doi:10.1093/bioinformatics/bti093.

de Lichtenberg, U., Wernersson, R., Jensen, T. S., Nielsen, H. B., Fausbøll, A., Schmidt, P., Hansen, F. B., Knudsen, S., and Brunak, S. (2005b). New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, 22(15):1191–1201.

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773.

Gauthier, N. P., Larsen, M. E., Wernersson, R., de Lichtenberg, U., Jensen, L. J., Brunak, S., and Jensen, T. S. (2008). Cyclebase.org–a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res*, 36(Database issue):D854–D859.

Gelder, R. N. V., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D., and Eberwine, J. H. (1990). Amplified rna synthesized from limited quantities of heterogeneous cdna. *Proc Natl Acad Sci U S A*, 87(5):1663–1667.

Güimil, R., Beier, M., Scheffler, M., Rebscher, H., Funk, J., Wixmerten, A., Baum, M., Hermann, C., Tahedl, H., Moschel, E., Obermeier, F., Sommer, I., Büchner, D., Viehweger, R., Burgmaier, J., Stähler, C. F., Müller, M., and Stähler, P. F. (2003). Geniom technology–the benchtop array facility. *Nucleosides Nucleotides Nucleic Acids*, 22(5-8):1721–1723.

Hubbell, E., Liu, W.-M., and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31(4):e15.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–36.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680.

Maniatis, T., Fritsch, E., and Sambrook, J. (1982). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbour Laboratory Press.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Segal, E., Yelensky, R., and Koller, D. (2003). Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19 Suppl 1:i273–i282.

Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol*, 17(10):974–978.

Stengele, K.-P., Bühler, J., Bühler, S., Kvassiouk, E., Green, R., Prykota, T., and Pfleiderer, W. (2005). Recent highlights on photolithic oligonucleotide array in situ synthesis. *Nucleosides Nucleotides Nucleic Acids*, 24(5-7):891–896.

Wernersson, R., Juncker, A. S., and Nielsen, H. B. (2007). Probe selection for DNA microarrays using OligoWiz. *Nat Protoc*, 2(11):2677–2691.

Willenbrock, H., Juncker, A. S., Schmiegelow, K., Knudsen, S., and Ryder, L. P. (2004). Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using dna microarrays. *Leukemia*, 18(7):1270–1277.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielsen, H. B., Saxild, H.-H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol*, 3(9):research0048.