KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

STATISTICAL ANALYSIS OF MICROARRAY DATA: APPLICATIONS IN PLATFORM COMPARISON, COMPENDIUM DATA, AND ARRAY CGH

Promotoren:
Prof. dr. ir. Y. Moreau
Prof. dr. ir. B. De Moor

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Joke ALLEMEERSCH

December 2006

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

STATISTICAL ANALYSIS OF MICROARRAY DATA: APPLICATIONS IN PLATFORM COMPARISON, COMPENDIUM DATA, AND ARRAY CGH

Jury:
Prof. dr. ir. H. Neuckermans, voorzitter
Prof. dr. ir. Y. Moreau, promotor
Prof. dr. ir. B. De Moor, co-promotor
Prof. dr. F. Holstege
Prof. dr. M. Kuiper
Prof. P. Rouzé
Prof. dr. ir. E. Schrevens
Prof. dr. ir. J. Vandewalle
Prof. dr. ir. J. Vermeesch

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Joke ALLEMEERSCH

U.D.C. 681.3*J3, 519.23 December 2006

D/2006/7515/97

ISBN 978-90-5682-763-2

Voorwoord

Traditioneel begint elk doctoraat terecht met een woord van dank en ook hier mag dat niet ontbreken.

Wanneer Prof. Yves Moreau me vroeg om te komen werken in de Bioinformatica groep op de data-analyse van microarrays, ben ik daar op ingegaan zonder goed te weten wat Bioinformatica was, laat staan wat een microarray was. Misschien gelukkig maar... Nu, een viertal jaar later, ben ik blij dat ik toen "ja" gezegd heb. Yves, bedankt om me in deze wereld binnen te loodsen. Je enthousiasme en gedrevenheid werkten aanstekelijk. Ook een woord van dank voor mijn co-promotor, Prof. Bart De Moor voor de kansen, de steun en de middelen die nodig waren om aan dit doctoraat te werken.

I also thank the members of the jury and in particular the reading committee, Prof. Pierre Rouzé, Prof. Eddie Schrevens, and Prof. Joos Vandewalle, for their reading efforts and their valuable remarks.

The CAGE project was a good opportunity for nice collaborations. First of all, I have to acknowledge Prof. Pierre Hilson and Prof. Martin Kuiper for the close collaboration, while working on the benchmark of the CATMA array, which was a fruitful experience for me. Naming all persons involved in the CAGE project would result in a long list, therefore I will only mention some of them. Thank you Helen Parkinson, for struggling together with the MAGE-ML files, Wolfram Brenner for the pleasant collaboration during your stay here in Leuven, Gert Sclep for all the confusing emails, Dawn Little, Tom Bogaert, Paul Van Hummelen, and

last but not least Steffen Durinck, my colleague at ESAT. You described us in your own PhD text as a “great team”. And, although you were running away all the time, I must admit that there is some truth in it. I really enjoyed our close collaboration and I hope that we can stay in touch and, who knows, perhaps our paths will cross again.

Gedurende het laatste jaar, werkte ik ook nauw samen met Femke Hannes en Prof. Joris Vermeesch. Allebei hartelijk bedankt voor de aangename samenwerking.

Veel dank gaat uit naar de ganse Bioinformatica groep — Cynthia, Karen, Steven, Stein, Bert, Gert en de ganse groep — voor de dagelijkse werksfeer: bedankt! Ook van harte bedankt aan de mensen die instaan voor de praktische regelingen en dan in het bijzonder Ida, Bart en Ilse.

Onrechtstreeks, maar zeker zo belangrijk voor dit werk is de invloed van mijn achterban, waar ik steeds terecht kon voor de noodzakelijke ontspanning en afleiding. Christoph, Kirsten, Mario, Aurore, Jozefien, Nathalie, iedereen in de Kon. St. Martinusharmonie in Tessenderlo, waarbij ik al een 15-tal jaar elke vrijdagavond een muzikale uitlaatklep vind, het gezelschap van in de Duitse lessen, de cello les,... Allemaal ne dikke merci!

Jeroen en Sofie, ik denk dat we ons de “bende van de Lombaarden” mogen noemen. Dankzij jullie is het bij ons een gezellige boel, waar het altijd leuk is om ’s avonds thuis te komen. Hopelijk kunnen we nog een paar jaartjes in elkaars gezelschap wonen!

Papa, mama, Simon & Veerle, dankzij jullie kan ik met een goed gevoel zeggen dat ik uit een heel warm nest kom. Jullie stonden altijd voor mij klaar. Niets was te veel voor jullie.

Steven, jongen, mercikes voor al je steun. Dankzij jou heb ik in alle rust dit doctoraat kunnen afwerken. We gaan er samen nog iets moois van maken...

Joke

Abstract

Since the mid-1990s microarrays have become a well-established technology to measure when, where, and to what extent a gene is expressed, on a genomewide scale. Whereas early experiments included only a few hybridizations, the tendency is now growing towards large compendium projects producing hundreds of samples and providing information on the gene expression at different developmental stages, under different environmental conditions or of different mutants.
This Ph.D. has been carried out within the framework of the CAGE project, a European demonstration project that aimed at composing an atlas of gene expression of *Arabidopsis thaliana* throughout its life cycle and under a variety of stress conditions. As a microarray platform, the *Complete Transcriptome MicroArray* or CATMA array was developed within this project. Its utility had not yet been proven and, therefore, a dedicated experiment was set up to benchmark the CATMA array against two well-established platforms. Different aspects of the platforms are compared in this thesis and the results for the CATMA array are promising.
The main contribution of this Ph.D. to the CAGE project is the preprocessing of the microarray data. Within the CAGE project, microarray data exchange and storage was done in MAGE-ML format, so that CAGE is a well-annotated, MIAME-compliant compendium. To facilitate data exchange in MAGE-ML format, a software package `RMAGEML` is presented, enabling to import MAGE-ML files in the statistical environment R and to update the MAGE-ML files with preprocessed values. The package is now part of the Bioconductor package, which is a major open source tool for the statistical analysis of microarray data.

Using this RMAGEML package, a data preprocessing pipeline is developed for an automated preprocessing and quality assessment of the CAGE data. With this high-throughput data preprocessing pipeline, a sizable set of more than 2,000 hybridizations is preprocessed, ready for an in-depth analysis.
To conclude, a nice example shows the improvement a well-chosen experimental design can bring. ArrayCGH is a microarray technology that can be used to detect aberrations in the ploidy of DNA segments in the genome of patients with congenital anomalies. In this Ph.D., I present a tool to analyze arrayCGH loop designs in which three patients are placed in a loop design, which is advantageous over the classical dye-swap approach.

Korte inhoud

Gedurende de laatste 10 jaar zijn microroosters geëvolueerd tot een gevestigde techniek voor het meten van gen expressie, voor duizenden genen in parallel. Oorspronkelijk werden microrooster experimenten opgezet als kleinschalige experimenten, bestaande uit een gelimiteerd aantal hybridizaties. Tegenwoordig verschuift de aanpak van microrooster experimenten meer en meer naar grootschalige compendium experimenten, waarbij honderden stalen gehybridiseerd worden en die informatie verschaffen over gen expressie in verschillende ontwikkelingsstadia, onder bepaalde omgevingsinvloeden, of van verschillende mutanten.

Dit doctoraat kadert binnen het CAGE project, een Europees demonstratie project, dat streefde naar het samenstellen van een atlas van gen expressie van de plant *Arabidopsis thaliana* gedurende zijn groei cyclus en onder een reeks van stress condities. Als microrooster platform werd gekozen voor de *Complete Transcriptome MicroArray* of CATMA microrooster. De capaciteiten van dit platform waren nog niet bewezen en daarom werd er een experiment opgezet dat toeliet om de CATMA array te toetsen aan twee gevestigde microrooster platformen. Verschillende aspecten van de platformen werden vergeleken in dt werk en de resultaten van de CATMA array bleken veelbelovend te zijn.
De voornaamste bijdrage van dit doctoraat aan het CAGE project was het normaliseren van de microrooster data. Binnen het CAGE project, werd voor de data uitwisseling en bewaring geopteerd voor het MAGE-ML formaat. Dit garandeert de goede annotatie van de CAGE experimenten, zodat het project voldoet aan de MIAME vereisten. Om de data uitwisseling in

MAGE-ML formaat te vereenvoudigen, hebben we een software pakket `RMAGEML` gebouwd, dat toelaat om data opgeslagen in MAGE-ML formaat te importeren in de statistische omgeving van R en om verwerkte data toe te voegen aan de oorspronkelijke MAGE-ML file. Dit pakket maakt nu deel uit van de Bioconductor pakketten.
Dit `RMAGEML` pakket is een belangrijk onderdeel van een data verwerkings pijplijn, die hier ontwikkeld werd voor een automatische normalisatie en kwaliteitscontrole van de CAGE data. Met deze data verwerkings pijplijn, werd een aanzienlijke data set van meer dan 2.000 hybridizaties klaargemaakt voor verdere analyse.
Ten slotte tonen we in een mooi voorbeeld de verbetering die een goed gekozen experimenteel design kan brengen. Array CGH is een microrooster, ontwikkeld voor het detecteren van chromosomale afwijkingen. Dit doctoraat toont een analyse tool voor de analyse van array CGH loop designs, waarbij drie patiënten in een loop design vergeleken worden, wat een significante verbetering is ten opzichte van de klassieke twee-aan-twee vergelijkingen.

Contents

Nederlandse samenvatting

Statistische verwerking van microrooster data: toepassingen in platform vergelijkingen, compendium data en array CGH

Hoofdstuk 1: Inleiding

Het einde van grote sequentie analyse projecten betekende het begin van het ontcijferen van de informatie verborgen in het DNA. In het DNA liggen de *genen*, die alle erfelijke informatie bepalen, verborgen. Eén belangrijke stap is het lokaliseren van de genen in de DNA sequentie. De volgende uitdaging bestaat dan uit het toewijzen van functies aan elk van deze genen.
Een belangrijk hulpmiddel hierbij zijn de *DNA microroosters* (*microarrays*), die ons in staat stellen om te meten waar, wanneer en in welke mate een gen actief is, en dit voor duizenden genen in parallel. Een microrooster experiment brengt een massa data voort en de analyse van deze microrooster data is bijna een statistische discipline op zich geworden. In deze thesis worden een aantal aspecten van microrooster data analyse belicht. De thesis is als volgt opgebouwd.

Hoofdstuk 2 en 3 geven een gedetailleerde inleiding op het onderwerp. In Hoofdstuk 2 worden de, in deze context, belangrijke begrippen uit de cel biologie geïntroduceerd. Via het centrale dogma, leiden we het begrip gen en, meer specifiek, genexpressie in. Het idee achter microroosters en de analyse van microrooster experimenten wordt ook voorgesteld. Hoofdstuk 3 wordt dan direct een stuk technischer. Er bestaan verschillende microrooster platformen met elk hun eigen karakteristieken. We stellen twee groepen microroosters voor, de microrooster met twee kanalen en met een enkel kanaal. Beide microrooster types hebben hun eigen normalisatie vereisten, die we kort bespreken.

In Hoofdstuk 4 maken we de vergelijking tussen drie microrooster platformen, die gebruikt worden voor de plant *Arabidopsis thaliana*. Een experiment speciaal opgezet voor deze vergelijking laat ons toe om de verschillende aspecten van de platformen te vergelijken.

Vervolgens beschrijven we in Hoofdstuk 5 een compendium project, namelijk het *Compendium of Arabidopsis Gene Expression* of CAGE project. Dit is een Europees project met als doel het opbouwen van een genexpressie compendium van de plant *Arabidopsis thaliana* in de verschillende groei stadia en voor een reeks stress condities. Uiteindelijke doel was de productie van een 2.000 biologische stalen, telkens twee maal gehybridiseerd op, in totaal, 4.000 microroosters. De stalen worden geproduceerd in acht laboratoria in Europa. De bioinformatica groep in ESAT was verantwoordelijk voor het normaliseren van deze data.
Om deze data, en ook alle andere gepubliceerde data, op een betekenisvolle manier publiek te maken, is het belangrijk dat de data voldoende geannoteerd is. Daarom werd de data binnen het CAGE project opgeslagen in MAGE-ML formaat, een XML taal, ontwikkeld door de Microarray Gene Expression Database groep als een standaard voor het beschrijven van microrooster experimenten. Het CAGE project was één van de eerste projecten, die het MAGE-ML formaat actief gebruikten (i.e., niet enkel om data op te slaan, maar ook voor de data communicatie), en er bestond nog geen software om data opgeslagen in MAGE-ML formaat te importeren in statistische dataverwerkingsprogramma's. In dit werk wordt er een stukje software gepresenteerd, die het mogelijk maakt om data te

extraheren uit MAGE-ML formaat en om verwerkte data toe te voegen aan de MAGE-ML files.
Voor het verwerken van de data van 4.000 hybridisaties hebben we een automatische dataverwerkingspijplijn ontwikkeld. Deze pijplijn start van de data, zoals opgeslagen in MAGE-ML formaat, extraheert de nodige informatie voor de normalisatie van de data en voegt de genormaliseerde data toe aan de oorspronkelijke MAGE-ML file. Ondertussen worden de nodige figuren en statistieken gegenereerd voor kwaliteitscontrole van de verschillende hybridisaties.

Een andere toepassing van de microrooster technologie is *array CGH*, een microrooster platform voor de detectie van chromosomale afwijkingen. Detectie van deze afwijkingen geeft inzicht in het ontstaan en de ontwikkeling van de gerelateerde pathologieën. In Hoofdstuk 6 stellen we een data analyse tool voor die loop design experimenten verwerkt. Dit design heeft voordelen ten opzichte van de klassieke paarsgewijze vergelijking van patiënten, waarbij een test patiënt vergeleken wordt met een normale, referentie patiënt. Twee methodes voor de analyse van dit loop design worden voorgesteld en met elkaar vergeleken. De te verkiezen methode is geïmplemeteerd in een webapplicatie.

Alle analyses, uitgevoerd in dit werk, werden met de open-source, statistische software *R* (`http://www.r-project.org`) uitgevoerd. *R* is een lopend project, waaraan iedereen code kan toevoegen. Een succesvol voorbeeld hiervan is Bioconductor (`http://www.bioconductor.org`), een reeks pakketten voor de analyse van genomische data. Bioconductor is een populaire tool geworden voor de analyse van microrooster data.

Hoofdstuk 2: Microrooster technologie voor het meten van genexpressie

Genen zijn stukjes DNA die de nodige informatie bevatten om eiwitten te synthetiseren. Het *centrale dogma* beschrijft dit proces in twee stappen: *transcriptie* (het omzetten van DNA in mRNA) en *translatie*, waarbij het mRNA vertaald wordt in een eiwit. Bijgevolg, bepalen de hoeveelheid en

het type DNA dat gekopieerd wordt in RNA, welke eiwitten aangemaakt worden. En dit is exact wat microroosters zullen meten. Indien het DNA van een gen gekopieerd is als RNA, dan is dit gen tot expressie gekomen en microroosters meten *genexpressie*.
Microroosters plaatsen duizenden cDNAs of oligonucleotides (de *proben*) op een plaatje en laten hierop een te analyseren staal over lopen. Doordat complementaire DNA strengen specifiek met elkaar binden (i.e., *hybridiseren*), kan men de genexpressie status voor alle (duizenden) genen, vertegenwoordigd met een probe op de microrooster, meten.
Het gelijktijdig meten van deze duizenden genen leidt tot een massa gegevens. De analyse van deze data gebeurt in een aantal stappen. De data wordt eerst genormaliseerd, zodat verschillen in de data van niet-biologische aard weggewerkt worden. Vervolgens worden lijsten samengesteld met genen die een verschil in expressieniveau vertonen tussen twee of meerdere stalen. Voor deze genen kan men dan de expressieprofielen weergeven en met elkaar vergelijken (bijvoorbeeld, door middel van clustering). Op deze manier kan men genen identificeren, die een rol spelen in bepaalde processen en eventueel een functie associëren.

Hoofdstuk 3: Technische aspecten van DNA microrooster technologie

Er bestaan verschillende microrooster platformen. Ze kunnen op de eerste plaats opgesplitst worden als microroosters met één of twee kanalen. Een verdere opsplitsing gebeurt dan op basis van de gebruikte proben.
Bij de twee kanalen microroosters, hybridiseren twee stalen, gelabeld met twee verschillende kleuren (zie Figuur 3.1, pagina 36). Het best gekend, in deze categorie, zijn waarschijnlijk de cDNA microroosters, waarbij PCR geamplificeerde cDNAs gespot worden op de microroosters.
Een beter alternatief zijn de *long oligonucleotide* microroosters. Gebaseerd op de sequentie informatie alleen, kan men oligonucleotides ontwerpen die specifieker voor het gen zijn dan de complete cDNAs, zodat cross-hybridisatie vermeden kan worden, en tegelijkertijd voldoende lang zijn om enkel te binden met het desbetreffende gen. Deze oligonucleotides worden *in silico* gesynthetiseerd en op de array geprint of *in situ* gesyn-

thetiseerd. Op deze manier maakt men ook geen gebruik van PCR producten, wat cross-contaminatie vermijdt, maar de techniek is vrij duur. Een derde platform dat we vermelden is de *Complete Arabidopsis Transcriptome MicroArray* of CATMA, een microrooster ontwikkeld voor de studie van genexpressie in de plant *Arabidopsis thaliana*. Deze plant wordt vaak als model organisme gebruikt omwille van de korte levenscyclus. Deze CATMA array is het resultaat van een Europees project dat *Gene-specific Sequence Tags* of GSTs voor alle gekende en voorspelde genen wilden samenstellen. De GSTs hebben een lengte van 150 à 500bp en minder dan 70% overeenkomst met eender welke sequentie in het *Arabidopsis* genoom. Later werd deze set nog uitgebreid naar minder specifieke GSTs, zodat ook genen, behorend tot een genfamilie opgenomen werden. Met deze GSTs werd de CATMA array gespot. Voor deze GSTs werden ook PCR primers ontworpen zodat cross-contaminatie vermeden wordt. Dit werd gedaan door aan de $5'$ uiteindes van de oligonucleotide specifieke primers, een paar primers te hechten, gebaseerd op de coördinaten van de GST in de 384 well plate, namelijk een combinatie van 16 primers $(r1, \ldots, r16)$ en 24 primers $(c1, \ldots, c24)$.
Voor de normalisatie van deze data voeren we een achtergrond correctie uit en fitten we een Loess regressie lijn door de MA-plot, per print tip. De log-ratios worden dan nog uitgemiddeld over de *dye-swap* (i.e., hybridisatie waarbij de staal-kanaal combinatie omgewisseld is).

Bij de microroosters met één kanaal (zie Figuur 3.7, pagina 50), bespreken we de short oligonucleotide microroosters van *Affymetrix*. Op een Affymetrix chip wordt een gen niet meer gemeten door één DNA streng, maar door een *probe set*, bestaande uit 11 à 20 probe paren. Elk *probe paar* bestaat uit een oligonucleotide van lengte 25bp (*perfect match*) en een tweede, gelijkaardige oligonucleotide, waarbij enkel de middelste nucleotide veranderd is (*mismatch*). De metingen van de probe sets worden door Affymetrix standaard gecombineerd in één expressiewaarde met *MicroArray Suite 5.0* (MAS 5.0). In Bioconductor bestaat er ook een alternatief, die de mismatch waardes niet in rekening brengt, namelijk *Robust Multi-array Average* (RMA).

Hoofdstuk 4: Benchmark van de CATMA array

Door het zorgvuldig opzetten van de GST collectie had men hoge verwachtingen over de kwaliteit van de CATMA array. Dit hoofdstuk is gewijd aan de vergelijking van de CATMA array met twee commerciële platformen, Agilent en Affymetrix. In de studie nemen we het standpunt van de typische microrooster gebruiker in: we zoeken een gevestigde service provider (VIB-MAF, SeviceXS en NASC voor CATMA, Agilent en Affymetrix, respectievelijk) en vertrouwen op de gangbare analyse methodes.

Het experiment werd als volgt opgezet. Aan een zelfde staal RNA werden zeven paren spike RNAs met gekende concentratie toegevoegd. Deze concentratie bedroeg 10.000 cpc[1] voor spike paar 1 en verminderde telkens met factor 10 tot 0,1 cpc voor spike paar 6 en 0 voor spike paar 7. Dit is spike mix 1. Spike mix 2 wordt dan gemaakt door spike 2 tot 7 een concentratie te geven van 10.000 gaande tot 0,1 en spike 1 een concentratie van 0. Analoog worden er zo 7 spike mixen gemaakt, tot elke spike in elke concentratie gemeten wordt (zie Tabel 4.3, pagina 79). Een referentie spike mix bevat elke spike met een concentratie van 100 cpc.

Voor de microroorsters met twee kanalen worden de zeven spike mixen gehybridiseerd ten opzichte van de referentie spike mix, met een dye-swap, wat het totaal op 14 hybridisaties brengt. Voor het Affymetrix platform worden de zeven spike mixen gehybridiseerd en de referentie spike mix wordt op een aparte, achtste slide gehybridiseerd (zie Figuur 4.2, pagina 78). De data op de CATMA array werden genormaliseerd met achtergrond correctie en een Loess regressie per print tip. Voor Agilent en Affymetrix gebruikten we de bijgeleverde expressie waardes. Op Affymetrix pasten we naast MAS 5.0 ook nog RMA toe.

Dit design stelt ons in staat om verschillende aspecten van de platformen te bekijken, zoals het bereik van de metingen, zoals getoond in Figuur 4.3 op pagina 81. Hierop scoorde de CATMA array goed — qua sensitiviteit kregen alle platformen problemen tussen de 10 en 1 cpc, maar voor de hoge intensiteiten had CATMA geen enkel probleem, terwijl zowel Affymetrix als Agilent saturatie verschijnselen vertoonden. Ook uit statistische testen bleek dat Agilent niet in staat was om voor een aantal

[1] copies per cell

spikes te discrimineren tussen een concentratie van 1.000 en 10.000, terwijl CATMA hier geen problemen ondervond. Voor alle platformen kon geen cross-hybridisatie van de spikes gevonden worden, ook al werden die in heel hoge concentratie toegevoegd. Gebaseerd op het achtergrond staal, konden we de *in vivo* coverage, het percentage valse positieven en de signaal-ruis relatie bekijken. Voor de *in vivo* coverage — gemeten als het aantal genen met een signaal boven de achtergrond — hadden Affymetrix en CATMA een sterke overeenkomst, terwijl bij Agilent meer dan 90% van de genen een signaal boven de achtergrond vertoonden, zodat deze statistiek weinig betekenis heeft. Verder toonde de signaal-ruis relatie een sterk verschil tussen de MAS 5.0 en de RMA normalisatie, die een veel hogere reproduceerbaarheid, onafhankelijk van de intensiteit, had.
In het algemeen kunnen we stellen dat CATMA gemakkelijk de vergelijking doorstaat en een uitstekend alternatief is voor de commerciële platformen. Deze platform vergelijking werd gepubliceerd in Allemeersch et al. (2005).

Hoofdstuk 5: Het CAGE project

In November 2002 startte het *Compendium of Gene Expression* of CAGE project, een Europees demonstratie project van 3 jaar, in samenwerking met acht laboratoria en 2 bioinformatica groepen. De acht laboratoria zouden 1.000 biologische stalen produceren, met een biologische herhaling. Deze stalen bevatten verschillende plantonderdelen van 3 ecotypes, een aantal stress condities en mutanten. Elk laboratorium kon ongeveer de helft van de stalen definiëren in functie van hun eigen onderzoek. De stalen zouden gehyhridiseerd worden met een technische herhaling; wat resulteert in 4.000 hybridisaties. Als platform werd de CATMA array gebruikt. De grootste bijdrage van onze groep was het preprocessen en de kwaliteitscontrole van de data.
Binnen het CAGE project, werd geopteerd voor een referentie design, waarbij elk staal ten opzichte van hetzelfde referentie staal gehybridiseerd wordt. Als referentie wordt een mix van de 16 primers $r1, \ldots, r16$ gebuikt. Omdat aan elke GST één van die 16 primers is toegevoegd, zal dit referentie staal op elke probe binden. Hierdoor heeft het geen betekenis

meer om alle log-ratios te normaliseren rond 0 met een Loess regressie. In de plaats zullen we algemene lineaire modellen ("General Linear Models" (GLM)) gebruiken die de effecten van de verschillen tussen de 16 primers in het referentie kanaal (zie Figuur 5.3 op pagina 120) en de print-tip effecten in beide kanalen verwijderen.
Deze within-slide normalisatie is geïmplementeerd in een automatische dataverwerkingspijplijn. Omdat er binnen het CAGE project gekozen werd voor het gebruik van het MAGE-ML formaat om de data op te slaan en uit te wisselen en omdat we voor de data analyse gekozen hebben voor het statistische programma R, was de eerste taak het maken van een import functie voor de MAGE-ML files in R. Dit deel werd gemaakt als een zelfstandig R-pakket `RMAGEML`, dat niet alleen MAGE-ML files kan importeren, maar ook, bijvoorbeeld, genormaliseerde waardes kan toevoegen aan de MAGE-ML files. Het pakket is ondertussen opgenomen bij de Bioconductor pakketten (`www.bioconductor.org`) en gepubliceerd in Durinck et al. (2004).
De pijplijn wordt ondersteund door een lokale MySQL databank, die bijhoudt welke experimenten genormaliseerd zijn en waar de bijhorende files opgeslagen zijn. Om de pijplijn te laten lopen, wordt een Perl script opgeroepen en dit script controleert of er een nieuw experiment beschikbaar is, door de genormaliseerde experimenten opgeslagen in de lokale databank te vergelijken met de lijst beschikbare experimenten op de ftp site van de European Bioinformatics Institute (EBI, `www.ebi.ac.uk`). Indien er een nieuw experiment gevonden wordt, wordt dit gedownload. Vervolgens roept het Perl script een R-script op, die de data normaliseert en de databank aanvult. De MAGE-ML file wordt geupdate met de genormaliseerde waardes en er wordt een HTML pagina per hybridisatie gemaakt, die statistieken en figuren bevat, die toelaten voor een kwaliteitscontrole. Het Perl script plaatst vervolgens de MAGE-ML file met de genormaliseerde waardes terug op de ftp site van de EBI en checkt of er nog experimenten wachten op normalisatie.
De data productie ging veel trager dan aanvankelijk verwacht werd. De eerste data zou geproduceerd worden na de eerste 6 maanden van het project, maar kwam uiteindelijk pas in de 31ste maand van het project (zie Figuur 5.6, pagina 129). Een data analyse van het compendium lag dan ook niet meer binnen het tijdsbestek. Een kleine analyse, waarbij

twee, dezelfde tijdreeksexperimenten van bladontwikkeling, geproduceerd door twee verschillende partners, met elkaar vergeleken worden, tonen wel dat de genexpressie patronen voor de significante genen gelijkaardige patronen vertonen.
Het CAGE project was wel een aanleiding om samen te werken met verschillende partners voor de data analyse van hun onderzoeksexperimenten. Een voorbeeld van een dergelijke analyse is een experiment in samenwerking met de VIB-PSB dat het effect van licht stress nagaat op catalase deficiënte planten, onder normale en hoge concentraties van CO_2. Met mixed model technieken, zoals voorgesteld in Wolfinger et al. (2001), werden de significante genen eruit gefilterd. Clustering van deze genexpressieprofielen leidde tot een groep genen, die duidelijk geactiveerd worden door fotorespiratorisch H_2O_2. Een biologische validatie van het experiment is nog niet gebeurd.

Hoofdstuk 6: Analyse van array CGH loop design experimenten

Dit hoofdstuk is gewijd aan een andere toepassing van microroosters, *array CGH*, voor het detecteren van chromosomale afwijkingen bij patiënten door het vergelijken van genomisch DNA. Typisch wordt in een dergelijk experiment, gelijkaardig aan de klassieke microrooster experimenten met twee kanalen, een test patiënt vergeleken met een normale, referentie patiënt op één slide en een tweede slide wordt dan gebruikt voor de dye-swap. Een groot nadeel van deze methode is de normale patiënt ook afwijkingen kan vertonen en deze kan men niet onderscheiden van de afwijkingen van de test patiënt. Een duplicatie van een kloon bij de test patiënt kan men niet onderscheiden van een deletie bij de referentie patiënt, en omgekeerd. Het design dat hier wordt voorgesteld is een loopdesign, waarbij 3 patiënten met elkaar vergeleken worden, zoals getoond in Figuur 6.1, op pagina 148. Voor de data analyse van dit loopdesign stellen we twee methodes voor: de mixed model aanpak (Wolfinger et al. (2001)) en LIMMA (Smyth (2004)). De mixed model aanpak schat de effecten op basis van de intensiteiten, terwijl LIMMA

werkt met de log-ratios van de intensiteiten. Beide methodes werden vergeleken op basis van de *signal-to-noise ratio* (SN)

$$SN_{\text{dupl/del}} = \frac{|\text{mean}_{\text{dupl/del}} - \text{mean}_{\text{non-aberrant}}|}{\sqrt{\frac{1}{2}\left(\text{var}_{\text{dupl/del}} + \text{var}_{\text{non-aberrant}}\right)}}$$

en het aantal valse positieven en valse negatieven. Voor beide statistieken bleek dat LIMMA de beste keuze was. De methode werd ook nog verder verfijnd, door het onderscheid te maken tussen een volledige en een partiële deletie of duplicatie.
De methode werd geïmplementeerd als een webapplicatie, die de gebruikers toestaat om de gpr files te uploaden en een R-script op te roepen die de LIMMA methode uitvoert en een HTML pagina genereert met de resultaten.

Hoofdstuk 7: Conclusies

Deze thesis kadert voor het grootste deel binnen het Europees demonstratie project CAGE. Een eerste bijdrage bestond uit een performantie studie van de CATMA array, door deze te vergelijken met de twee commerciële platformen van Agilent en Affymetrix. Het experiment demonstreert dat de CATMA array op zijn minst een evenwaardig platform is.
De grootste bijdrage van ESAT was de ontwikkeling van een automatische dataverwerkingspijplijn en kwaliteitscontrole van de data. Omdat binnen het CAGE project geopteerd werd voor het MAGE-ML formaat voor data uitwisseling en omdat de dataverwerking in R gedaan werd, was een eerste, belangrijke stap het maken van import functies voor files in MAGE-ML formaat naar de statistische omgeving van R. Dit resulteerde in een Bioconductor pakket `RMAGEML` (Durinck et al. (2004)).
De dataverwerkingspijplijn maakte het mogelijk om de hybridisaties geproduceerd binnen het CAGE project op een automatische manier te normaliseren. Het is waarschijnlijk de eerste dataverwerkingspijplijn die start van data in MAGE-ML formaat en MAGE-ML files kan exporteren met genormaliseerde waardes.
Voor elke hybridisatie werd er ook een HTML pagina gegenereerd voor de kwaliteitscontrole. Het manueel nagaan van de kwaliteit van elke

hybridisatie en het verhelpen van de upload fouten en onnauwkeurigheden in de MAGE-ML codering heeft voor een significante verbetering van de data set gezorgd.
De data productie was trager dan verwacht en de data analyse van de eigenlijke data in compendium moet dan ook nog starten. Om efficiënt te werk te gaan, zal er nauw samengewerkt moeten worden met biologen. De eerste belangrijke stappen zullen zijn om nog meer inzicht te krijgen in de kwaliteit van de data, door bijvoorbeeld het gedrag van gekende genen na te gaan, en om de data te groeperen in grote blokken van vergelijkbare experimenten.

Een tweede toepassing van microroosters was het analyseren van loop design experimenten op array CGH. Twee methodes werden hierop vergeleken en LIMMA (Smyth (2004)) lijkt ons de meest aangewezen methode. Deze analyse methode werd geïmplemeteerd in een webapplicatie.
Voor deze tool zijn nog uitbreidingen mogelijk. Eens er een voldoende grote data set voor een bepaald array versie aanwezig is, zou men kloon specifieke standaard deviaties kunnen implementeren in de t-test, gebaseerd op alle data. De tool houdt nu ook nog geen rekening met de classificatie van de naburige klonen. Indien deze in rekening gebracht zouden worden, zou men ook afwijkende regio's kunnen definiëren naast individuële klonen.

Glossary

AQBC	Adaptive Quality Based Clustering, 136
CATMA	Complete *Arabidopsis* Transcriptome MicroArray, 36
cDNA	complementary DNA, 19
CGH	Comparative Genomic Hybridization, 29
DNA	deoxyribonucleic acid, 12
EST	Expressed Sequence Tag, 32
FDR	false discovery rate, 93
FISH	Fluorescence In Situ Hybridization, 28
FP	False Positive, 153
gDNA	genomic DNA, 29
GLM	General Linear Model, 62
GST	Gene-specific Sequence Tag, 36
IM	Ideal Mismatch, 50

IVT	in vitro transcription, 47
LIMMA	Linear Models for MicroArray data, 58
MAGE-ML	MicroArray Gene Expression Markup Language, 25
MAGE-OM	MicroArray Gene Expression Object Model, 25
MAS 5.0	MicroArray Suite 5.0, 49
MGED	Microarray Gene Expression Database group, 25
MIAME	Minimum Information About a Microarray Experiment, 25
MM	MisMatch, 47
mRNA	messenger RNA, 15
PCR	Polymerase Chain Reaction, 19
PM	Perfect Match, 47
REML	restricted maximum likelihood, 65
RMA	Robust Multi-array Average, 52
RNA	ribonucleotide acid, 14
rRNA	ribosomal RNA, 16
RT	reverse transcription, 18
SN	Signal-to-Noise ratio, 149
TAIR	The *Arabidopsis* Information Resource, 36
TIGR	The Institute for Genome Research, 70
TP	True Positive, 151
tRNA	transfer RNA, 16

Publication list

- Hilson P., **Allemeersch J.**, Altmann T., Aubourg S., Avon A., Beynon J., Bhalero R. P., Bitton F., Caboche M., Cannoot B., Chardakov V., Cognet-Holliger C., Colot V., Crowe M., Darimont C., Durinck S., Eickhoff H., Falcon de Languevialle A., Farmer E. E., Grant M., Kuiper M. T. R., Lehrach H., Leon C., Leyva A., Lundenberg J., Lurin C., Moreau Y., Nietfeld W., Serizet C., Tabrett A., Taconnat L., Thareau V., Van Hummelen P., Vercruysse S., Vuylsteke M., Weingartner M., Weisbeek P. J., Wirta V., Wittink F. R. A., Zabeau M., Small I. Versatile gene-specific sequence tags for Arabidopsis functional genomics: Transcript profiling and reverse genetics applications. Genome Research, vol. 14, no. 10b, Oct. 2004, pp. 2176-2189.

- Durinck S., **Allemeersch J.**, Carey V. J., Moreau Y., De Moor B. Importing MAGEML format microarray data into BioConductor. Bioinformatics, vol. 20, no. 18, Dec. 2004, pp. 3641-3642.

- **Allemeersch J.**, Durinck S., Vanderhaeghen R., Alard P., Maes R., Seeuws K., Bogaert T., Coddens K., Deschouwer K., Van Hummelen P., Vuylsteke M., Moreau Y., Kwekkeboom J., Wijfjes A. H. M., May S., Beynon J., Hilson P., Kuiper M.T.R. Benchmarking the CATMA microarray: a novel tool for Arabidopsis transcriptome analysis. Plant Physiology, vol. 137, Feb. 2005, pp. 588-601.

- Denolet E., De Gendt K., **Allemeersch J.**, Engelen K., Marchal K., Van Hummelen P., Tan K. A. L., Sharpe R. M., Saunders P. T. K., Swinnen J. V., Verhoeven G. The effect of a Sertoli cell-selective knockout of the

androgen. Molecular Endocrinology, vol. 20, no. 2, Feb. 2006., pp. 321–334.

- **Allemeersch J.**, Van Vooren S., Hannes F., De Moor B., Vermeesch J., Moreau Y. An experimental loop design improves the detection of congenital chromosomal aberrations by array CGH. Internal report 06-190, *Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven* (Leuven, Belgium), 2006.

Chapter 1

Introduction

This opening chapter gives a general introduction to microarrays and addresses in short the different aspects of microarrays that will be discussed in the course of the thesis. The chapter concludes with an outline of the thesis and an overview of the contents of the following chapters.

The completion of sequencing programs, as for example the Human Genome Project (The International Human Genome Sequencing Consortium (2001)), was not the end, but merely the start of the unraveling of the information hidden in the DNA. Buried within the DNA sequences are the *genes* (i.e., DNA sequences that code for proteins), which determine almost all the inherited characteristics of species. The set of genes that are expressed in a cell gives an indication of the state of the cell (i.e., its location, developmental stage, shape, stress response, etc). Gaining insight in the location of those genes within the genome, their functions and mutual interactions, will lead to a higher understanding of the functioning of each organism.

An important tool to gain insight in gene functions, which will have a central position in this work, are *DNA microarrays*. They enable to study gene expression — namely to measure when, where, and to what extent a gene is active — for thousands of genes simultaneously. Microarrays were introduced in the 1990s in various forms (i.e., nylon macroarrays with radioactive detection (Gress et al. (1992)), glass microarrays with fluorescent detection (Schena et al. (1995)), and oligonucleotide chips with fluorescent detection (Lockhart et al. (1996)), Jordan (2002)) and caused a revolution in functional genomics. They allow to screen thousands of genes in parallel, in contrast to the earlier techniques which could only focus on a few genes at a time. Therefore, microarrays have become a popular research tool and the use of microarrays grows exponentially (see Figure 1.1).

In its early days, microarray experiments were set up as small-scale experiments. They included only a limited number of hybridizations. Nowadays, the tendency is growing towards large compendia projects, that produce a large number of hybridizations and provide information on the gene expression at different developmental stages, under different environmental conditions, or of different mutants (Moreau et al. 2003). More and more, this data is made available to the community.
To ensure that the wealth of data pouring out of such compendia is turned into meaningful knowledge, is a great challenge, which has been accepted by the science of computational biology or bioinformatics. Bioinformatics combines techniques from applied mathematics, informatics, and statistics, to facilitate the handling of these huge amounts of data.
This thesis will focus mainly on the statistical aspect of microarray data analysis. Currently, there exists no standardized way of microarray data analysis — a statistical tower of Babel (Allison et al. (2006)) — and it does not look as if such standardization will be obtained in the near future, but the microarray community is striving for it. Recently, a US-wide initiative, MAQC (`http://www.nature.com/nbt/focus/maqc/index.html`), has been taken, aiming to provide quality control tools to the microarray community and to develop guidelines for microarray data analysis. Carefully selecting the appropriate tools for preprocessing and analysis

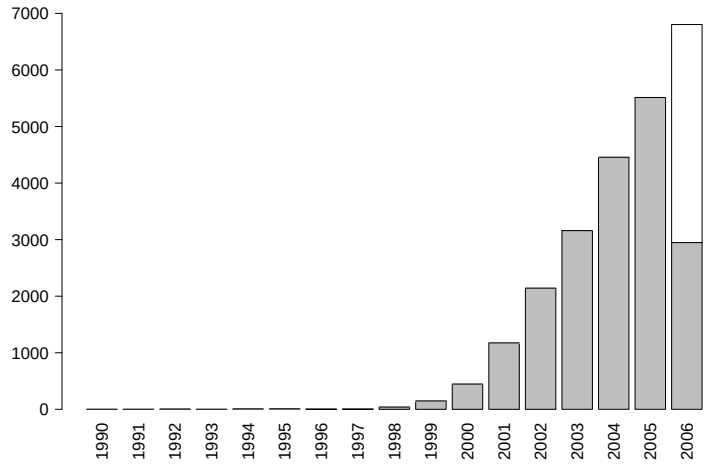


Figure 1.1: **Number of publications on microarray related research.** The histogram shows a rapid increase in number of publications involving microarrays over the last 10 years. The numbers correspond to the number of publications containing 'microarray', 'microarrays', 'micro array', or 'micro arrays' in the titles or abstracts as stored in the pubmed data base (`www.pubmed.gov`). The number of publications for 2006 was only counted until May, 2006, and then extrapolated. The numbers were provided by Steven Van Vooren.

of microarrays is vital to distinguish biological information from chance variation and hence, to avoid misleading results.

Deliverables and achievements

The majority of the work presented in this thesis has been done within the framework of the *Compendium of Arabidopsis Gene Expression* or CAGE project, a European Demonstration project, that aimed at building a compendium of gene expressions of *Arabidopsis thaliana* throughout its life cycle and under a variety of stress conditions. The ultimate goal

was the production of 2,000 biological samples, hybridized on 4,000 microarrays. The samples were grown and hybridized by eight different laboratories in Europe.

Different microarray platforms exist, each requiring its specific normalization needs (see Chapter 3). Within the CAGE project, the Complete *Arabidopsis* Transcriptome MicroArray or CATMA array was chosen. This platform owns certain advantages over the more classical cDNA platform, as the probes are designed to guarantee a higher gene specificity and to avoid cross-contamination (Hilson et al. (2004), Chapter 3). Whether CATMA is a good alternative to the more expensive, commercial platforms had not yet been demonstrated at the start of this thesis.
Therefore, one of the first tasks of this work was to benchmark the CATMA array against two commercial, oligonucleotide-based platforms. In an experiment, set up for this purpose, the same sample was hybridized on all three platforms. The sample was chosen such that it allowed to assess and compare all aspects of the resulting expression values in detail. Personal contribution to this benchmarking experiment was the preprocessing of the CATMA and Affymetrix data (Section 4.5), and the actual statistical analysis of the platform comparison. The latter allowed to draw conclusions on the dynamic range and sensitivity, *in vivo* coverage, specificity, signal reproducibility, false positive rate, and false negative rate (presented in Sections 4.6 - 4.11). A complete overview of this benchmarking is shown in Chapter 4 and is published in Allemeersch et al. (2005).

The size of large compendium projects, such as the CAGE project, bears consequences towards data preprocessing and storage. To communicate between all different partners involved in the project, and to make the data really usable to the community, all data has to be well-annotated so that researchers can analyze the experiment appropriately, interpret the results correctly and reproduce the experiment. The *MicroArray Gene Expression Markup Language* or MAGE-ML, an XML language, has been developed by the Microarray Gene Expression Database group (MGED; `http://www.mged.org`) as a standard for microarray data description. The CAGE project was one of the first projects that wished to use this

MAGE-ML format actively (i.e., not only to store the data, but also to extract the data again). However, no facilities were provided to extract the data and to import it in a statistical environment. In this work, a tool enabling to import data in the statistical environment R and to add preprocessed data to the MAGE-ML files is presented. This tool is available as an independent R-package, RMAGEML, which has been added to the Bioconductor packages. The work has been published in Durinck et al. (2004).

This RMAGEML package has been used to construct a data preprocessing pipeline for the CAGE project. The preprocessing of the 4,000 microarrays requires an automated approach. Therefore a data preprocessing pipeline is presented that starts from the data, as stored in MAGE-ML format, normalizes the data within slide, according to the experimental design that was used, and updates the MAGE-ML file with the corrected intensities. In the meantime, images and statistics for quality assessment are generated and made available to the partners. A database system is also updated to keep track of all preprocessed data. This data preprocessing pipeline is presented in Chapter 5.
Preprocessing of data depends on the design that has been used in an experiment. For the majority of the hybridizations in the CAGE project, a special kind of reference design is used; all samples are hybridized against an artificial reference sample, that produces a constant signal in the reference channel. This design made the typically used normalization approach (i.e., Loess normalization) inappropriate and forced us to find an alternative preprocessing method. We propose in this work a within-array normalization, using General Linear Models (Chapter 5). This normalization, along with a Loess normalization for the classical dye-swap experiments, has been implemented in the preprocessing pipeline.

Starting from the data generated in the preprocessing pipeline, the more exciting work can start: the actual analysis of the data generated in the CAGE project. However, as there was a serious delay in the data production, this analysis could not be completed within the framework of this thesis. Therefore, we have to restrict ourselves to a short preview of a comparison between two partners in Chapter 5. The chapter concludes with an experiment on the influence of high-light stress on catalase

deficient plants, in collaboration with one of the CAGE partners.

The last topic, outside of the scope of the CAGE project, is the analysis of arrayCGH data. *ArrayCGH* is a microarray platform for the detection of chromosomal deletions and duplications. The detection of such aberrations in specific patients enables us to gain insight in the origination and development of diseases. In this work, a specific analysis tool for array CGH loop designs is presented and demonstrated. This design places three patients in a loop design, which has advantages over the classical two-by-two, dye-swap comparisons. Two different analysis methods are introduced (i.e., mixed models on the absolute intensities and a linear model of the log-ratios of the intensities (LIMMA)). Both methods are compared, based on signal-to-noise ratios and true and false positive rates. LIMMA turned out to be preferable and is implemented in a web-based application.

A more schematic overview of the contents and the organization of the thesis is shown in Figure 1.2 and in Table 1.1.

All analyses in R with BioConductor

All analyses shown in this work have been performed with the free, open-source statistical software *R* (`http://www.r-project.org/`), an implementation of the statistical programming language S. A second, commercial implementation is S-Plus. The initiative for *R* was taken in 1995 by Ross Ihaka and Robert Gentleman (hence the name *R*) at the Department of Statistics of the University of Auckland in Auckland, New Zealand (Ihaka and Gentleman (1996)). *R* is an ongoing project and a large group of individuals has contributed to it, by adding or debugging *R* code.
The community can also add packages that group a number of classes and functions for a specific task. An excellent and successful example is the Bioconductor project (`http://www.bioconductor.org`; Gentleman et al. (2004)), where statisticians and bioinformaticians have developed and are still developing a number of packages for the analysis of genomic data. The project started in 2001 and takes now the lead in the

analysis of microarray data. Bioconductor comprises analysis tools and graphical methods for the preprocessing of microarray data, identification of differentially expressed genes, and graph theoretical analysis. For annotation, it provides links to the public databases and mappings between different probe identifiers.
The analyses presented in this work will make use of Bioconductor and show its strength.

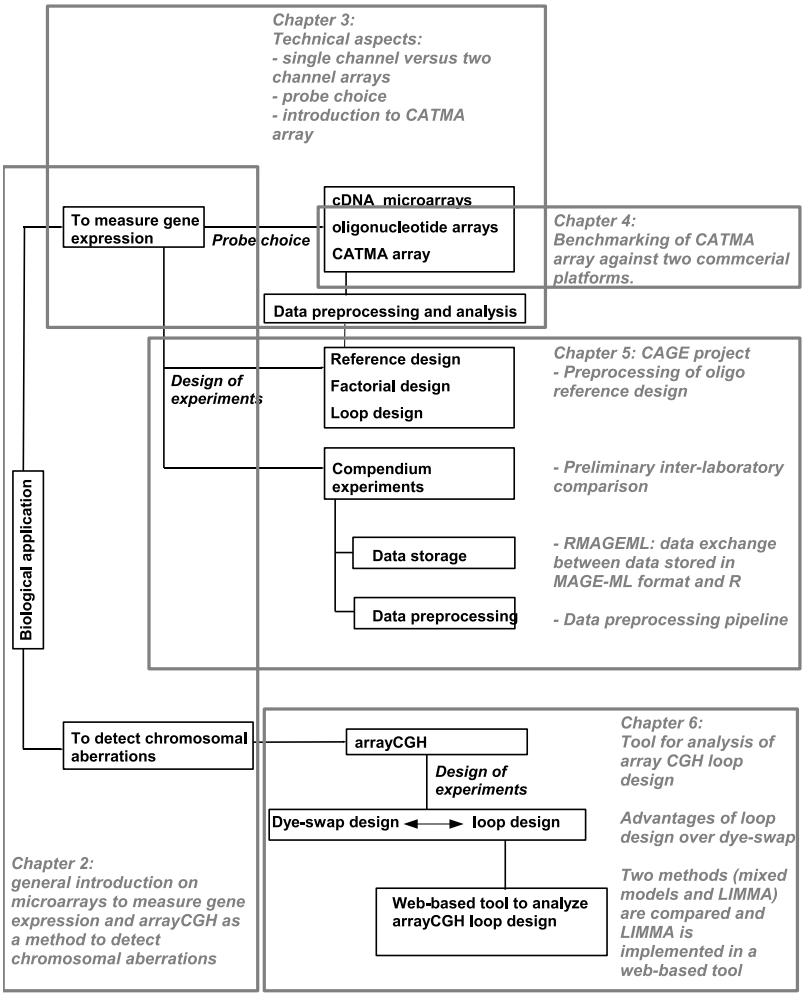


Figure 1.2: **Organization of the thesis.** This figure provides an overview of the contents of the thesis and how it is organized in the different chapters.

Topics & achievements	Chapter	Publication
General introduction on basic concepts in cell biology, the idea of microarrays to measure gene expression, and arrayCGH to detect copy number variations	2	
Technical aspects of microarrays to measure gene expression. Discussion of two and single channel arrays, with a number of alternatives for probe design. Introduction of the CATMA array, an array designed for *Arabidopsis* transcriptome analysis	3	Hilson et al. (2004)
Benchmarking of the CATMA array against two commercial oligonucleotide platforms (i.e., Agilent and Affymetrix), to assess dynamic range, sensitivity, specificity, reproducibility, and false discovery rate	4	Allemeersch et al. (2005)
Presentation of Compendium of Arabidopsis Gene Expression or CAGE project. Contributions: - Choice of appropriate preprocessing method for the oligo reference design - Data preprocessing pipeline for an automated quality assessment and data preprocessing - RMAGEML: an R package to facilitate communication between data stored in MAGE-ML format and the R environment for data analysis - Preliminary inter-laboratory comparison of data generated in CAGE - Analysis of a research experiment within CAGE to assess the influence of high-light stress on catalase deficient *Arabidopsis* plants under normal and high concentration of CO_2	5	Durinck et al. (2004)
ArrayCGH to detect chromosomal aberrations. Discussion of the advantages of a loop design over the classical dye-swap design. Development of a web-based analysis tool of arrayCGH loop designs. This work is done in close collaboration with the Department of Human Genetics. Contributions: - Proposal of two analysis methods: mixed models and LIMMA - Benchmarking of both methods	6	

Table 1.1: Contents and achievements in this thesis.

Chapter 2

Microarray technologies to measure gene expression

In this introductory chapter, the reader gets acquainted with all concepts that will be used in this work. A general introduction to gene expression is given. Starting with the structure of the DNA and via the Central Dogma, the concept of gene expression will be introduced. Gene expression will be measured with microarray technology, a high throughput technology to measure gene expression for thousands of genes simultaneously. We will stress on the complexity of the data generated by microarrays and point at the necessity for standardized data formats, to facilitate the interpretation and integration of the data.
A second, completely distinct class of arrays, array CGH, is introduced at the end of the chapter, as a technique to detect chromosomal aberrations.

2.1 Cell biology in a nutshell

This section introduces the concepts in cell biology essential to understand gene expression. The structure of DNA is explained, and its mechanism for the coding of proteins.

2.1.1 The structure of DNA

In the 1940s *deoxyribonucleic acid* (DNA) was conjectured to be the carrier of the genetic information in organisms (Avery et al. (1944)). But the mechanism by which the DNA gives instructions to the cell and how the information contained in the DNA was passed on from a cell to its daughter cells was not clear, until the determination of the double helix structure of the DNA by James Watson and Francis Crick in 1953.
In that model a DNA molecule consists of two long polynucleotide chains, the *DNA strands*, built out of four types of nucleotides (see Figure 2.1). These consist each of a sugar, deoxyribose, attached to a single phosphate group and a base, which can be adenine (A), cytosine (C), guanine (G), or thymine (T). The two DNA strands are each formed by a chain of alternating sugars and phosphates. These two chains are held together by hydrogen bonds between the bases. Because of the chemical structure of the bases, hydrogen bonds can exist exclusively between the pairs A and T and between C and G. This is called *complementary base pairing*. As adenine (A) and guanine (G) consist of two rings, while cytosine (C) and thymine (T) of one single ring, complementary base pairing leads to base pairs of equal width and keeps the two DNA strands at equal distance. The two strands wind around each other and form a double helix.
In each strand the subunits of the nucleotides are lined up in the same direction, which gives a polarity to the strand. According to this polarity, one end of the strand is called the $3'$ end and the other end is the $5'$ end, which corresponds to the numbering of the carbon atoms of the sugar molecule. The complementary strand has then the opposite direction, from $5'$ to $3'$ (see Figure 2.2).

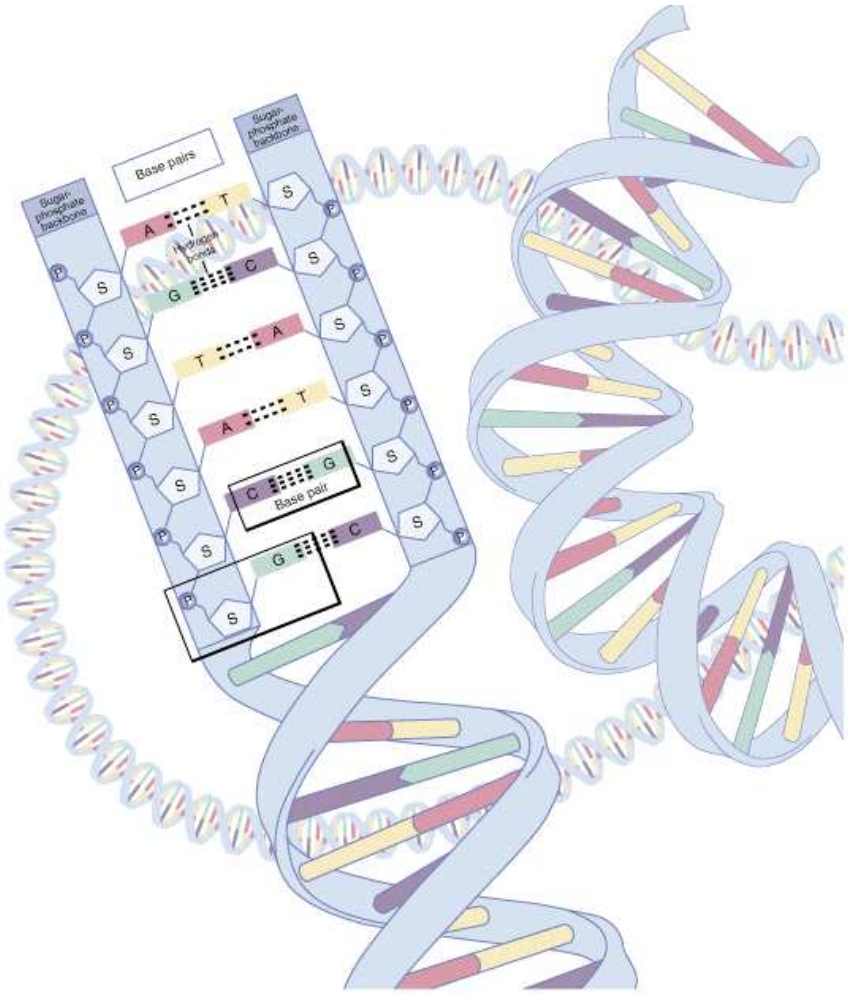
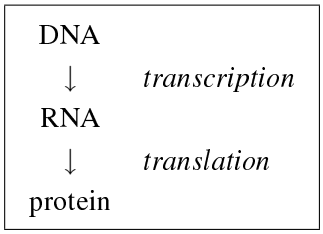


Figure 2.1: **The DNA molecule.** The double helix of the DNA is shown along with details of how the bases, sugars and phosphates connect to form the structure of the molecule. DNA is a double-stranded molecule twisted into a helix. Each strand, comprised of a sugar-phosphate backbone and attached bases, is connected to a complementary strand. The bases are adenine (A), thymine (T), cytosine (C), and guanine (G). A and T are connected by two hydrogen bonds. G and C are connected by three hydrogen bonds. The figure was obtained from the National Human Genome Research Institute (`http://www.genome.gov/glossary.cfm`).

2.1.2 DNA codes the formation of proteins

The information contained in the DNA instructs the synthesis of proteins. A protein can be seen as a long chain built from a selection of 20 amino acids; this chain is folded in a 3D structure. The proteins execute most of the functions in the cell.
Pieces of DNA that contain the necessary information to synthesize a protein are called *protein coding genes*. The transformation of genes into proteins can roughly be described in two steps: *transcription*, which transcribes DNA into RNA, and *translation*, which translates RNA into proteins (Figure 2.3). This fundamental principle



is often called the *central dogma* of molecular biology. Both steps will be explained here into more detail.
Next to the protein coding genes, there exists also a second type of genes in the genome: the *non-coding genes*. Non-coding RNA genes encode functional RNA molecules. Many of these RNAs are involved in the control of gene expression, particularly protein synthesis.

Transcription: From DNA to RNA

The *transcription* process copies a piece of the DNA information. Therefore the enzyme *RNA polymerase* moves along the DNA and unwinds and opens the DNA helix in front of it. One of these strands (in the direction from 3′ to 5′) serves then as a template and with this template a new strand is formed by complementary base pairing with incoming ribonucleotides. This new single stranded chain of ribonucleotides is called *ribonucleotide acid* or RNA. Next to the fact that RNA is single stranded, it differs also from DNA by its content. The nucleotides contain the sugar ribose instead

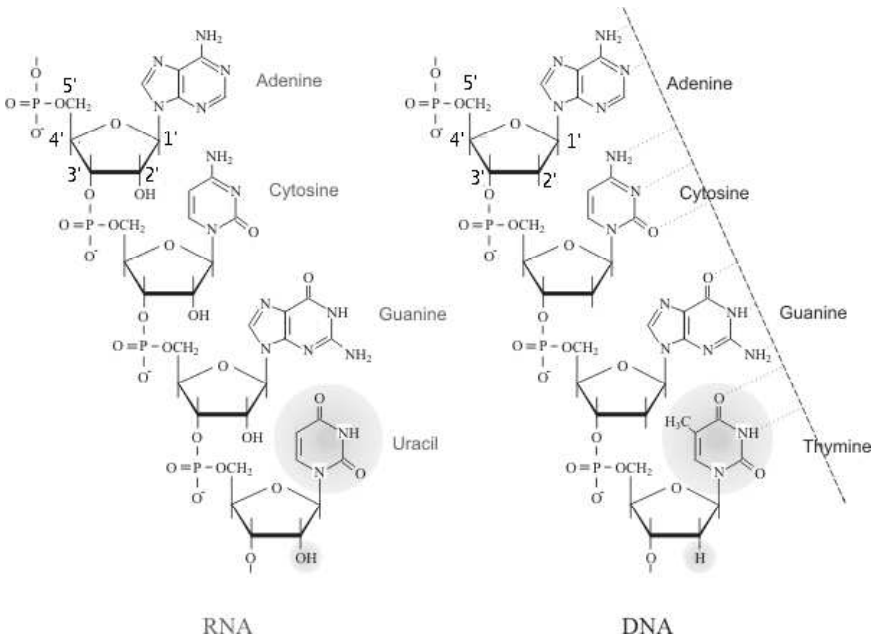


Figure 2.2: **The structure of the RNA and DNA molecule.** The chemical structure of both RNA and DNA is displayed. In contrast to DNA, RNA contains the base uracil instead of thymine and it is single stranded. The figure can be found at `http://www.ktf-split.hr/glossary/en_o.php?def=nucleic%20acid`.

of deoxyribose and RNA contains also the bases adenine (A), cytosine (C), and guanine (G), but the fourth base is uracil (U) instead of thymine. As thymine, uracil can also only base-pair with adenine. However, the strength of the binding between uracil and adenine is much lower than that between thymine and adenine. As a result, RNA does not form stable double helices, but only partially hybridizes with itself (see Figure 2.2).

Different kinds of RNA can be produced. The majority of the RNA molecules will specify the amino acid sequence of the proteins and is called *messenger RNA* or mRNA. But there exists also RNA that has a function on its own, as we will see in the translation step.

For eukaryotes, the transcription process takes place in the nucleus of the cell. The translation step, in which the actual protein is formed, will take place outside of the nucleus, in the cytoplasm, but before the mRNA is exported from the nucleus, the mRNA is *processed*. At the $5'$ end, an atypical nucleotide is attached (the base guanine along with a methyl group) and at the $3'$ end, the RNA is cut off and a series of A nucleotides are added. This is called the *poly(A) tail*. These two additions make the mRNA more stable and therefore more feasible to export the mRNA from the nucleus. Apart from these small changes, people also discovered in the 1970s that only a small part of the mRNA sequence is coding for a protein. Before the mRNA leaves the nucleus, *RNA splicing* takes place. In this process, the non-coding sequence parts (*introns*) are cut out of the primary mRNA and only the coding parts (*exons*) are retained and get into the cytoplasm. In prokaryote cells, which have no nucleus, the protein synthesis is less complicated. Both processes — transcription and translation — occur at the same place in the cell and there is no notice of introns and exons. Hence there is no processing step of the mRNA and the translation process often starts before the transcription step is actually finished.

Translation: From RNA to proteins

The *translation* step can be described as a decoding step in which the mRNA strand, composed of 4 nucleotides, is translated into a chain built from 20 different amino acids. The code behind the translation is that each group of three nucleotides (*codon*) codes for one specific amino acid. *Transfer RNA* (tRNA) is a small molecule that can recognize and bind to a specific codon, again by complementary base pairing. Each tRNA is then also linked to the amino acid corresponding to the codon (see Figure 2.3). The deciphering is then effected in the cytoplasm by the *ribosome*—a big complex consisting out of more than 50 proteins and a number of RNA molecules (*ribosomal RNA* or rRNA). The ribosome moves along the mRNA strand starting from the $5'$ end. For each codon it captures the complementary tRNA and binds its amino acids to form the protein chain.

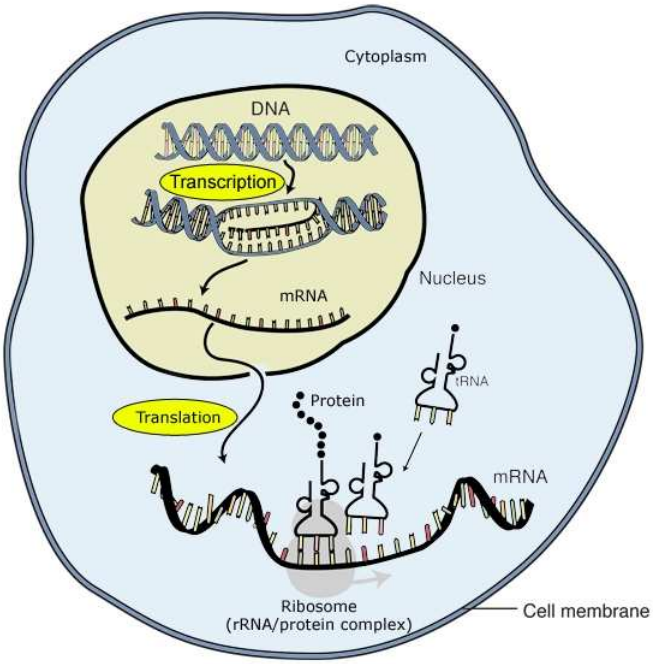


Figure 2.3: **Transcription and translation.** For eukaryote cells, transcription takes place in the nucleus of the cell. During transcription DNA is transcribed to RNA. This RNA is exported out of the nucleus into the cytoplasm, where the mRNA is translated into a protein. The figure was obtained from the National Human Genome Research Institute (`http://www.genome.gov/glossary.cfm`).

Gene expression

The concise description of the transcription and translation process illustrates how the information contained in the DNA of a gene is transformed into the construction of the proteins, which are essential for all cell processes. Hence, the type and the amount of copied information from the cell influences which proteins are produced and therefore, indirectly also the characteristics or the phenotype of the cell and consequently of the organism. If the DNA of a gene is transcribed into RNA, then this gene is

called *expressed*.
The initiation of the transcription by the RNA polymerase is tightly regulated by regulatory proteins, called *transcription factors*. They activate or repress the *gene expression* (i.e., the level of transcription of the DNA of the gene). Gene expression can be disturbed by modifications of the DNA sequence (insertions and deletions — likely to affect the protein sequence as well) or by other chemical modifications that do not change the nucleotide sequence itself (*epigenetics*). For example, most cancers involve the epigenetic silencing of genes that normally control cell proliferation.
Microarrays quantify the gene expression. And this is done on a global scale: the transcription abundance is measured for thousands of genes simultaneously. Whole-genome arrays enable biologists even to study the role of all known and predicted genes in a genome at once.
One of the fundamental criticisms on the microarray technology was that DNA microarrays measure the gene expression at transcription level and not the actual protein concentrations, which seem to be more directly related to the cell functions than the mRNA expression levels. However, measuring the expression at protein level genomewide is more difficult. Therefore measuring the gene expression at transcription level is still valuable as a resource for studying expression profiles. It provides us also with different, but therefore not much less valuable information.

2.2 Experimenting with DNA

Recent technological developments allow us to study the DNA in a way that was completely new and caused a revolution. We can sequence the DNA, isolate the DNA of a specific gene from a genome and replicate this DNA as many times as we wish. These techniques gave a new dimension to the study of the DNA and enabled microarray technology, and hence the study of expression of the genes, their function, and their interactions with each other, on a large scale.
As an example, we will introduce two commonly used methods, Reverse Transcription (RT) and Polymerase Chain Reaction (PCR). These two techniques will also be referred to when we explain the production of microarrays in Chapter 3.

2.2.1 Reverse transcription

Reverse transcription (RT) is a process in which double stranded DNA is formed from mRNA (i.e., we reverse a part of the central dogma). In general, synthesis of a new DNA strand requires a *primer*, a nucleotide sequence, that can serve as a starting point. A primer binds on the RNA and an enzyme will then add nucleotides to this existing strand. In case of RT, we need a primer that can bind on the mRNA to initiate the reverse transcription. As all mRNAs have a poly(A) tail (Section 2.1.2), an oligo(dT) primer, a chain of Ts, will recognize the mRNA in the solution. The enzyme *reverse transcriptase* is added and will build the first DNA strand, nucleotide per nucleotide, starting from the primer and complementary to the mRNA strand. The resulting strand of spliced DNA is called *complementary DNA* or cDNA.

2.2.2 Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) amplifies or generates a large number of copies of a DNA sequence. In general, PCR starts from double stranded DNA (see Figure 2.4). In a first step the DNA is heated and the two strands are separated. Two primers, each complementary to one of the two DNA strands and at the opposite side of each other, are added in a large amount. When the DNA cools down, the primers will bind to the two DNA strands. The enzyme *DNA polymerase* is added and the DNA is synthesized starting from the two primers. The DNA is heated again and after a number of PCR rounds, exponentially many DNA molecules are produced.

2.3 DNA microarrays

Already since the mid-1970s, we were capable of measuring gene expression, with techniques called Southern blotting and later with Northern blotting. *Southern blotting* is used to recognize a DNA sequences and uses a piece of DNA as probe, whereas *Northern blotting* uses a piece of messenger RNA as probe and is applied to recognize RNA sequences. For both methods the probe is radioactively labeled and distributed over a gel

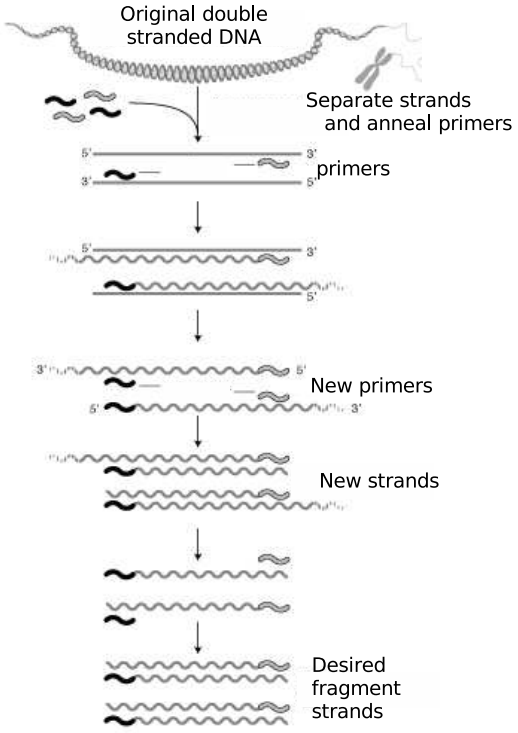


Figure 2.4: **Polymerase Chain Reaction.** In a first step the DNA is heated and the two strings are separated. Two primers are added in large amount and when the DNA cools down, these primers bind to the two DNA strands. Starting from the two primers, the DNA is synthesized. These steps are repeated several times. The desired piece of DNA is obtained and the amount of this DNA will grow exponentially. The figure was obtained from the National Human Genome Research Institute (`http://www.genome.gov/glossary.cfm`).

containing a sample of RNA or DNA. The complementary base pairing property of DNA makes that this oligonucleotide probe binds solely on its complementary strand. Measuring the amount of radiation, gives an

indication of the amount probe present in the RNA or DNA sample. Hence, it was possible to measure quantitative differences of expression for a selected gene.

A *microarray* can be understood as performing thousands of Southern or Northern blottings in parallel. Instead of distributing one probe over a gel with RNA or DNA, thousands of probes are now fixed on a solid surface and the RNA sample is spread over these probes. In general, a microarray can be described as a chip with up to 45,000 spots on a slide. Each spot contains DNA material (for details, see Chapter 3) of a known gene. In an experiment, RNA is extracted from a biological sample. This RNA is fluorescently labeled and brought into contact with the probes on the microarray. The genes will bind, or hybridize, exclusively to the probes on the microarray with a complementary sequence. The excess material is washed off. The microarray is then scanned and the fluorescence signal is measured. These intensities give an indication of the RNA levels in the biological sample. With this vague description a first introduction to the idea of microarrays is given. In Chapter 3, a more detailed description will be given, by presenting different microarray platforms. However, a complete overview of all possible microarray techniques with the different protocol choices for all steps in the process, is beyond the scope of this work.

2.3.1 Applications of microarray experiments

The power of microarrays to analyze thousands of genes in parallel increased the speed of experimental progress significantly. Over the past few years, the number of probes printed on the array increased and high density arrays now allow to measure gene expression genomewide — analysis of the entire human genome can now be done in one single run.
Microarrays are used in all fields of biology, for plants, animals and humans, and this for a variety of biological questions. Expression profiles can be compared between organisms at different developmental stages, under different environmental stress conditions, or in different disease states. The general goal of all these experiments is to find the function, the reg-

ulation of the genes and their interaction with other genes. Assessing the function of genes is mainly obtained by making the assumption that genes that share approximately the same expression patterns, are likely to have a similar biological function. Therefore, the classical output of microarray experiments consists of a number of clusters, showing genes with a similar behavior under different conditions.

2.3.2 Different analysis steps in a microarray experiment

The standard data analysis of a microarray experiment to obtain these clusters with similar expression profiles can be described in a few steps (Figure 2.5). The analysis starts from the data, as they come out of the scanner. These 'raw' data are tab-delimited files, that contain the intensities and a number of other characteristics of the spots, as spots size, a quality flag for each spot and so on. Before this data can be actually analyzed, some quality assessment and normalization steps are required.
The quality assessment can help to discover serious quality problems, or even mistakes that occurred in one of the preceding steps. If the quality assessment does not disclose any serious irregularities, that require reperforming one or more hybridizations, the analysis can continue and the data can be normalized.
Normalization serves to remove all bias in the data that is of a non-biological nature. Consider, for example, an experiment with 2 samples, each hybridized on a slide. By comparing the different slides, we want to ensure ourselves that the detected differences in expression relate to the difference between the samples and cannot be ascribed to the difference between the slides. Normalization can partly be done by a careful design of the experiment. If the experiment is, for example, set up in such a way that the use of the arrays is well balanced for the different samples, with a sufficient number of repeats for each sample, taking the average over all repeats will reduce these array effects significantly. Additional examples will be given in Paragraph 3.4. Next to an appropriate design of experiments, additional normalization or preprocessing steps have to be performed between and within arrays. These are platform specific and will be discussed in detail in Chapter 3.
Once the data are sufficiently preprocessed, one can start to detect which

genes are differentially expressed (i.e., have different expression levels for the different samples). To assess this, the classical statistical tools (e.g., t-test and General Linear models) can be used. A specific problem with this kind of analyses is that we perform these tests for all genes on the slide, which means thousands of tests at once. Therefore special care has to be taken for multiple testing. If we want to select a list of significantly differentially expressed genes, we want to control the false positive rate (i.e., the proportion of genes that are falsely called differentially expressed). Corrections to control the false positive rate are often rather conservative, so that a number of truly significant genes are missed. An alternative is to control the false discovery rate, which is the expected proportion of false positives among the tests found to be significant. We will see an example of that in Section 4.10. The eventual list of differentially expressed genes will then feed the clusters.
These analysis steps are performed by a statistician. For the interpretation of the clusters and the biological mechanism behind these expression profiles, statisticians have to go back to the biologist. They will try to find their way through this massive amount of data. Bioinformaticians try to provide tools that can help the biologists. Huge databases (OMIM, Gene Ontology (`http://www.geneontology.org/`)) are set up to collect gene annotation information from previous experiments. Another way to increase the power of the data analysis of an experiment is to combine the new data with data from previously performed experiments (Aerts et al. (2006); Moreau et al. (2003); Rhodes et al. (2002)).

2.3.3 Publicly available microarray data

One can use the published research results of an experiment, but often, people prefer to obtain the complete original data set. Making this high-throughput data available is highly desirable, as it not only allows scientists to replicate an experiment, it also permits them to add this publicly available data to their own data, which enriches their data and enables a faster growth of knowledge in the field.
The means to make this data available was often limited to referring to a web site, mentioned in the paper. But, as every author has its own data formats, data is fragmented and it becomes hard to compare and combine the

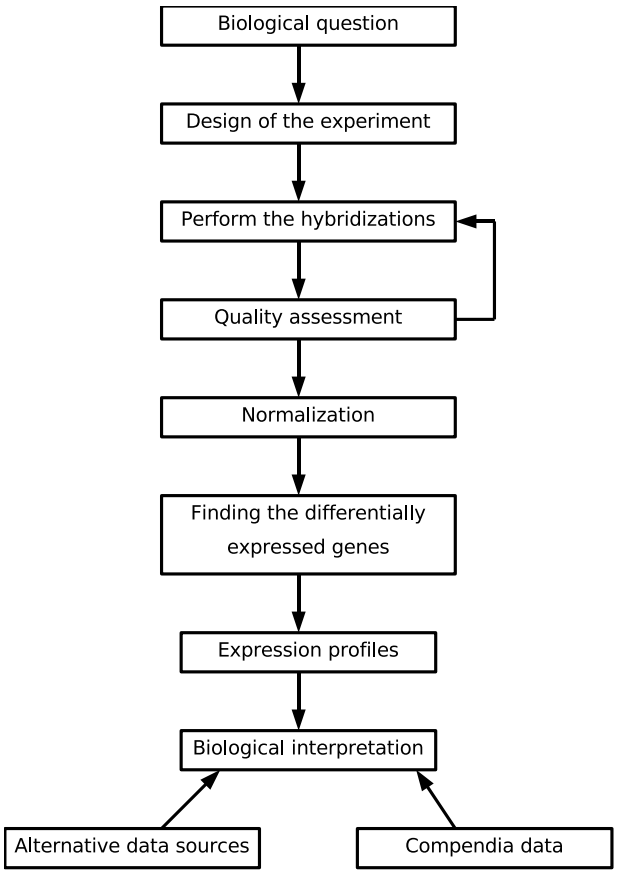


Figure 2.5: **Workflow in the analysis of a microarray experiment.** Starting from the specific biological question, addressed by the experiment, a carefully chosen design of the experiment is set up. Based on this design, the required hybridizations are performed. If the quality assessment detects no inferior hybridizations, data can be analyzed. For the resulting list of differentially expressed genes, the expression profiles can be assessed and clustered. Alternative data sources and microarray compendia data can assist for the biological interpretation of the expression profiles.

different data sources. To make this data meaningful for the community, the authors need to describe the experiment in full detail. This includes a description of the sample (i.e., its cell type and the environmental conditions to which is was subjected). There exists also a variety of microarray platforms. Hence, the platform that has been used and its probes on the platform have to be described. Although it is preferable to receive the data before any analysis has been applied to it, data is often processed and only the resulting expression values are available, lacking quality measures and indications for reliability. Therefore, at least all necessary data analysis steps to obtain the processed data require a detailed description. Another shortcoming is the maintenance of the authors' web sites, that cannot be guaranteed.

Minimum Information About a Microarray Experiment (MIAME). A first, important step towards standardization was initiated by the Microarray Gene Expression Database group (MGED; `http://www.mged.org`), which aimed to define the minimum information that is required to interpret the results of the experiment and to enable the reproduction of the experiment. They presented a document called *The Minimum Information About a Microarray Experiment* or MIAME (Brazma et al. (2001)), determining all information that should be provided to describe an experiment. The ultimate purpose of MIAME is to supply a structure that can be used to build public microarray databases in a way that they are meaningful to the community and enable data base queries.

MicroArray Gene Expression (MAGE). Based on the MIAME requirements, a data structure, called *MicroArray Gene Expression* or MAGE, has been developed (Spellman et al. (2002)). This comprises an object model, MAGE-OM, that consists of a number of 'packages', each describing a particular aspect of the microarray experiments. For example, the `AuditAndSecurity_package` contains the information on the contact that created or modified the data. In such a package a number of 'classes' are grouped. The `AuditAndSecurity_package` contains, for example, the classes `Organization` and `Person`. The classes can be described as a set of attributes and associations with other classes. The class `Person` has amongst others the attributes `lastName`,

`firstName`, and `address`.
To store this type of data, the *eXtensible Markup Language* or XML language was chosen. XML is a markup language, comparable to HTML, but, in contrast to HTML, it is specifically suited for the storage of the description of data, without any directions for its presentation. The data is structured with tags and attributes, that allow to extract the data rapidly from the XML file. The vocabulary that can be used in the XML is often defined in the *Document Type Definition* (DTD) — it declares the structure of the documents via a tag and attribute list, containing the allowable tags and attributes, respectively, along with a specification of its possible contents.
The MAGE object model has been translated into a DTD. This XML representation of the MAGE-OM is called the *MicroArray Gene Expression Markup Language* or MAGE-ML. A short example is shown in Figure 2.6.
Based on MIAME guidelines and MAGE, public microarray repositories have been built. At the moment, the two main databases are Gene Expression Omnibus (GEO; `http://www.ncbi.nlm.nih.gov/geo/`; Barrett et al. (2005)) at the National Center for Biotechnology Information (NCBI) and ArrayExpress (`http://www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html`; Brazma et al. (2003)) at the European Bioinformatics Institute (EBI). By submitting data to these public repositories, both the maintenance and the completeness of the data sets is guaranteed.

The use of MIAME is stimulated by the fact that prestigious journals as Nature, Cell, and The Lancet require MIAME compliant data sets as a condition for publication (DeFrancesco (2002)). Nature and Cell require even that authors submit their microarray data to one of the public repositories.

2.4 Alternative techniques to assess gene expression

Beside microarrays, a whole range of alternative techniques to measure gene expression on a genome-wide scale are available. Whereas microarray analysis is hybridization-based, others are sequence- or fragment-based. ESTs, SAGE, and MPSS are examples of sequence-based

```
   <AuditAndSecurity_package>
        <Contact_assnlist>
            <Person
                identifier="PERS:Wolfram_Brenner:CAGE_MPI"
                address="Ihnestrasse 73"
                phone="++49-(0)30-8413-1697"
                email="brenner@molgen.mpg.de"
                lastName="Brenner" firstName="Wolfram">
                <Roles_assnlist>
                    <OntologyEntry category="Roles" value="submitter"/>
                </Roles_assnlist>
                <Affiliation_assnref>
                    <Organization_ref identifier="ORG:MPI fr molekulare
Genetik:Lehrach"/>
                </Affiliation_assnref>
            </Person>
            <Organization
                identifier="ORG:MPI fr molekulare Genetik:Lehrach"
name="Lehrach">
                <Parent_assnref>
                    <Organization_ref identifier="ORG:MPI fr molekulare
Genetik"/>
                </Parent_assnref>
            </Organization>
            <Organization
                identifier="ORG:MPI fr molekulare Genetik"
                name="MPI fr molekulare Genetik" address="Ihnestrasse 73,
Berlin, Berlin, 14195, Germany"/>
        </Contact_assnlist>
    </AuditAndSecurity_package>
```

Figure 2.6: **MAGE - ML example.** A small piece of the MAGE-ML description of an experiment is shown, namely the `AuditAndSecurity` package. This part specifies for example the submitter of the experiment. The attributes of each class are listed with their corresponding values. Associations with other classes can be recognized by the tags with `assn` added to it. For example, `Affiliation` is an association to `Person`. As also `ref` has been added, it references to an Organization, which can be found by the `identifier` in the tag `Organization_ref`. This piece of MAGE-ML code is obtained from an experiment within the CAGE project (Section 5.1).

techniques, whereas cDNA-AFLP is fragment-based.

Gene expression measurement by high-throughput sequencing of Expressed Sequence Tags (ESTs; Okubo and Matsubara (1997)) involves counting of ESTs that are sequenced per gene. As this does not rely on previous sequence information, it is a valuable technique for the discovery of new genes. However, EST sequencing is laborious and expensive. Serial Analysis of Gene Expression or SAGE (Velculescu et al. (1995)) reduces the DNA sequencing effort by sequencing concatenated tags de-

rived from transcripts. SAGE is based on counting sequence tags of 14 bp from cDNA libraries. Contrary to EST, SAGE requires that the genome sequence of the organism or a substantial cDNA sequence database is available in order to identify the corresponding genes. To facilitate target identification, the LongSAGE method was developed by Saha et al. (2002). LongSAGE generates 21 bp tags, which allow unique assignment of tags to genomic sequences. However, quantification of lowly expressed genes requires sequencing of a large number of tags, which implies a high cost. Massively Parallel Signature Sequencing or MPSS (Brenner et al. (2000)) improves SAGE as it is a parallel sequencing method that can generate 100-1000 short sequence signatures in one single analysis. It also generates longer (16-20 bp) signatures to make gene identification more accurate. However the method is technical demanding. The cDNA-AFLP (Bachem et al. (1996)) technique applies the standard Amplified Fragment Length Polymorphism or AFLP (Vos et al. (1995)) protocol, as described for genomic DNA, on a cDNA template. This low cost procedure involves cleavage of the cDNA population by two restriction enzymes, followed by adaptor ligation to these fragments to allow for PCR amplification. The amplified fragments are then presented as a banding pattern on a sequencing gel. The differences in the intensity of the bands provide a good measure of the relative differences in the levels of gene expression. cDNA-AFLP does not require prior sequence information. Separately obtained datasets, however, cannot readily be compared.

2.5 CGH arrays

In this section a completely distinct class of arrays, CGH array, is introduced. CGH array is used in cytogenetics to detect chromosomal deviations. But to introduce this, we will go back in history.
In 1956, for the first time, the correct number of chromosomes was mentioned. Tijo and Levan suggested that a human cell contains 23 pairs of chromosomes (Figure 2.7) (Tijo and Levan (1956)). At the end of the same year this number was independently confirmed by Ford and Hamerton (Ford and Hamerton (1956)). Very rapidly chromosomal deviations from this rule were observed. Some cells contained more than two sets

Figure 2.7: **Human karyotyping.** From this picture it was established that human cells count $2n = 46$ chromosomes. This picture was reproduced from Trask (2002).

of chromosomes (*polyploidy*), others had an extra or a missing copy of one or more specific chromosomes (*aneuploidy*). Next to aberrations in chromosome number, also deviations in the chromosome structure have been observed. Segments of the chromosome can be removed or have an additional copy (i.e., *deletions* and *duplications*, respectively). The study of the detection of such chromosomal abnormalities is called *cytogenetics*. The link between specific chromosomal abnormalities and diseases was soon made, increasing the importance of cytogenetics in medicine. Already in 1959, the observation was made that trisomy 21 (i.e., three copies of chromosome 21 instead of two) causes Down Syndrome (Lejeune et al. (1959)). In fact, chromosomal deviations appear quite often; about 1 out of the 100-200 newborns have a chromosomal abnormality (Shaffer and Bejjani (2004)). Hence, detection of these aberrations and their influence on the phenotype becomes an important and interesting field.

Various techniques have been developed to detect chromosomal deviations (Oostlander et al. (2004)). At the end of the 1950s, it was possible to capture cells in their metaphase stage. At this second phase of mitosis the chromosomes are well condensed and spread from each other (Fig-

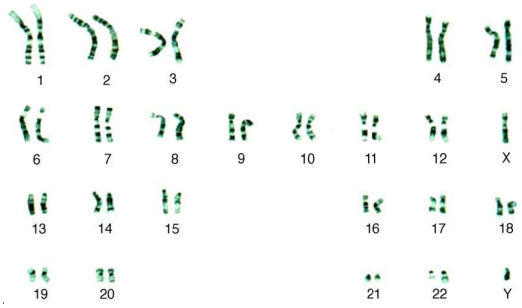


Figure 2.8: **Chromosomal banding.** After treatment a banding pattern becomes visible, which enables identification of every single chromosome (Smeets 2004).

ure 2.7) and it is then easy to arrange them into pairs (*karyotyping*). It was this technique that enabled to count chromosomes and to observe numeric chromosomal aberrations.

Structural chromosomal abnormalities remained invisible until the late 1960s. With chromosomal banding techniques (Figure 2.8) the metaphase chromosomes were stained such that dark and light bands become visible along the length of the chromosome. This allows not only detection of numeric chromosomal abnormalities, but also for detection of large structural aberrations. Alterations of 3-5Mb are detectable; for alterations smaller than 3Mb, the resolution of this method is not refined enough.

In the 1980s *fluorescence in situ hybridization* (FISH) was developed. Cloned segments of genomic DNA of specific chromosome regions are fluorescently labeled (Figure 2.9). These probes hybridize to their complementary sequences and produce a fluorescent signal at these specific locations on the human chromosomes. Fluorescent light gives a higher resolution and allows to test for a number of chromosomal locations at a time, as different fluorochromes can be used. But, for efficient use, you need prior knowledge of the type and location of the possible chromosomal abnormalities. The technique is also time-consuming as only a small number of segments (i.e., a minute part of the genome) can be tested at a time.

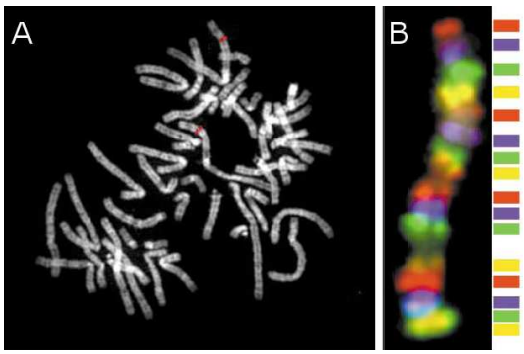


Figure 2.9: **FISH.** (A) FISH colors a specific DNA sequence, in this case a segment of 150 kb of chromosome 1, with a fluorescent signal (red). (B) Multicolour FISH: Several genomic sequences are analyzed simultaneously. Both pictures are reproduced from Trask (2002).

In 1992 *comparative genomic hybridization* (CGH) was introduced by Kallioniemi et al. (1992). *Genomic DNA* (gDNA) (i.e., DNA sequences that include exons and introns, coding and noncoding regions) of both a test and a normal, reference sample is isolated and labeled with fluorescent dyes in red and green, respectively. The DNAs hybridize to normal human metaphase chromosomes and ratios of the intensities of test versus reference signals display the chromosomal abnormalities. If a chromosomal region is present in both the test and reference sample, both samples will hybridize and we obtain a ratio of one. In case there is a deletion, only the reference sample will hybridize and we get a negative log-ratio. Similarly the log-ratio will be positive if a segment is duplicated. The CGH technique has still its limitations, as it makes use of metaphase chromosomes, which limits its resolution.
Recently these problems are solved by replacing the metaphase chromosomes by cloned DNA segments on an array as targets for the hybridization. These *CGH arrays* use arrays of up to $\pm$40,000 human clones, spread over all chromosomes in the genome. With CGH array not only the resolution has increased, you also do not need any prior information about the expected deletion or duplication, as was the case

with FISH, and it is less labor intensive. We will further elaborate on the array CGH technology in Chapter 6.

Chapter 3

Technical aspects of DNA microarray technologies

Microarrays can be split mainly into two classes, the two-channel arrays and the single-channel arrays. For two-channel arrays, two labeled samples are hybridized onto one single slide, resulting into two separate intensities. These intensity values are often reported as $\log_2$*-ratios. Whereas for single-channel arrays, only one sample is hybridized and this results in more or less absolute measurements. This major difference implies that the data coming from both types of platforms has to be treated distinctly and requires specific normalization methods. A further categorization can be made based on the probes used on the slides. The choice of the probes has an influence on the quality of the measurements, and has consequences towards normalization. In the following sections the different microarray types and their specific normalization methods will be presented. We restrict ourselves to only those techniques that were used in this work. At the end of the chapter an introduction to finding differentially expressed genes is given.*

3.1 Two-channel arrays

As two-channel arrays are used most in this work, we will first focus on these. For two-channel arrays, two samples are hybridized on the microarray, labeled with two different dyes. A schematic overview of these steps is shown in Figure 3.1.

3.1.1 Different probes

In a first section the classical cDNA microarrays will be described. The probes used on a cDNA microarray are complementary DNA (cDNA) clones. We will discuss two additional types of two-channel microarrays with alternative probe sets, namely the long oligonucleotide array from Agilent and the GST based CATMA microarray for *Arabidopsis thaliana.*

cDNA microarrays

A first step to build a microarray is to select the probes, that will be printed on the array. For cDNA microarrays, the probes can correspond to known genes, short (200 to 500 base pairs) DNA sequences that are part of a cDNA, i.e. an *expressed sequence tag* (EST), or cDNAs from libraries of interest. The actual probes are cDNA clones obtained from mRNA through reverse transcription (Section 2.2.1). To make sufficient cDNA clones to print on a microarray, the cDNA clones are amplified with PCR (Section 2.2.2). Therefore a second primer (complementary to the cDNA strand) is added and the DNA is amplified by many rounds of PCR. The PCR product is then spotted on the array by a set of pins. These pins dip into the PCR product, take a small amount of PCR product and drop it on the microarray surface. The cDNA probes in the PCR product are double stranded, therefore the array is heated, so that the DNA is separated and can bind to complementary strings.
As cDNA microarrays are two-channel arrays, two samples, an experimental and a reference sample, will be hybridized to the array. Hence, cDNAs are synthesized from the mRNA of the experimental sample and from the mRNA of the reference sample. These two samples are labeled with a red and green fluorescent dye, called *Cy5* and *Cy3*, respectively. There exists a number of methods for labeling: an overview can be found in Yang et al.

(2000). The labeled experimental sample is mixed together with reference sample and deposited on the array. If a gene is present in one or both of the samples, then it will bind to its complementary cDNA probe, again based on the complementary base pairing property. If it is present in both samples, the spot will emit a Cy3 and a Cy5 fluorescent signal. If it is present in only one of the samples, either a Cy5 or a Cy3 intensity will be measured, depending on in which sample it was expressed. If it is absent for both samples, it will not hybridize and no signal will be emitted.
Before the intensities can be measured, the array is washed to remove the unbound sample. A scanner is then used to measure the Cy3 and Cy5 signals. These signals then have to be preprocessed. This is described in Section 3.1.2.

The advantage of cDNA microarrays is the low cost, compared to alternative microarrays as oligonucleotide arrays (Sections 3.1.1 and 3.2.1), and therefore it was often used in the academic world. However, it has some serious shortcomings, which have eroded their attractiveness and by now they have essentially fallen out of favor. As the EST libraries are often far from complete and represent a fraction of the genes in a species, it is difficult to obtain full-genome coverage. The cDNA clones are also amplified by PCR from clones that grow in bacterial cultures and these cultures are often stored in well plates, each typically containing 96 or 384 wells. But bacteria can contaminate other wells and this cross-contamination leads to clone sets that contain other sequences than they are assumed to contain (Knight (2001)). Gene families also share a high degree of identity and cDNA probes are not guaranteed to be gene specific. It is possible that one of the genes in a family is expressed at a low level, while another member of the family is expressed at a high level. Because of the high sequence similarity, this highly expressed gene can also bind to the probe, designed for the low-expressed gene, although its sequence is not completely complementary. This phenomenon is called *cross-hybridization*. In such case, the expression of the low-expressed gene will be concealed by the high expression of the cognate gene.
An alternative for cDNA microarrays, that can bypass these shortcomings, is long oligonucleotide arrays.

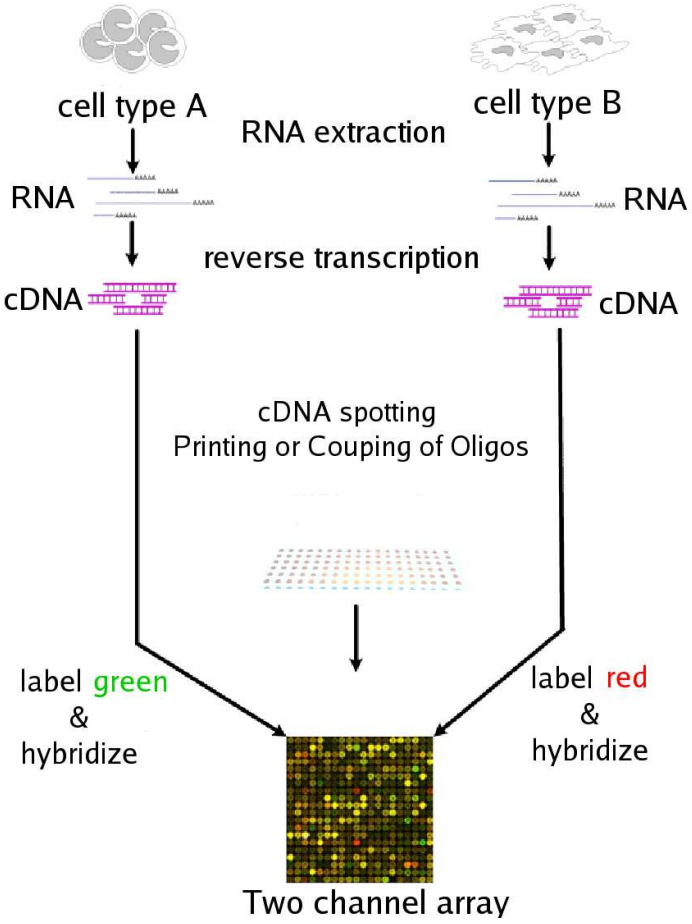


Figure 3.1: **Two channel microarray experiment.** For two different cell types A and B the RNA is extracted and cDNA is synthesized from the mRNA. The two samples are labeled in Cy3 and Cy5 and hybridized on the two-channel microarray. If a gene is expressed in the sample and present on the array, the gene will bind to the corresponding probe and a Cy3 or Cy5 signal will be emitted, depending on in which samples the gene is expressed.

Long oligonucleotide platforms

Nowadays many genomes have been sequenced and the availability of this sequence data allows people to design a microarray *in silico*. This led to a whole new class of microarrays, the oligonucleotide microarrays. Two types can be distinguished, the short and the long oligonucleotide arrays. As long oligonucleotide arrays are also two-channel arrays, their normalization is highly comparable to the cDNA microarrays and therefore, we will discuss them in this section. For short oligonucleotide arrays, we refer to Section 3.2.1.

Oligonucleotide arrays are created by pre-synthesizing the oligonucleotides *in silico* and then depositing them on the array, or by synthesizing the oligonucleotides *in situ*, directly on the surface of the array. Based on the sequence information alone, regions within genes can be selected in such a way that they offer greater specificity than the full-length cDNA clones (i.e., they are better able to distinguish closely related sequences and therefore limit the potential for cross-hybridization). In addition, the selected oligonucleotides are sufficiently long to avoid unrelated binding (cfr. short oligonucleotides, Section 3.2.1). All oligonucleotides are also designed to have the same length, which guarantees more consistent hybridization conditions for every single gene on the array.

A study by Hughes et al. (2001) shows that the ideal length for oligonucleotides to guarantee a satisfactory degree of hybridization specificity and sensitivity is obtained for 60-mers.

Agilent (`http://www.agilent.com`) provides a commercial long oligonucleotide platform, based on oligonucleotides of length 60 base pairs. The slides are printed with Agilent's non-contact industrial inkjet printing process. With this inkjet printing technology, any selected oligonucleotide can be synthesized *in situ*, directly on the glass microarray surface. This inkjet printing technology deposits oligo monomers onto specially prepared glass slides, and builds based on digital sequence files the oligonucleotides base-by-base. Therefore they bypass the need for PCR products. The print tip makes no contact with the slide surface and prevents irregularities due to surface contact, resulting in more spot uniformity. The use of non contact printers is not limited to oligonucleotide arrays (e.g., piezoelectric printers), but it is less frequently used for cDNA microarrays, mainly due to

cost issues and the amount of probe required for printing.

GST arrays - CATMA

The third type of two-channel arrays that we will introduce, is the *Complete Arabidopsis Transcriptome MicroArray* or CATMA array, solely designed for the study of the *Arabidopsis thaliana* plant.

Arabidopsis thaliana is a small plant that is widely used as a model organism in plant biology, as it has important advantages for research. It has a short life cycle: it takes 6 weeks from starting into growth until mature seeds. Its genome is quite small (about 120Mb) and it has 5 chromosomes. Detailed information on *Arabidopsis* can be found at The *Arabidopsis* Information Resource (TAIR; `www.arabidopsis.org`). The picture is obtained from `http://www.arabidopsis.org/images/arabi_bw1tr.gif`.

The CATMA array is the result of a collaborative project joining the efforts and resources of laboratories in eight European countries. The project (`http://www.catma.org`) started in 2000 and aimed to produce *Gene-specific Sequence Tags* (GSTs) for all known and predicted genes in the genome sequence (Hilson et al. (2004)). First, the *Arabidopsis* genome was reannotated with the EuGène gene prediction software (Schiex et al. (2001)), leading to a set of about 29,600 predicted genes. For these predicted genes, GSTs and their specific PCR primers were designed with Specific Primers and Amplicon Design Software (SPADS; Thareau et al. (2003)), in such a way that the GSTs have a length of 150-500 bp and less than 70% identity with any other part of the *Arabidopsis* genome (Thareau et al. (2003)). In a first round, this resulted in a set of 21,120 *in silico* GSTs. All information on the GSTs is accessible through the CATMA database (Crowe et al. (2003); `http://www.catma.org`; also relayed by other *Arabidopsis* web sites). Since then, this collection has been updated with 3,456 additional *in silico* tags (24,576 in total; Hilson et al. (2004)). These additional GSTs are derived from genes belonging to gene families and therefore its nucleotide sequence is too similar to other family members to design a specific GST. These genes were added

by using less stringent parameters, as for example increasing the identity cutoff of 70%. Furthermore, improvements of the *Arabidopsis* genome annotations and the gene prediction software changed the original gene set. A third round, based on more recent *Arabidopsis* genome annotations, generated by EuGène gene prediction software and augmented with TIGR 5 gene models from The Institute for Genomic Research (TIGR; `http://www.tigr.org/`), led to the design of an additional set of about 5,760 GSTs, and in total a set of about 30,000 GSTs.
These GST sets are excellent probes for the production of spotted arrays. They are not only carefully designed, guaranteeing an improved sensitivity and specificity. Also their PCR primers are designed in such a way that cross contamination is avoided. The GSTs were first amplified from the genomic DNA with oligonucleotide specific primers, but to the $5'$ end of these primers an extension is added (see Figure 3.2). This pair of extension sequences is a combination of one of twenty-four arbitrary PCR primers and one of sixteen arbitrary primers. The allocation of a pair of these primers is based on the coordinates on the 384 well plate (16 rows $\times$ 24 columns). The second amplification round of the complete GST collection is then executed with these 40 $(= 24 + 16)$ primers. In this way, each GST is amplified by one, unique primer pair, corresponding to its place on the well plate, and the cross-contamination problem, mentioned for the spotted cDNA arrays, is avoided.

The GST amplicons as well as the arrays printed with the CATMA GSTs are available from the Nottingham *Arabidopsis* Stock Centre (NASC, `http://nasc.nott.ac.uk/`). The GST amplicons can easily be reamplified and subsets can be selected to print dedicated arrays. Furthermore, the GSTs can be cloned and used for other functional studies (Hilson et al. (2004)). Because its probes are designed from the complete genome sequence rather than selected from available cDNA or EST collections, it minimizes homologies between probes, and maximizes the genome coverage. Therefore, it definitely is an ideal low-cost option for in-house spotting. In Chapter 4, we will assess whether it can stand the comparison with the commercial, oligonucleotide-based arrays, by benchmarking its sensitivity, specificity, and coverage in an experiment designed specifically for this purpose.

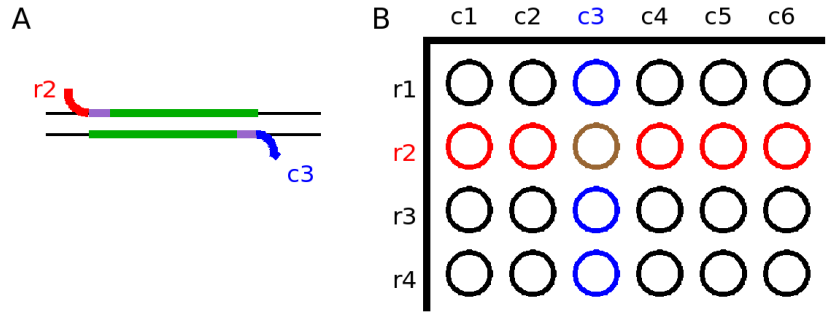
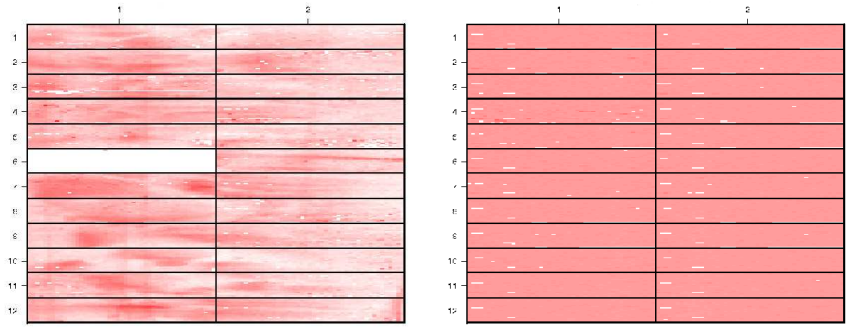


Figure 3.2: **The first amplification round of the GSTs.** The GSTs are first amplified with GST specific primers, each extended at the $5'$ with a specific primer pair. This primer pair is chosen according to the location of the GST on the 384 well plate. Forty arbitrary primers are chosen, one for each column (c1-c24) and one for each row (r1 - r16). For example, the GST in the second row and the third column will have the primer pair (r2, c3). This figure is based on Figure 3 in Hilson et al. (2004).

3.1.2 Quality assessments and normalization of two-channel arrays

The scanner provides a number of statistics for each spot. In all analyses of two-channel arrays, we will make use of the mean foreground and the median background intensities of the Cy3 and Cy5 channel (Yang et al. (2000)). These intensities will be abbreviated as Rf and Rb for the Cy5 foreground and background intensity, respectively, and, analogously, Gf and Gb for the Cy3 foreground and background intensity. We will also extract the local standard deviations of these statistics. The last feature that we will use are the flags, which are scanner and user dependent. The user can flag spots of poor quality (e.g., spots that are absent, too small, uneven hybridization (such as spots with a doughnut shape)). Some scanners also flag for *saturated* spots (i.e., the spot pixel intensity values exceed the detection range and their measured intensities reached a plateau).
The spots on an array are also grouped in bigger blocks, which we will call *grids*. For spotted arrays, as cDNA arrays and the CATMA array, each



(a) (b)

Figure 3.3: **Image of the Cy5 background intensities.** The $\log_2$ Cy5 background intensities are plotted for two slides, in order to visualize spatial effects. (a) One of the print tips shows a defect. (b) The image of the background intensities shows little irregularity and this seems to be a nice hybridization. The images were created with the function `maImage`, which is part of the `marray` package (Bioconductor). The data shown in the figure is part of the data produced within the CAGE project (Chapter 5).

grid is spotted with a specific print tip. Hence there can exist differences between the grids, caused by the print tip. We will call this the *print-tip effects*.

Images of the intensities

A color image of the values of a statistic for each spot on the slide can show spatial effects in the data and expose entire regions of poor quality. For example, by plotting the background intensities, one can visualize dust particles or fingerprints. For spotted arrays, if one of the print tips works abnormally or the cover slip is badly removed, this also becomes visible. An example is shown in Figure 3.3.

Signals significantly different from background

Another measure to assess the hybridization sensitivity is by computing the percentage of probes on an array that report a hybridization signal. There are several possible rules to call a spot *above the background hybridization level*. One can compare the Cy3 and Cy5 foreground intensities (Fg) with the local background measurements (Bg). Often simple rules are applied, such as for example $\mathrm{Fg} > 1.5\mathrm{Bg}$. In this work, we wanted to include the standard deviation of the intensities into the decision rule. If not mentioned otherwise, we have adopted the following decision rule. A signal was considered *above background* if it fulfilled the following criterion for both channels:

$$\mathrm{Fg} > \mathrm{Bg} + 2\mathrm{sd}(\mathrm{Bg}), \tag{3.1}$$

i.e., a signal or foreground intensity (Fg) is called *significant* if it is larger than the background intensity (Bg) plus two times the standard deviation of background. Under the assumption that the intensities measured for all pixels follow a normal distribution, this formula can be interpreted as requiring that the foreground intensity lies outside the 97.5% confidence interval of the background intensities. The ratio

$$\frac{\mathrm{Fg} - \mathrm{Bg}}{\mathrm{sd}(\mathrm{Bg})}$$

is often called the *signal-to-noise ratio* and therefore this rule puts a threshold on the signal-to-noise ratio. At one point we will use a more stringent rule, in which also the standard deviation of the foreground is taken into account

$$\mathrm{Fg} > \mathrm{Bg} + 2\sqrt{\frac{\mathrm{var}(\mathrm{Bg})}{2} + \frac{\mathrm{var}(\mathrm{Fg})}{2}}, \tag{3.2}$$

i.e., a signal or foreground intensity is called significant if it is larger than the background intensity plus two times the standard deviation of background and foreground, computed as the square root of the average of their variances. If the microarray contains *negative controls* (spots that contain DNA that should not hybridize), one can also compare the foreground intensities of the spots with those of the negative controls, instead of the local background.

Correlation plots

Another diagnostic plot is a visualization of the correlations between the different samples in an experiment. This can help to detect anomalies in the samples, as, for example, two identical hybridizations in one experiment or two swapped hybridizations. One can also visualize the non-biological effects as different array batches used within one experiment, at what time the hybridizations were done, or who performed the different hybridizations. An example is shown in Figure 3.4.

Background subtraction and log-transformation

The measured Cy3 and Cy5 foreground intensities include also a contribution that is not due to the hybridization of the sample to the probe, and therefore, it is common practice to subtract for each spot an estimate of the background intensity from the foreground intensity. However, there is still a lot of discussion about whether background subtraction is advisable or not. In this work, we will apply a background subtraction, unless it is explicitly mentioned differently. Different methods exist to subtract this background (Yang et al. (2000)). Here, *local background subtraction* will be applied, so the background will be estimated locally from the area close around the spot.
We will typically work with the base 2 log transformed data. This transformation evens out strongly skewed data and makes the data more normally distributed. This also reduces the dependency of the variation of the intensities on the magnitude of the intensities. Therefore, we will work with $\log_2$ intensities, defined as

$$G = \log_2(\mathrm{Gf} - \mathrm{Gb}) \text{ and } R = \log_2(\mathrm{Rf} - \mathrm{Rb}). \tag{3.3}$$

Loess Normalization

If one considers a *self-self experiment* (i.e., two identical samples are hybridized on one single slide) and one plots the $\log_2$ Cy5 and $\log_2$ Cy3 intensity values (R and G, respectively) versus each other, one expects them to lie along the diagonal. But, as shown in Figure 3.5(a), this is not the case. This effect is caused in essence by the effect of the two different dyes that

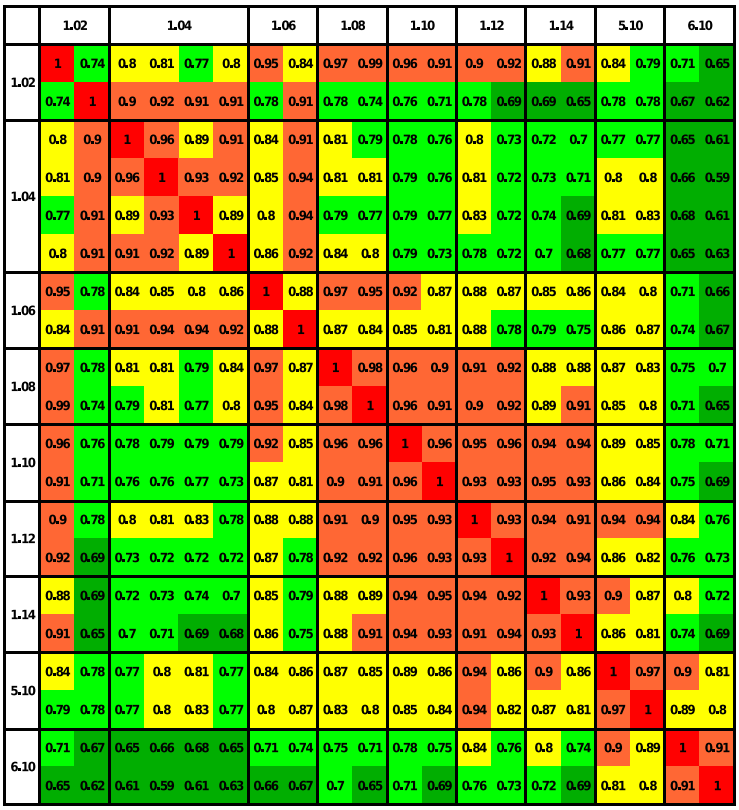


Figure 3.4: **Visualization of the correlation between the samples.** This panel shows the correlation between the samples of a time course experiment. The experiment compares leaf samples of *Arabidopsis Columbia* at different developmental stages (Table 5.2); each time hybridized against a common reference (see Section 5.2.1). Correlation was computed between the $\log_2$ ratios of the measured intensities above background (according to Equation 3.1). Correlations are colored from red (high correlation) to dark green (low correlation (i.e., below 0.70)). From the graph it is clear that the sample at developmental stage 1.06 correlates better with samples at 1.02 and 1.04 than with the replicate sample at stage 1.06. Similarly, one of the samples at stage 1.02 correlates better with the samples at 1.06 and 1.08 than with its replicates. This was indicative of a possible swap of two samples.

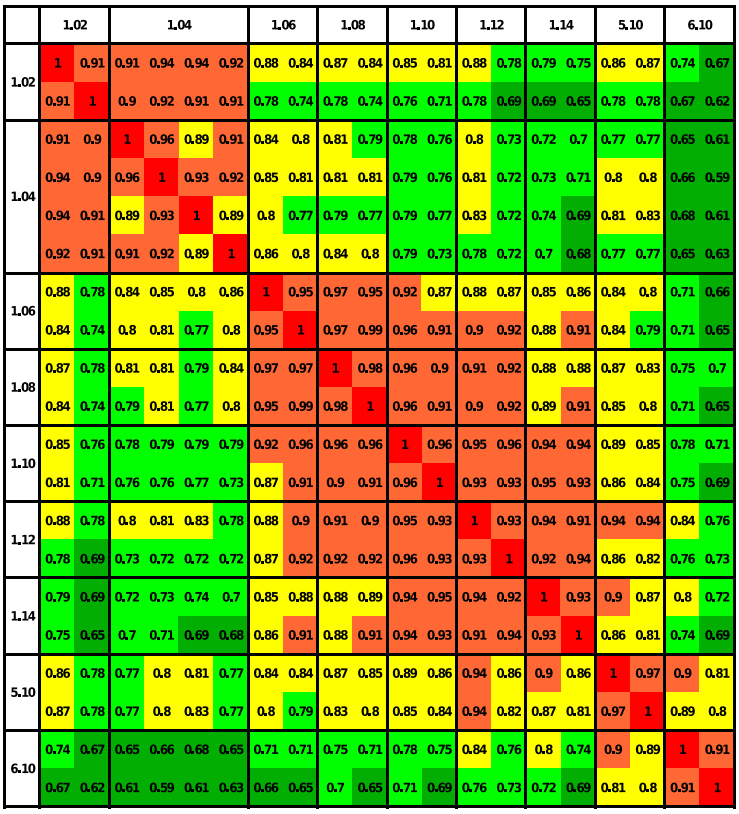


Figure 3.4: *(continued)* **Visualization of the correlation between the samples.** If the two possible erroneous samples are exchanged, the correlation plot look as is expected from such experiment. Later on this mistake was confirmed by the lab where the experiment was done. This data is part of the data produced within the CAGE project (Chapter 5).

were used, because, biologically, there is no difference between the samples. Typically, red intensities (Cy5 - channel) tend to be lower than the green (Cy3) intensities. Instead of plotting intensities of the two-channels versus each other, this plot is typically rotated clockwise over 45° and then,

after scaling, the log ratios are plotted versus the mean intensities, namely

$$M = \log_2 \left(\frac{\text{Rf} - \text{Rb}}{\text{Gf} - \text{Gb}} \right) = R - G \tag{3.4}$$

versus

$$A = \log_2 \sqrt{(\text{Rf} - \text{Rb})(\text{Gf} - \text{Gb})} = \frac{1}{2}(R + G), \tag{3.5}$$

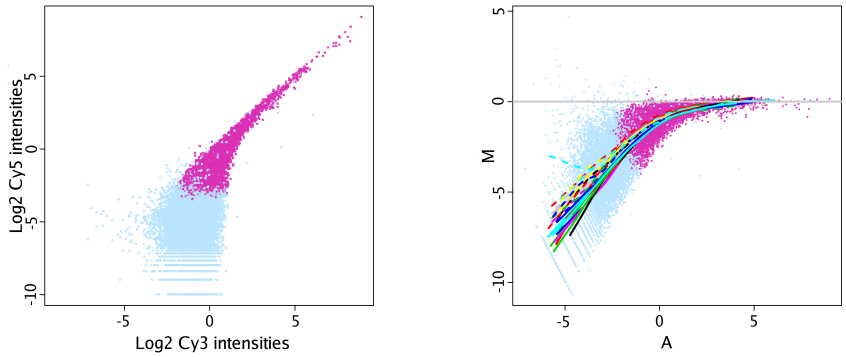
with R and G as defined in 3.3. This plot of M (Minus) versus A (Add) is generally called an *MA-plot* (see Figure 3.5(b)). This MA-plot also shows the dye effects, as one expects that the plot is centered around $M = 0$, which is clearly not the case. Based on these M and A values, we will normalize the data for this dye effect by fitting a robust locally weighted regression, or *Loess* regression, for M based on the values of A (Cleveland (1979)). Loess is a local fitting method, that fits at a value A_i a value $\hat{M}_i$ that is based on the data points in the neighborhood of A_i. In our particular case, this neighborhood will include a portion of 40% of the data points. The choice of this portion has an influence on the degree of smoothness. Loess is weighted in the sense that it weights the different data points according to their distance from the point A_i. The weight function at a point x within the neighborhood of A_i is defined as

$$w(x) = \left(1 - \left| \frac{x - A_i}{d(A_i)} \right|^3 \right)^3,$$

where $d(A_i)$ equals the maximal distance within the span. One can interpret this as a weighted regression, based on the points within a sliding window as shown in Figure 3.6.
To remove the print-tip effects, one can also split up the data into groups, printed by the same print tip. By fitting separate Loess lines for each group and by correcting the intensity by its corresponding Loess lines, not only the dye effect will be removed, but data is then also corrected for the print-tip effect. Of course, this can only work if the array is sufficiently large.

Dye swap

Another additional normalization method to remove the dye effects, is by performing a *dye swap*. This technical replication of the experiment consists of duplicating the labeling and hybridization step, but with the dyes



(a) Cy3 intensity (G) versus Cy5 intensity (R)

(b) MA-plot

Figure 3.5: **MA-plot.** These figures show data of a self-self experiment (i.e., in both channels, the same sample is hybridized) and in this particular case it is a whole plant sample of *Arabidopsis thaliana* at developmental stage 1.04 (Table 5.2). This data is part of the data produced within the CAGE project (Chapter 5). (a) This panel displays a scatter plot of the background subtracted $\log_2$ intensities of the Cy5 and Cy3 channel (i.e., R and G, respectively). (b) In this panel, the plot has been rotated over $45°$ and becomes after scaling a typical MA-plot, as described in Equations 3.4 and 3.5. In both plots, light blue spots correspond to spots below background, while violet spots are above background (according to Equation 3.1). Through this data point cloud, Loess lines are drawn for all data, grouped per print tip. Correction according to the different Loess lines will provide us with the Loess normalized intensities.

swapped. Suppose for example that a test sample is compared with a reference sample, then the first hybridization measures for example the test sample in Cy5, while the reference sample in Cy3. For the second hybridization, the test sample is then labeled in Cy3, while the reference sample in Cy5. The log-ratio of the test versus the reference sample is

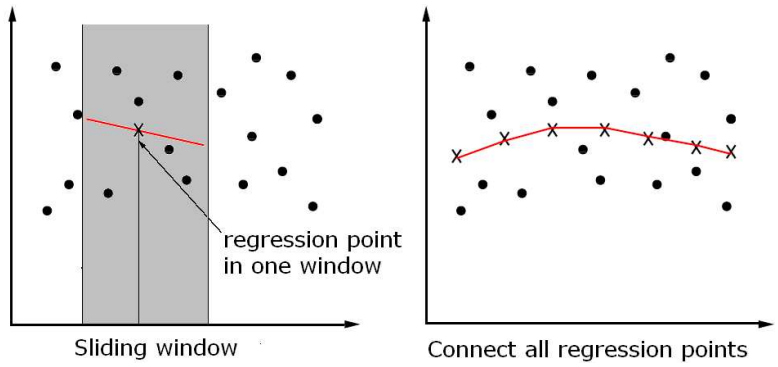


Figure 3.6: **Loess regression.** At each point an estimate is obtained from the points within a window centered around that point, obtained by linear or quadratic, weighted regression. Afterwards all estimates are connected and a loess regression line is obtained.

then computed as

$$\frac{1}{2}\left[\log_2\left(\frac{R^1_{\text{test}}}{G^1_{\text{reference}}}\right) - \log_2\left(\frac{R^2_{\text{reference}}}{G^2_{\text{test}}}\right)\right],$$

where R and G are the background corrected intensities and the superscript indicates the hybridization.

3.2 Single-channel arrays

A second class of microarrays are the short oligonucleotide arrays. Short oligonucleotide arrays are single-channel arrays (i.e., they use only one dye). Therefore, there will be no issue of dye effects.
Affymetrix (`http://www.affymetrix.com`) is probably the most used short oligonucleotide platform and with its specific probe design, it requires a completely different normalization strategy.

3.2.1 Short oligonucleotides: Affymetrix

These short oligonucleotides are synthesized on the slide by using a set of masks. By covering the slide with a mask, a selection of positions on

the chip is exposed and light will then be used to activate these unprotected sites. At these activated positions, nucleotides will bind, resulting in the synthesis of one nucleotide at the positions chosen with the mask. A new mask is then applied and the same process is repeated. After several rounds, the desired set of probes is obtained. In this way a lot of different masks are required and this makes the Affymetrix chip a rather expensive chip.
On an Affymetrix chip, a gene is not represented anymore by one single DNA strand, but by a *probe set*, consisting of 11 to 20 probe pairs. Each *probe pair* is composed of two short oligonucleotides of length 25bp. One matches with a part of the sequence of the gene and is called the *perfect match* (PM). The second oligonucleotide has the same sequence as the perfect match, except for a single mismatch in the middle of the oligonucleotide (at the 13^{th} position), and is therefore called the *mismatch* (MM) probe. These mismatch probes are assumed to measure the nonspecific binding and are therefore often used as a kind of background correction. Disadvantage of this setup is that these oligonucleotides are so short that they are sometimes not gene specific.
The sample is prepared by first obtaining double-stranded cDNA from the mRNA sample via reverse transcription (Section 2.2.1). From this cDNA, cRNA is synthesized via *in vitro transcription* (IVT) (i.e., transcription within a laboratory mixture that contains all the necessary components). The nucleotides used to perform this transcription are biotinylated, so that the cRNA is labeled. Afterwards the cRNA is fragmented into smaller pieces and hybridized on the array (see Figure 3.7).

3.2.2 Normalization of Affymetrix chips

The probe design and the fact that they are single-channel arrays require of course a completely different preprocessing. In this work, both MicroArray Suite 5.0 (MAS 5.0) and Robust Multi-array Average (RMA) will be applied to normalize the Affymetrix data. At the moment, these two are probably the most popular normalization methods. There are a number of alternatives as, for example, Model Based Expression Index (MBEI, Li and Wong (2001)), GCRMA (Wu and Irizarry (2005)), but we refer to literature for discussion of those methods and restrict ourselves to MAS 5.0 and

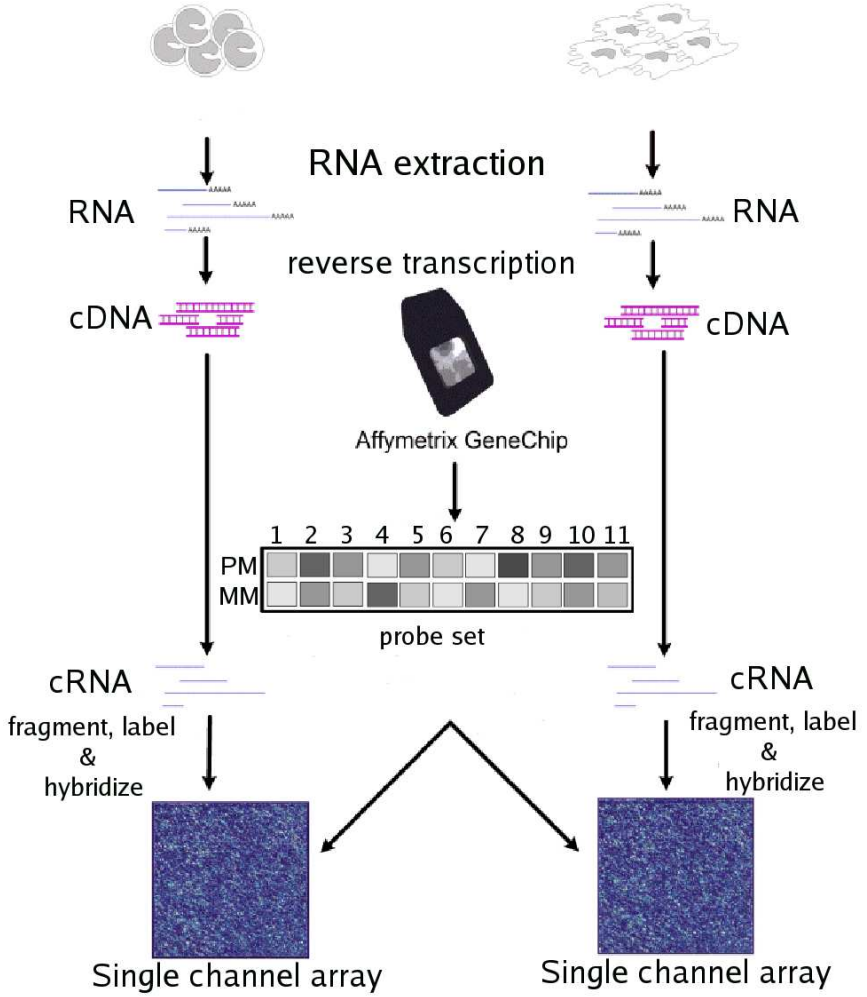


Figure 3.7: **Single channel microarray experiments - Affymetrix:** For two different cell types, the RNA is extracted and cRNA is obtained from the mRNA, via RT and IVT. The two samples are each hybridized on a separate Affymetrix chip.

RMA.

Normalization can be described in a few steps:

1. Background subtraction.
2. Computing the expression value: The set of intensities of the different probe pairs $\{(PM_{ij}, MM_{ij})|j = 1, \ldots, n_i\}$ have to be combined into one single *expression value* x_i for each gene i.
3. Normalization step.

These steps will be described for both methods, MAS 5.0 and RMA.

MicroArray Suite 5.0

Background subtraction The MAS 5.0 algorithm applies a background correction. Therefore it breaks the array up into, by default, 16 rectangular zones and it computes for each region the 2nd percentile signal of the cells. This is then the background for that zone. The local background $b(x, y)$ for a cell (x, y) is defined as an average background, weighted according to the distance between its array coordinates and the centers of the 16 different zones. One also computes a local noise value $n(x, y)$ as the standard deviation of the lowest 2% cell intensities. And the adjusted intensity, noted as $a(x, y)$, is then defined as

$$a(x, y) = \max(I(x, y) - b(x, y), 0.5n(x, y))^1,$$

where $I(x, y)$ is the maximum of the intensity at (x, y) and 0.5. For simplicity, we will denote the set of background corrected MM and PM values also as MM and PM, respectively.

Computing the expression value In this step, we will combine the set of background-adjusted intensities of the probe pairs $\{(PM_{ij}, MM_{ij})|j = 1, \ldots, n_i\}$ into one single expression value x_i for each probe set i.
Until 2001, Affymetrix used *AvDiff* as a default to compute the expression values. This AvDiff approach was included in the previous version,

[1]The constants, given in the equations are the default values as applied in MAS 5.0, but they can be changed.

MAS 4.0. AvDiff computes the expression measurement as an average over all PM_{ij} values, corrected by the mismatch value MM_{ij}:

$$x_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \Delta_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} (PM_{ij} - MM_{ij}) .$$

To make the average more robust towards outliers, the algorithm computed the average for a subset, consisting of those differences Δ_{ij} within 3 standard deviations from the mean of $\Delta_{(2)}, \ldots, \Delta_{(n_i-1)}$, where $\Delta_{(k)}$ is the k^{th} smallest difference.
However, for about one third of the measurements, the MM value is actually higher than the PM value (Irizarry et al. (2003)) and, hence, the expression value can become a negative number with this computation method. Also, a linear scale is not optimal as the variance then heavily depends on the intensity. MAS 5.0 (Affymetrix (2001)) attempts to bypass these shortcomings. First, the MM-values are replaced by an *Ideal Mismatch* (IM) value, which equals the MM value if $PM > MM$, but has been corrected to be smaller than PM if $PM \leq MM$. Therefore a kind of robust average log-ratio, called the *biweight specific background* (SB), is computed as

$$SB_i = T_{\text{bi}} \left(\{ \log_2(PM_{ij}) - \log_2(MM_{ij}) | j = 1, \ldots, n_i \} \right) , \qquad (3.6)$$

where T_{bi} denotes Tukey's Biweight average.
Tukey's biweight average computes a robust average that is unaffected by outliers. Suppose, in general, one wants to compute this robust average for a set of points $a = \{a_k | k = 1, \ldots, n\}$. Hence in our specific case, a equals $\{\log_2(PM_{ij}) - \log_2(MM_{ij}) | j = 1, \ldots, n_i\}$. First, you have to compute the median M of the values a_k and for each data point a_k, the absolute distance from this median M is computed. From these absolute deviations the median S is calculated, which gives an indication of the spread of a_k around their median. With these statistics, a distance[2] is computed for each data value as

$$u_k = \frac{a_k - M}{5S + 0.0001}.$$

[2]The constants, given in the equations are the default values as applied in MAS 5.0, but they can be changed.

From these distances, the weights are computed as

$$w(u_k) = \begin{cases} \left(1 - u_k^2\right)^2 & \text{for } |u_k| \leq 1 \\ 0 & \text{for } |u_k| > 1. \end{cases} \tag{3.7}$$

The final biweight estimate is then computed as a weighted average

$$T_{\text{bi}}(\{a_k | k = 1, \ldots, n\}) = \frac{1}{\sum_k w(u_k)} \sum_k w(u_k) a_k. \tag{3.8}$$

Formula 3.6 computes a robust average log-ratio of PM versus MM and if this ratio is sufficiently large, most values PM_{ij} are larger than MM_{ij} and one can use this average SB_i to compute the ideal mismatch value for the few spots that have $MM_{ij} \geq PM_{ij}$. Hence the Ideal Mismatch is computed based on the probe set. In case SB_i is also small, the Ideal Mismatch value converges to the Perfect Match value. More precisely, the Ideal Mismatch value is computed as

$$IM_{ij} = \begin{cases} MM_{ij} & \text{if } MM_{ij} < PM_{ij}, \\ \frac{PM_{ij}}{2^{SB_i}} & \text{for } MM_{ij} \geq PM_{ij} \text{ and } SB_i > 0.03, \\ \frac{PM_{ij}}{2^{\left(\frac{0.03}{1+\frac{0.03-SB_i}{10}}\right)}} & \text{for } MM_{ij} \geq PM_{ij} \text{ and } SB_i \leq 0.03. \end{cases} \tag{3.9}$$

With this Ideal Mismatch value, the log expression measurement for each gene i is then computed as the robust average (i.e., Tukey's biweight estimate) of the log-transformed PM values corrected with the IM values

$$x_i = T_{\text{bi}}\left(\{\log_2\left(PM_{ij} - IM_{ij}\right) | j = 1, \ldots, n_i\}\right). \tag{3.10}$$

Normalization. The expression measurements are normalized by scaling the expression measurements for each chip to an equal mean. Therefore a *Trimmed Mean* is computed as the average of all observations $\{2^{x_i}\}$, after removing the lowest 2% and the upper 2% of the data. To scale the data to a target value Sc, we define the *Scaling Factor* as

$$\text{Sf} = \frac{\text{Sc}}{\text{Trimmed Mean}}. \tag{3.11}$$

In our data sets this target value Sc will equal 100. The reported MAS 5.0 expression measurement is then defined as

$$\text{MAS 5.0 expression value}(i) = \text{Sf} * 2^{x_i}. \tag{3.12}$$

In this work, MAS 5.0 expression values will always refer to the expression values as defined in Equation 3.12. Within MAS 5.0, there is limited flexibility in the normalization procedure and a description of the alternatives can be found in Affymetrix (2002).

Robust Multi-array Average

Robust Multi-array Average (RMA, Irizarry et al. (2003)) is an almost equally popular method to apply background correction, to normalize the data between the slides, and to compute the expression measurements. Contrary to MAS 5.0, this RMA approach does not make use of the MM-values, as the MM-values are likely to measure both non-specific binding and signal. Therefore, they add noise to the data and can lead to a bias in the corrected PM signal. The RMA method is available as part of the `affy` package within Bioconductor.

Background subtraction. The background is estimated globally for each array, by assuming that the PM values for probe set i and probe j can be described as $PM_{ij} = b_{ij} + s_{ij}$ with an exponential signal $s_{ij} \sim \text{Exp}(\alpha)$, a normal background $b_{ij} \sim N(\mu, \sigma^2)$. Given the Perfect Match values and estimates of α, μ, and σ, the signal values can be computed as $E[s_{ij}|PM_{ij}]$. Under the given assumptions, this equals

$$E[s_{ij}|PM_{ij}] = a + \sigma \frac{\phi\left(\frac{a}{\sigma}\right) - \phi\left(\frac{PM_{ij}-a}{\sigma}\right)}{\Phi\left(\frac{a}{\sigma}\right) + \Phi\left(\frac{PM_{ij}-a}{\sigma}\right) - 1},$$

with $a = PM_{ij} - \mu - \sigma^2\alpha$ and ϕ and Φ are the density and the cumulative distribution function of the normal distribution, respectively. The estimates for α, μ, and σ are estimated globally from the data in an *ad hoc* way.

Normalization. RMA forces the background corrected probe intensities to have the same distribution for each array, by applying *quantile normalization*. Therefore, all data is placed in a matrix in such way that each column corresponds to a chip and each row to a gene. Data are sorted for each column in increasing order and the average is taken for each row. These averages replace the original values in all columns. The last step is to unsort the columns to their original order. For example, on all chips the smallest value after quantile normalization will be the average of the original smallest values over all chips.

Computing the expression value. Once the PM-values are background corrected and normalized, a linear model is fitted to their $\log_2$-transformed values for each probe set i:

$$\log_2 (PM_{aj}) = \mu_a + \alpha_j + \epsilon_{aj},$$

where μ_a is the global, overall effect of the probe set on array a, α_j represents the probe effect for the j^{th} probe in the probe set with the constraint that $\sum_j \alpha_j = 0$ and ϵ_{aj} is the residual for the j^{th} probe on array a. Our interest is, of course, the estimates of the log expression levels μ_a for array a. This effect is estimated in a robust way, by using *median polish* (Tukey (1949)). Therefore, the expression values are placed in a matrix; rows correspond for example to the arrays and columns to the probes. The medians of the rows are subtracted from the corresponding rows and next the medians of the columns are subtracted from their columns. These steps are repeated iteratively until all row and column medians are zero. The obtained, residual matrix is then subtracted from the original matrix. These are the fitted values and if you take the average for each row, you obtain for each array an estimate for the probe set.

3.2.3 Quality assessment of Affymetrix chips

Affymetrix provides a number of guidelines to judge whether the hybridizations within an experiment are satisfactory. This quality assessment is typically done after the MAS 5.0 or RMA normalization step.

Visualization methods

RNA-degradation plots RNA-degradation plots help to assess the RNA quality. To this end, the individual probes in a probe set are ordered according to their location relative to the $5'$ end of the targeted RNA molecule. Since RNA degradation typically starts from the $5'$ end of the molecule, we would expect probe intensities to be systematically lower at that end of a probe set when compared to the $3'$ end. On each chip, the PM-values are combined according to their location in the probe set and averaged over the probe sets. These values are then plotted against their location. It is important that these RNA degradation lines are all more or less parallel. An example of an RNA degradation plot is shown in Figure 3.8.

MA plot For two-channel arrays, MA-plots were a useful tool to detect aberrant behavior in the intensities. Affymetrix is a single channel platform, and hence, there is no straightforward way to make an MA-plot. Therefore, MA-plots for Affymetrix chips are defined in alternative ways. The expression measurements are plotted against a synthetic reference array, which is created by taking the probe-wise medians over all hybridizations in the experiment. Alternatively, one makes the MA-plots of each pairwise comparison.

Density plots of the PM values Histograms of the PM values enable to detect arrays with overall higher or lower intensities or can indicate saturation effects. These cause a small "blob" at the high intensities.

Quantitative quality assessment

The main quantitative indications for the quality of the arrays can be summarized in one single graph. An example is shown in Figure 3.9. In this graph, each line (separated by dashed lines) corresponds to one of the chips. The names of the chips are written on the left side.

Detection calls MAS 5.0 computes the Present and Absent calls for all spots, comparable to Section 3.1.2 for the two-channel arrays. This is done

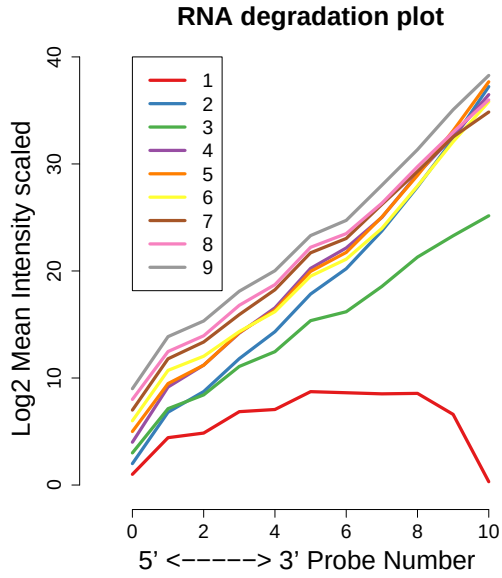


Figure 3.8: **RNA-degradation plot.** This plot shows the RNA-degradation effect for an experiment with 9 Affymetrix chips. The PM values of the 9 hybridizations are base 2 log transformed, averaged by their location in the probe sets over all probe sets. They are scaled to have a standard deviation of 1 to make the trend more visible. They show all a similar increasing trend, except for the first hybridization, which has a strongly deviating behavior. This plot is produced with the `AffyRNAdeg` and `plotAffyRNAdeg` functions of the `affy` package from Bioconductor.

by computing a *discriminating score* for each probe pair (PM_{ij}, MM_{ij}), defined as

$$R_{ij} = \frac{PM_{ij} - MM_{ij}}{PM_{ij} + MM_{ij}}.$$

If an MM_{ij} value becomes small compared to the PM_{ij} value, this ratio R_{ij} will converge to 1. In case the MM_{ij} value is comparable to or larger than the PM_{ij} value, this discriminating score R_{ij} will be small or nega-

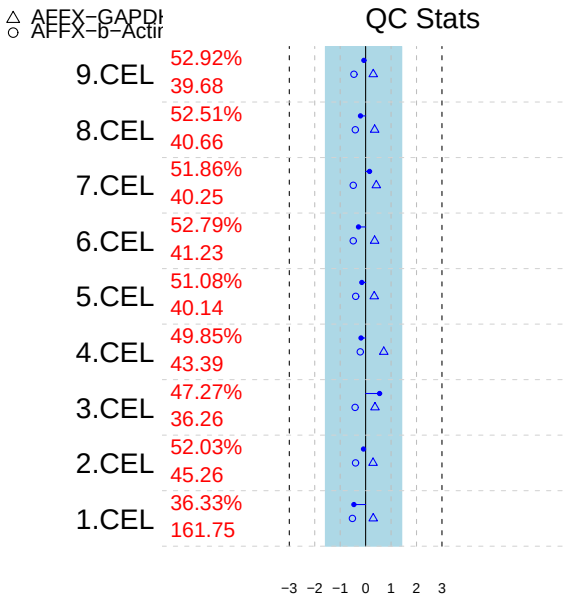


Figure 3.9: **Quality statistics plot.** This plot summarizes the quality assessment statistics for an experiment with 9 Affymetrix chips. A detailed description of the various parameters can be found in Section 3.2.3. In this case both the background and the percentage of Present calls are colored red; this is caused by the fact that the first hybridization has a low percentage of Present calls and a high background. The Scaling factors and the $3'/5'$ ratios for β-actin and GAPDH are within the expected ranges. Combined with the information in Figure 3.8, Hybridization 1 seems to have a deviating behavior. This plot is produced with the function `qc` of the `simpleaffy` package from Bioconductor.

tive. A *detection p-value* p_i for probe i is then computed by testing whether $\text{median}_j(R_{ij})$ equals a small value τ or is larger than τ. The default value for τ is 0.015. This test is done with a one-sided Wilcoxon's signed rank test (Affymetrix (2002)). Depending on the detection p-value p_i, a probe

is called *Present*, *Absent*, or *Marginal* according to the following rule[3]

$$\begin{cases} p_i < \alpha_1 & \text{Present call} \\ \alpha_1 \leq p_i < \alpha_2 & \text{Marginal call} \\ \alpha_2 \leq p_i & \text{Absent call} \end{cases} \tag{3.13}$$

The percentages of Present calls are indicated for each hybridization in Figure 3.9, as the top left number. These numbers should be more or less comparable. In the figure, the percentages are colored red if there is a spread of more than 10% across the whole experiment, which can indicate that at least one of the hybridizations is of inferior quality.

Average background Similar to the percentage of present calls, also the average background should be comparable across the slides. Differences in the background can indicate different amounts of cRNA or differences in hybridization efficiency which resulted in a brighter chip. The background values are colored red if the range of the values is larger than 20 units.

The scaling factor The scaling factor (as defined in Equation 3.11) shows the amount of scaling necessary to bring the mean expression level to an equal mean and therefore, it also reflects the overall expression level on an array. Affymetrix suggests that the scaling factors should be less than three-fold different from the overall average scaling factor. The scaling factors are drawn as a line from the zero-fold line of the image to its scaling factor value and they should fall within the blue strip on the image, which indicates the three-fold of the overall mean scaling factor.

The $3'/5'$ ratios for β-actin and GAPDH To assess the RNA sample quality, one can make use of the genes β-actin and GAPDH. These genes are expressed in most cell types. These are also relatively long genes and most Affymetrix chips contain separate probesets for the $3'$, the middle and the $5'$ region. The ratio of the probe set at the $3'$ to the probe set at the $5'$ end gives an indication of the RNA degradation or inefficient

[3]The default values for α_1 and α_2 are 0.04 and 0.06, respectively.

transcription of the cRNA. Affymetrix states that the $3'/5'$ ratio for β-actin should be lower than 3. As the GAPDH gene is smaller, the cut-off value for GAPDH is set at 1.25. In Figure 3.9, these $3'/5'$ ratios for β-actin and GAPDH are indicated with triangles and circles, respectively. The dashed lines display the values from -3 to 3.

The hybridization controls: BioB, BioC, BioD, and CreX Prior to the hybridization, transcripts derived from completely unrelated material are spiked in. In this way nothing else should hybridize to their probe sets and their intensity is a measure for the hybridization and scanning quality. The transcripts BioB, BioC, BioD, and CreX are added in increasing concentration ($1.5\,pM$[4], $5\,pM$, $25\,pM$, and $100\,pM$). The concentration of BioB corresponds to the lower detection limit, and Affymetrix suggests that it should be called Present in about 50% of the chips. The other three transcripts should always be called Present and their expression measurements should increase.

3.3 Finding differentially expressed genes

In this section, a number of methods to detect differentially expressed genes will be presented. We will not provide a complete overview of all possible methods, but we will restrict ourselves to those techniques that were applied in this work.

3.3.1 LIMMA or Linear Models for MicroArray data

This method represents the design of any microarray experiment in terms of a linear model fitted for each gene separately. To test for differential expression, the gene-specific variance estimates are improved in a Bayesian way by using the information from all genes. The method is implemented in a Bioconductor package called `limma`. We will give here an overview of the method; a more detailed description can be found in Smyth (2004).

[4] pico molar

Describing the experimental design with a linear model

To explain this into more detail, we focus first on the construction of the linear model. Therefore we have to make the distinction between two and single-channel arrays. First, in case of two-channel arrays, the log-ratios $y_g = \log_2(R_g) - \log_2(G_g)$ of the Cy5 intensity R_g over the Cy3 intensity G_g are modeled for each gene g. Consider for example the following loop design

$$\begin{array}{ccc} & C & \\ \swarrow & & \nwarrow \\ A & \longrightarrow & B \end{array}$$

in which two contrasts (e.g., $(B_g - A_g)$ and $(C_g - B_g)$) have to be estimated for each gene g, as the contrast of the third slide can be expressed as a linear combination of these two contrasts. The log-ratios (y_{g1}, y_{g2}, y_{g3}) from the three microarrays can then be used to estimate the contrasts $(B_g - A_g)$ and $(C_g - B_g)$ as

$$E\left[\begin{pmatrix} y_{g1} \\ y_{g2} \\ y_{g3} \end{pmatrix}\right] = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} B_g - A_g \\ C_g - B_g \end{pmatrix}. \tag{3.14}$$

Or, in general, we can write it as a matrix

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$$

representing the design of the experiment and therefore called the *design matrix*, and *coefficient vectors* $\alpha_g = \begin{pmatrix} B_g - A_g & C_g - B_g \end{pmatrix}^T$, containing the effects for gene g.

In the case of single-channel arrays, instead of the log-ratios, the log-intensities are measured. For example, consider an experiment of 5 slides, in which sample A is measured three times and sample B is measured

twice, the design matrix and the effects can then be written as

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } \alpha_g = \begin{pmatrix} A_g \\ B_g \end{pmatrix}.$$

Analogously, all designs can be described with a design matrix X and an effect vector α_g, in such a way that

$$y_g = X\alpha_g + \varepsilon_g,$$

where ε is a noise term. The contrasts or the independent variables will be estimated for each gene from the dependent variables (i.e., the log-ratios) via a linear regression. Further, we assume that the variance of y_g can be written as $\text{var}(y_g) = W_g\sigma_g^2$, with W_g a positive-definite weight matrix.
The model can be fitted with an ordinary least squares fit of the linear model for each gene and this is also what we will apply in this work.

Testing for differential expression

The contrasts of interest can be equal to the estimated effects, but sometimes additional comparisons between the RNA samples are of interest. Therefore we will denote all contrasts of interest as a vector β_g, consisting of linear combinations of the elements in α_g, described with a *contrast matrix* C:

$$\beta_g = C^T\alpha_g.$$

Suppose, for example, we have a loop design as described in 3.14 and our contrasts of interest are not only the effects $B_g - A_g$ and $C_g - B_g$, but also $A_g - C_g$, then we want to test

$$\beta_g = \begin{pmatrix} B_g - A_g \\ C_g - B_g \\ A_g - C_g \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} B_g - A_g \\ C_g - B_g \end{pmatrix} = C^T\alpha_g.$$

By fitting the model to our data, estimates $\hat{\alpha}_g$ for the coefficients α_g, s_g^2 for σ_g^2, and a positive definite matrix V_g, such that $\text{var}(\hat{\alpha}_g) = V_g s_g^2$, are obtained. From these, estimates for all contrasts of interest can be computed, as $\hat{\beta}_g = C^T \hat{\alpha}_g$ and $\text{var}(\hat{\beta}_g) = C^T V_g C s_g^2$. These contrasts are assumed to be approximately normal with mean β_g and covariance matrix $C^T V_g C \sigma_g^2$ and the residual variance s_g^2 is assumed to follow a scaled chi-square distribution

$$s_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2.$$

With an ordinary t-test, one can test whether an estimated contrast β_{gj} for contrast j and gene g equals 0, with the t-test statistic

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}},$$

where v_{gj} is the j^{th} diagonal element of the covariance matrix V_g. LIMMA uses an empirical Bayes t-test in the sense that the standard deviations are based on the data of all genes and not only on the data available for the specific gene. In this way, the test exploits the fact that the model is fitted in parallel for thousands of genes. This t-test is called the *moderated t-test*. A prior estimate for the standard deviation s_0 with d_0 degrees of freedom is estimated, based on the measurements of all genes (Smyth (2004)). This prior updates the gene specific standard deviation and the posterior

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

is used for the moderated t-test with test statistic

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}.$$

This $\tilde{t}_{gj}$ follows a t-distribution with $d_g + d_0$ degrees of freedom, which is narrower and reflects the fact that information has been borrowed from the ensemble of genes to make inferences about each single gene.

3.3.2 General Linear models

A second method to find differentially expressed genes is by fitting general linear models (Kerr et al. (2000), Wolfinger et al. (2001)). Prior to the specification of the models, used to analyze microarray experiments, we give a short description of general linear models.

An introduction to concepts in general linear models

General linear models (GLM) are probably the most used technique in statistics. The technique serves to find the statistical relation between one or more *explanatory* variables and a *response* variable. In this work, all explanatory variables will be treated in a qualitative way (e.g., two growth conditions of plants) and their values can be interpreted as labels that group the data. In some cases the explanatory variable can have a quantitative connotation, but then it is treated as a qualitative variable (e.g., low, medium, versus high temperature). No assumptions about the statistical relation between the factors and the response are made. The explanatory variables are called *factors* and their values are called the factor *levels*. A combination of such factor levels is called a *treatment*.
An important distinction between two types of factors have to be made. They can model a fixed effect or a random effect. A *fixed effect* is an effect for which the levels are chosen for their intrinsic importance. For example, one wants to compare the two growth conditions for plants. For a *random effect*, on the contrary, one wants to model a population of levels and there is no direct interest in the levels that are actually used. The levels are merely chosen to represent the population. For example, if different researchers lead to a difference in the measurements taken in an experiment, then one can chose, for example, five researchers and assess the effect of the different persons working on the experiment. In this case, the interest is not the effect of those five, particular persons, but merely to generalize to the effect of having different persons working on an experiment. If both fixed and random effects are used, we call the model a *mixed model*.
In the text, we will denote fixed factors by Greek letters, while random factors are denoted by Roman letters. Their subscripts will indicate the level. For example, a_i denotes the i^{th} level of a random factor a.
In the following section, a general linear model specific for the analysis of

a microarray experiment, as was proposed in Kerr et al. (2000) and Wolfinger et al. (2001), will be presented.

Microarray experiment in terms of a general linear model

Consider for example a classical dye-swap experiment, in which two samples are compared on two arrays. We can then introduce a factor a_i, that represents the array effect, with two levels $i = \{1, 2\}$ corresponding to Array 1 and 2. Secondly, we have a factor δ_j to measure the Dye effect, again with two levels corresponding to the dyes Cy3 and Cy5. Analogously, the sample effects can be modeled as a factor τ_k with $k = 1, 2$. For each gene g, we will measure log intensities y_{ijkg} for all index combinations $(i, j, k) \in \{(1, 1, 1), (1, 2, 2), (2, 1, 2), (2, 2, 1)\}$. Remark, that such a dye-swap experiment has a Latin square design. A model describing this experiment can be written as follows:

$$y_{ijkg} = \mu + a_i + \delta_j + \tau_k + \gamma_g + (a\gamma)_{ig} + (\tau\gamma)_{kg} + e_{ijkg}, \qquad (3.15)$$

where

y_{ijkg} are the log-intensity of gene g for treatment k, dye j, and array i,

μ is the global mean,

a is the main array effect (random effect; $i = 1, \ldots, \#$ arrays),

δ is the main dye effect (fixed effect; $j = 1, 2$),

τ is the main treatment or sample effect (fixed effect; $k = 1, \ldots, \#$ samples),

γ is the main gene effect for gene (fixed effect; $g = 1, \ldots, \#$ genes),

e is the random error effect.

All random effects a_i, $(a\gamma)_{ig}$, and e_{ijkg} are assumed to be normally distributed with mean 0 and variance σ_a^2, $\sigma_{a\gamma_g}^2$, and σ_e^2, respectively.
The effects a, δ, and τ are used to normalize the data for array, dye and sample effects. The main effect γ accounts for the average effects of the

individual genes and the interaction term $a\gamma$ accounts for the spot effect. All these effects have a normalization purpose and are not of primary interest.
The effect that is of interest is the interaction term $\tau\gamma$ of treatment $\times$ genes. These interaction effects $(\tau\gamma)_{ig}$ account for differences that can not be ascribed to the combination of the main effects of treatment i and gene g. If an interaction effect $(\tau\gamma)_{ig}$ is significantly different from 0, the gene g is differentially expressed. And this is what we will test for to find the list of differentially expressed genes.
Due to the large number of genes, it is computationally intensive to fit this model. Therefore, this model is split up into two parts, a normalization model and a gene-specific model (Wolfinger et al. (2001)). In the first step, the *normalization model*

$$y_{ijkg} = \mu + a_i + \delta_j + \tau_k + r_{ijkg} \tag{3.16}$$

is fitted. The interpretation of the effects is similar to the interpretation as in the global model 3.15. On these residuals r_{ijkg}, a *gene-specific model*

$$r_{ijkg} = \gamma_g + (a\gamma)_{ig} + (\tau\gamma)_{kg} + e_{ijkg}$$

is fitted for each gene. Again, the interpretation of the effects is similar to the effects defined in 3.15, but now they are fitted for each gene separately. Depending on the design of an experiment, this model can change. The idea behind the normalization and gene-specific mixed models was proposed in Wolfinger et al. (2001). Similar models, in which all factors were treated as fixed were proposed in Kerr et al. (2000).

Fitting the mixed model

In the case of a general linear model with only fixed effects, this model can be fitted as in R with the function `aov`. This will provide the least squares solution.
To fit mixed models in R, the library `nlme` and, more specifically, the function `lme` can be used. If we denote the fixed factors as a vector β and the random factors for the i^{th} subject as b_i, then we can model the response y_i for the i^{th} subject as

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

$$b_i \sim N(0, \Psi), \;\; \varepsilon \sim N(0, \sigma^2 I),$$

where X_i and Z_i are the fixed-effects and random-effects design matrix, respectively, and Ψ is a positive-definite variance-covariance matrix. Remark that, in case of a random effect, the intraclass correlations (i.e., correlation between objects that share the same random effect) are not zero, hence independence is no longer assumed.
Suppose, for example, we have a model with one fixed effect β_j with four levels ($j = 1, \ldots, 4$) and one random factor b_i with three levels ($i = 1, 2, 3$), defined as $y_{ij} = \beta_j + b_i + \varepsilon_{ij}$, then this model will be fitted in R, by default, as

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{34} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{34} \end{pmatrix}.$$

Hence, in R the first level of the fixed effects factor is considered as a kind of 'standard' and the remaining three levels of the fixed factor are compared to the first level. Throughout this work, this reference coding will be used. Other choices are possible. For example, in S-PLUS, by default, the *Helmert contrasts* are estimated (i.e., the i^{th} level is contrasted with the average of the preceding levels).

The distribution of b_i is determined by the matrix Ψ. If we denote the parameters in this matrix as a vector θ, then we want to estimate the parameters β, θ, and σ. These parameters can be fitted by using the *maximum likelihood* (ML) estimation. This comes down to maximizing the likelihood function, which is the probability density function of the data y, but regarded as a function of the parameters with the data fixed:

$$L(\beta, \theta, \sigma^2 | y) = p(y | \beta, \theta, \sigma^2).$$

This method, however, has the tendency to underestimate the variance and covariance parameters, and, therefore, it is often the habit to use *restricted maximum likelihood* (REML) estimates, in which marginal likelihoods are

used to estimate the variance components. The restricted maximum likelihood estimation, as defined by Laird and Ware (1982), is an empirical Bayesian approach, in which the variance components are estimated, by assuming a locally uniform prior distribution for β and by integrating these out of the likelihood:

$$L_R(\theta, \sigma^2|y) = \int L(\beta, \theta, \sigma^2|y)d\beta.$$

Maximizing this restricted likelihood provides estimates for θ and σ^2. With these parameters, estimates for β can be obtained via the ordinary ML estimation with known variance components. All details on the computations can be found in Pinheiro and Bates (2000).
Assessing the significance of the fixed effects is done with the Wald F-test. From the model

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$
$$b_i \sim N(0, \Psi), \;\; \varepsilon \sim N(0, \sigma^2 I),$$

follows that

$$y_i \sim N(X_i\beta, Z_i \Psi Z_i' + \sigma^2 I).$$

The hypothesis that

$$H_0 : L\beta = 0 \text{ versus } H_a : L\beta \neq 0$$

is tested with an approximate *Wald F-test* (i.e., a multivariate generalization of the t-test):

$$F = \frac{\left(\hat{\beta} - \beta\right)' L' \left[L \left(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\theta}, \hat{\sigma}) X_i\right) L'\right]^{-1} L \left(\hat{\beta} - \beta\right)}{\text{rank}(L)}$$

where $V_i = Z_i \Psi Z_i' + \sigma^2 I$.

After the introduction of all these concepts and statistical methods, we can start with the actual analysis of the data. As announced in Chapter 1, we will start with the benchmarking of the CATMA array against two commercial platforms, Agilent and Affymetrix.

Chapter 4

Benchmark of CATMA array

The Arabidopsis research community has been blessed with multiple independent resources for transcript profiling, both from commercial sources and academic core facilities. However, today, microarrays that do not carry probes for the majority of transcription units identified in the genome, in particular cDNA arrays, are quickly becoming obsolete. Therefore, this is an opportune moment to introduce the CATMA array (Section 3.1.1) as an alternative to the limited coverage cDNA and commercial, more expensive oligonucleotide arrays. The aim of this work was to describe, in detail, the performance of the CATMA array in comparison with the oligonucleotide-based platforms commercialized by Agilent (Arabidopsis 2 oligo array) and Affymetrix (ATH1 GeneChip probe array; Redman et al. 2004), and to present these results as a reference to the Arabidopsis research community. This work has been published in Allemeersch et al. (2005).

4.1 The CATMA project

The CATMA project (`http://www.catma.org`) was initiated by French and Belgian laboratories in December 1999 and joined by additional groups from Germany (July 2000), the Netherlands, Switzerland, the United Kingdom (October 2000), Spain (January 2001) and Sweden (September 2001). The project aimed to produce *Gene-specific Sequence Tags* (GSTs, Section 3.1.1) for all genes in the genome sequence (Hilson et al. (2004)). Before one can start to search for GSTs, the actual genes in the five *Arabidopsis* chromosomes have to be identified. The choices made for launching the project reflect the status of the knowledge in February 2001 when the structure of only a minority of *Arabidopsis* genes (about 2000) had been determined experimentally. Therefore the project also had to rely on gene prediction to identify the boundaries of each transcription unit and of the exon(s) within it.
At the start of the large scale GST synthesis within the CATMA project, the chromosome annotations published by the Arabidopsis Genome Initiative (AGI) sequencing centers were not homogeneous. Different tools had been adopted by different centers and had evolved over time. According to an evaluation of the gene prediction algorithms used for the annotation of the *Arabidopsis* nuclear genome, the EuGène package, developed by Schiex et al. (2001), offered the most reliable results. Therefore, a complete updated annotation of the *Arabidopsis* genome, provided by EuGène and based on a uniform set of parameters, was originally chosen to design the CATMA GSTs and this resulted in a set of 21,120 in silico GSTs (version 1). Since then, the collection has been updated with 3,456 additional in silico tags (24,576 in total; version 2). These additional GSTs are derived from genes belonging to gene families and therefore its nucleotide sequence is too similar to other family members to design a specific GST. These genes were added by using less stringent parameters, as, for example, increasing the identity cutoff of 70% and by taking into consideration added $3'$ untranslated regions (UTRs). A second group are GSTs for which the PCR amplification failed the first time and for which alternative primers were found. And also improvements of the *Arabidopsis* genome annotations and the gene prediction software changed the original gene set. A third additional GST set has now been created in

the framework of the CAGE project (Section 5.1), based on more recent *Arabidopsis* genome annotations, generated by EuGène gene prediction software and augmented with TIGR 5 gene models from The Institute for Genomic Research (TIGR; `http://www.tigr.org/`), led to the design of an additional set of about 5,760 GSTs, and in total a set of about 30,000 GSTs.
On 21st June 2002 the CATMA database was made public, allowing full searching of the first 21,120 validated GSTs (v1). In March 2004, the database was updated to total of 24,576 GSTs. The update history of the CATMA database and website is described in the CATMA status page (`http://www.catma.org/status.html`).
With these GSTs, the *Complete Arabidopsis Transcriptome MicroArray* or CATMA array was built.

4.2 The CATMA benchmark strategy

Several genome-scale microarrays are now available for *Arabidopsis* transcript profiling and choosing a particular platform will depend on various criteria including genome coverage, data quality, dynamic range, and sensitivity, as well as more practical factors such as availability, price, and logistics. We present here a detailed analysis of the main technical characteristics of the CATMA array, and compared them with the Agilent Arabidopsis 2 oligo array (Agilent array, Section 3.1.1) and the Affymetrix ATH1 genome array (Affymetrix array; Section 3.2.1). Together, these arrays cover the three probe types now used in genome-scale microarrays: PCR amplicons (150 to 500 bp, CATMA), the long oligonucleotides (60mer, Agilent) and the short oligonucleotide sets (25mer, Affymetrix).

For all three platforms, the RNA labeling, hybridization, scanning, and data extraction were performed by a laboratory offering routine microarray services with that particular platform, and following its standard protocols and processes: VIB-MAF microarray facility to process the CATMA arrays, ServiceXS (a service facility in The Netherlands) for Agilent and the Nottingham Arabidopsis Stock Center for Affymetrix. Hence, all datasets were produced independently by laboratories best positioned

to provide service with their particular platform. In all three cases, the platforms were equipped with the standard suite of hardware and software commercially distributed by the Amersham, Agilent, and Affymetrix companies, respectively. The whole analysis is done from the position of a regular customer, that chooses a reliable service provider and trusts the tools, as provided by the microarray companies.

Several studies have already described microarray platform comparison and quality assessment based on various approaches (Chudin et al. (2002); Kuo et al. (2002); Yuen et al. (2002); Lee et al. (2003); Nimgaonkar et al. (2003); Tan et al. (2003)). A common method for platform comparison is to determine the concordance of differential expression measurements between contrasted biological samples. Such studies both pointed to platform-specific expression differences (Kuo et al. (2002); Moreau et al. (2003); Tan et al. (2003)) or illustrated a broad concordance between different platforms (Barczak et al. (2003)).
We have chosen not to focus on gene-for-gene comparison of ratio reports between platforms, but rather on the comparative analysis of RNA samples designed specifically to test the hybridization characteristics of the platforms. Instead we have selected probes that gave no signal in a large number of experiments (data not shown). Prior to labeling, the corresponding RNA of these probes has been added to a single biological sample in a known concentration, i.e. spiked in. We have spiked the biological sample with a range of calibrated quantities, making it feasible to assess different aspects of the platform.

4.3 Coverage of the different platforms

Before we start with the actual assessment of the hybridization qualities of the platform, we will compare the coverage of the platforms (i.e., determine which genes are represented in each of the compared arrays). Therefore, the sequences of their respective DNA features, or probes, were analyzed with BLASTN, an algorithm for comparing biological sequences, against all the transcription units described in the *Arabidopsis* genome annotation provided in January 2004 by The Institute for Genome

Research (TIGR release 5.0).

The total number of array probes or probe sets was 18,981 (CATMA version 1), 22,072 (CATMA version 2), 21,500 (Agilent), and 22,763 (Affymetrix). At the time of the analysis, the CATMA GSTs were produced in two successive rounds and this *in silico* analysis presents both the data on CATMA v1 and CATMA v2. All hybridization data presented below were obtained with arrays printed with the initial version of the repertoire, CATMA version 1. Also, approximately 1,000 of the probe sets on Affymetrix ATH1 arrays permit cross-hybridization to one or more other closely related genes, thus allowing transcript detection of up to 24,000 genes.

The TIGR 5.0 genome annotation contains a total of 26,207 protein-coding genes. In addition, it describes genomic regions with homology to open reading frames of transposable elements[1] (2,355) and pseudogenes[2] (1,652), accounting for an additional 3,786 annotations and 29,993 annotations in total. The coverage is summarized in Table 4.1.
The probe design for all platforms was done with genome annotations predating TIGR 5.0. With the continued refinements in the gene prediction algorithms, some of these gene models have become obsolete. As a result, all platforms contain probes designed according to previous TIGR gene models that do not appear anymore in the latest release.
Of the 22,072 probes on the CATMA version 2 array, 21,019 probes match an AGI code. Because the same gene model is sometimes covered by multiple probes, this set covers a total of 19,910 of the TIGR5 AGI codes (including 575 pseudogenes or transposable elements). It is remarkable that a total of 1,053 *Arabidopsis* sequences have probes only on the CATMA array, and 996 of these are designed to target genes uniquely identified by the EuGène gene prediction software (Schiex et al. (2001)). An additional 57 probes consist of sequences designed to target TIGR3 gene models that do not appear anymore in TIGR5.

[1] DNA sequences that can move around to different positions within the genome of a single cell (i.e., *transposition*). In the process, they can cause mutations and change the amount of DNA in the genome.

[2] genes that once coded for a protein, but that have mutated and that no longer work

	CATMA v1	CATMA v2	Agilent Arabidopsis 2	Affymetrix ATH 1
Probes/probe sets	18,852	22,072	21,500	22,763
Transposable elements plus pseudogenes	363	575	572	946
TIGR 5.0	18,122	21,019	20,921	22,348
On TIGR annotation prior to 5.0	46	57	579	260
EuGène annotation	684	996	-	-
Organelle genomes	-	-	-	155

Table 4.1: **Overview of in silico coverage:** For all three platforms, the number of probes/probesets that match TIGR 5.0 genome annotation is given. As all platforms are designed with genome annotations of earlier TIGR gene models or with different gene models, as EuGène for CATMA, they contain also genes that are not present in the TIGR 5.0 gene models. Affymetrix is the only platform that contains probes for mitochondrial and chloroplast genes.

For Agilent, 20,921 of the 21,500 probes match a gene model with an AGI code. Again, because sometimes more than one probe targets a single gene model, these probes cover 21,090 TIGR5 genes, of which 572 sequences target pseudogenes or transposable elements. An additional 579 probes consist of sequences designed on the basis of the TIGR3 genome annotation, evidently concerning gene models that are absent from TIGR5.
For Affymetrix, 22,348 out of the 22,763 probe sets match a gene model with an AGI code. They cover 23,315 AGI codes of the TIGR5 model, including 946 pseudogenes and transposable elements. An additional set of 260 probes was designed against the TIGR3 models that did not appear anymore in the TIGR5 collection. Affymetrix also has 155 probe sets, targeting genes from mitochondrion and chloroplast. This is the only platform that contains probes for mitochondrial and chloroplast genes.
The coverage of TIGR 5.0 and the overlap of the genes between the different platforms is shown in Figure 4.1.

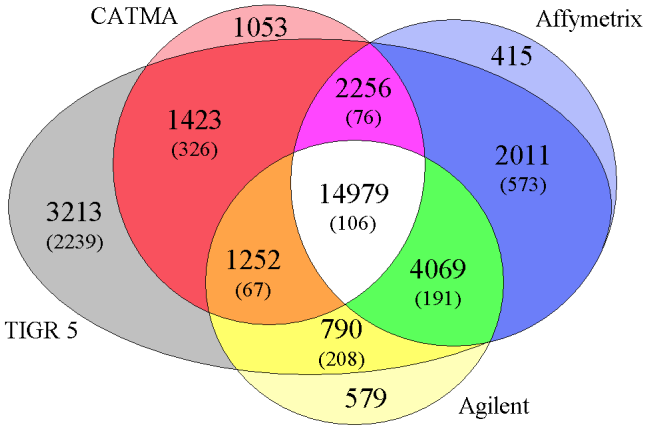


Figure 4.1: **The coverage of the TIGR 5.0 gene models by the different probe repertoires.** Probe numbers of the CATMA version 2 array, the Agilent Arabidopsis 2 oligo array and Affymetrix ATH1 genome array are superimposed on the TIGR5 annotation. Numbers between parentheses refer to non-protein-coding genes, pseudogenes and transposable elements. These three arrays have 14,979 genes in common. A total of 3,213 TIGR5 genes are not covered by any platform. Each platform contains probe sequences that are not supported by TIGR5. Part of these sequences relate to TIGR3 models that are not supported anymore in TIGR5 (57 for CATMA, 597 for Agilent, and 260 for Affymetrix). CATMA in addition contains 996 probes designed for genes predicted by an alternative genefinder EuGène (Schiex et al. (2001)), whereas Affymetrix contains 155 probe sets for organelle genes.

4.4 Design of the experiment

We chose to evaluate the performance of the three array types by performing the same standardized experiment on each of these platforms. The samples were constructed in an artificial way, making it feasible to assess different aspects of the platform. The base for all samples was a single batch of total RNA extracted from *Arabidopsis thaliana Columbia* (*Col*) whole shoots harvested at the developmental stage 1.04 (Boyes et al.

(2001), Table 5.2). To this shoot total RNA sample, *in vitro* synthesized polyadenylated RNA species (from now on referred to as *spike RNAs*) were added in known, well-chosen concentrations. The choice of the genes corresponding to the spike RNAs was not arbitrary. First of all, to be sure that we only measure the added quantities and no unknown signal from the shoot total RNA base sample, the spike RNAs were chosen in such a way that they gave no signal in previous experiments on CATMA and Affymetrix chips. Secondly, they have to be represented on all three arrays. However, a small mistake was made here, as one spike RNA was not represented on the Agilent chip. In this way, fourteen cDNA clones were selected (Table 4.2) and used as templates to synthesize polyadenylated spike RNAs. We assumed that 14 spikes would allow for an in-depth cross platform comparison and 14 is still a number that could practically be handled. As you see in the table, spike 6 was not represented on the Agilent chip.

These fourteen spikes were added in different concentrations to the shoot total RNA sample. Therefore, each spike RNA was mixed in equal amount with one of the other spike RNAs to obtain seven pairs of spike RNAs at equal concentration. They are labeled 'a' through 'g' in Figure 4.2. These seven spike RNA pairs were then combined to construct seven spike mixes in a design similar to a Latin square design. Each mix contained six of the seven spike pairs in staggered concentrations from 0.1 to 10,000 copies per cell (cpc), covering five logs (Table 4.3). As a result, all spike mixes contained equal quantities amounting to approximately 7.4% of the endogenous cellular poly(A)RNA content of in-vitro synthesized poly(A)RNA, assuming that a cell contained on average 300,000 transcripts. Spike 13 was eliminated from further analysis because its quality was insufficient (see Section 4.6).
To convert the spike hybridization signals to ratios an eighth sample was prepared, called the *reference sample*, consisting of the base shoot total RNA, but completed with all spike RNAs now at the same concentration of 100 cpc. Thereby, the comparison of the seven RNA samples to the reference sample should theoretically yield signal ratios ranging from 100-fold to 0.001-fold across the gene subset corresponding to the spike RNAs, and a signal ratio of 1 for all other genes. Hybridization series

Spike Name	At code	ATH1-121501	CATMA GST	Agilent Id
Spike 1	At1g48580	261302_at	CATMA1a39680	AT1G48580.1
Spike 2	At1g52560	262148_at	CATMA1a43590	AT1G52560.1
Spike 3	At3g62230	251252_at	CATMA3a55370	AT3G62230.1
Spike 4	At3g19920	257994_at	CATMA3a19540	AT3G19920.1
Spike 5	At3g17520	258347_at	CATMA3a16945	AT3G17520.1
Spike 6	At1g79760	261345_at	CATMA1a68900	
Spike 7	At5g06760	250648_at	CATMA5a05955	AT5G06760.1
Spike 8	At4g37780	253067_at	CATMA4a39290	AT4G37780.1
Spike 9	At4g37900	253012_at	CATMA4a39410	AT4G37900.1
Spike 10	At1g02360	259443_at	CATMA1a01350	AT1G02360.1
Spike 11	At3g49960	252238_at	CATMA3a43010	AT3G49960.1
Spike 12	At5g22380	249940_at	CATMA5a19840	AT5G22380.1
Spike 13	At5g66230	247134_at	CATMA5a61590	AT5G66230.1
Spike 14	At5g40420	249353_at	CATMA5a36085	AT5G40420.1

Table 4.2: **The 14 cDNA clones selected to generate the spike RNAs.** The number of the spike, as used in the text, is indicated, along with the gene At code, the probe set identifier on the Affymetrix chip, the CATMA GST identifier, and the probe identifier on the Agilent array.

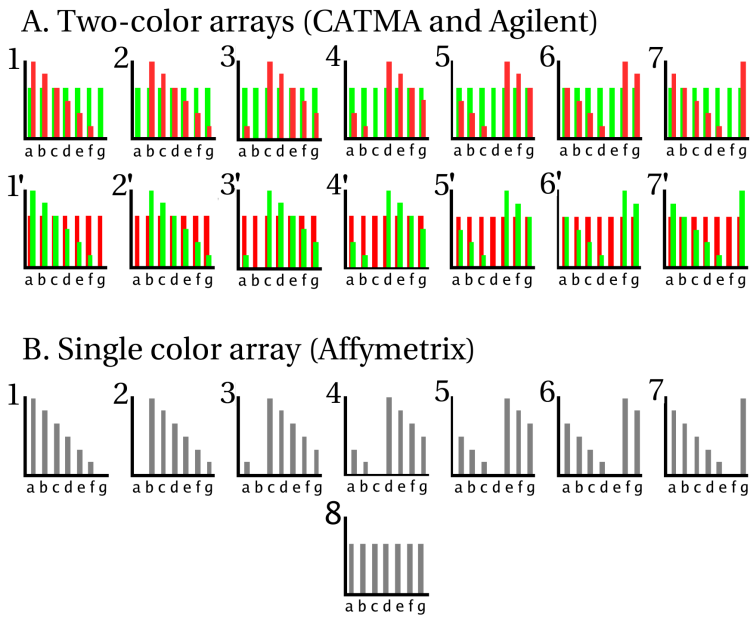


Figure 4.2: **Schematic representation of the experimental design.** Each graph represents the concentrations of the different spike RNA pairs in the RNA sample(s) hybridized to a single array. (A) Two-channel arrays (CATMA and Agilent). In series 1 to 7, the RNA samples containing the spike RNAs in staggered concentration were used as template to synthesize the Cy5 labeled targets, whereas the reference sample was used for the Cy3 labeled target. The dye-swaps were hybridized in the $1'$ to $7'$ series. Cy3 and Cy5 were co-hybridized. (B) Single-channel arrays (Affymetrix). The seven RNA samples containing the spike RNAs in staggered concentrations and the eighth reference sample were each hybridized on a single array.

were set up to perform all possible combinations with the available RNA samples. Therefore, a distinction has to be made between the two-channel arrays and the single-channel arrays. For two-channel arrays (CATMA and Agilent), each individual RNA sample was compared directly to the reference sample, and both dye-swaps were analyzed, resulting in 14 slides for each platform (Figure 4.2A). For the single-channel arrays

Spike number	Spike mix 1	Spike mix 2	Spike mix 3	Spike mix 4	Spike mix 5	Spike mix 6	Spike mix 7	Ref. mix
1, 8 (a)	10,000	0	0.1	1	10	100	1,000	100
2, 9 (b)	1,000	10,000	0	0.1	1	10	100	100
3, 10 (c)	100	1,000	10,000	0	0.1	1	10	100
4, 11 (d)	10	100	1,000	10,000	0	0.1	1	100
5, 12 (e)	1	10	100	1,000	10,000	0	0.1	100
6 (f)	0.1	1	10	100	1,000	10,000	0	100
7, 14 (g)	0	0.1	1	10	100	1,000	10,000	100

Table 4.3: **Concentration (copies per cell) of the 14 spike RNAs for the 7 different spike mixes and the reference mix.** Each spike RNA was calibrated and mixed in equal amount with one of the other spike RNAs to obtain seven pairs at equal concentration (labeled $a - g$).

(Affymetrix), each of the seven spike mixes were hybridized to one slide, and the reference sample was hybridized on an additional eighth slide (Figure 4.2B). Although the total number of hybridizations on Affymetrix arrays was only half that of the two-channel arrays, this fact actually reflects the practical application of the different platforms for a single observation. On two-channel arrays, one usually measures one probe per gene in a dye-swap; while on single-channel arrays a gene is measured as a probeset in a single hybridization.

4.5 Data acquisition and normalization

As the experiment involves different platforms, different normalization methods, specific for each platform, had to be used. The purpose of the study was to benchmark the CATMA array against two commercial platforms and therefore, the analysis was done from the position of a regular customer. Hence the typical, custom normalization methods, accepted as standard by the microarray data community, were applied. The CATMA array data were normalized after subtracting for each feature the median background intensity from the mean foreground intensity. Generally speaking, it is still common practice to include background subtraction in the normalization. However, it is a point of discussion and

for some of the analyses presented in this work, results will be shown for both approaches, with and without background subtraction. After background subtraction, the data were $\log_2$ transformed and normalized using the standard Loess normalization (Section 3.4), for each print tip separately. These Loess normalized $\log_2$ ratios were averaged over the two dye-swaps. The ratios were then computed as the exponential base 2 of that average.

Similarly, the $\log_{10}$ ratios calculated from Agilent array hybridizations as supplied by the service provider were averaged over the two dye-swaps and the final ratio was also expressed as the exponential base 10 (Agilent (2003)).

The raw Affymetrix data were preprocessed with two software packages. As with Agilent, the expression measurements as supplied by the service provider were used. They were computed with Affymetrix Microarray Suite (MAS) 5.0 (Affymetrix (2001); Section 3.2.2). In addition, the data were also normalized by ourselves with RMA (Irizarry et al. (2003); Section 3.2.2). Because the Affymetrix platform does not allow for direct, within chip, comparisons of two samples, ratios were calculated for the seven spike mixes relative to the eighth reference spike mix.

4.6 Dynamic range and sensitivity

Microarray data typically provide information about the level of transcripts relative to a reference. Therefore, it is critical to investigate the dynamic range of the different platforms (i.e., whether they display a linear dose-response relationship between transcript abundance and hybridization signal) and to determine the span of this dynamic range. In our experimental design, the spike RNA concentration range covered all biologically relevant transcript levels and this broad range enabled us to compare the dynamic range of all three platforms in a straightforward manner.
For all three platforms the ratios of the different spike RNAs were extracted as described in Section 4.5. By computing the ratios a mistake was discovered. Despite the quality control of all spike RNAs, the RNA concentration

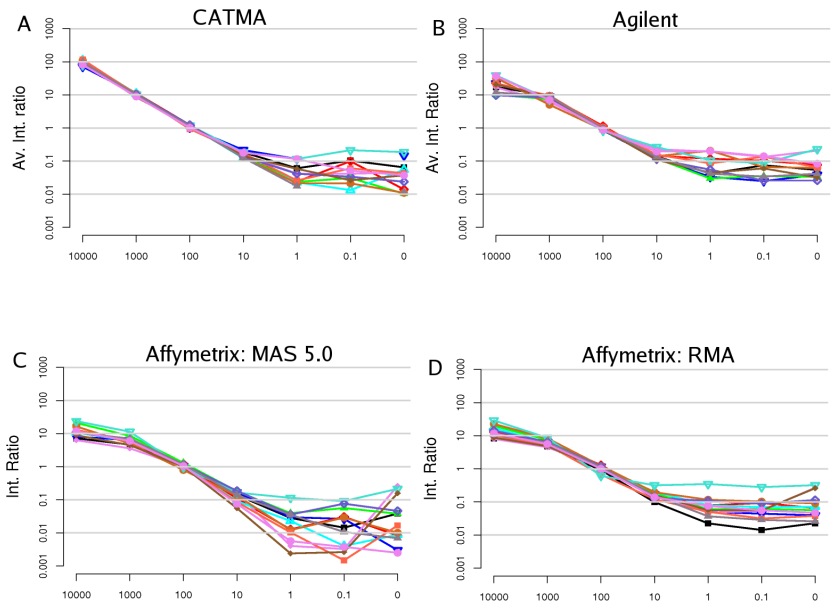


Figure 4.3: **Normalized intensity ratios.** The abscissa indicates the cell copy number equivalent in spike mix 1 to 7. The ordinate shows the resulting ratios relative to the reference mix (all at concentration of 100 cpc) for the different platforms. (A) CATMA. (B) Agilent. (C) Affymetrix with MAS 5.0 preprocessed data. (D) Affymetrix with RMA preprocessed data.

of spike 13 was overestimated and no meaningfull conclusions could be drawn for the analysis of this spike RNA. Therefore it was omitted from all subsequent analyses. The ratio measurements for all remaining 13 spike RNAs on all three platforms are shown in Figure 4.3.
Each panel in the figure is the summary of a complete hybridization series (14 arrays for both CATMA and Agilent; eight arrays for Affymetrix), where each curve represents the signal ratios associated with one of the 13 spike RNAs. The signal ratios are for each spike plotted left to right from the highest to the lowest concentration. In this way, the panels provide

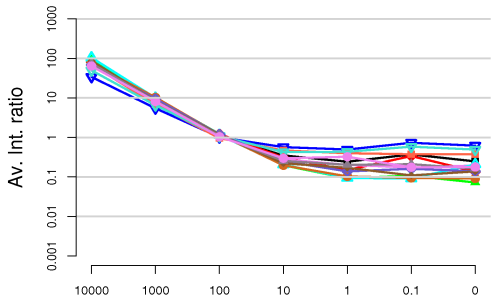


Figure 4.4: **Dose-response curves for the CATMA array.** Dose-response curve from the intensity measurements on CATMA, without background subtraction.

an overview of the hybridization dynamic range for all platforms. In all of them, the ratios calculated for samples at 100 cpc were close to 1, as was expected, because the reference sample contains all spike RNAs at that same concentration. CATMA arrays displayed a near perfect dynamic range over three logs (10,000 to 10 cpc), while Agilent and Affymetrix arrays had a somewhat wider spread of the curves with dynamic range seldom beyond two logs (1,000 to 10 cpc), depending on the spike RNA, and on the preprocessing method for Affymetrix. For CATMA, the dose-response curves were also made in case no background correction was applied. Figure 4.4 suggests that background subtraction gives better results than without background subtraction.

The high concentrations (ratios superior to 1) provide information on saturation effects (see Section 3.1.2), that can arise when the scan settings are so high that some pixels exceed the software limit. Clearly, only the CATMA platform reported accurately ratios for spike RNAs at the highest concentration (10,000 cpc; 100-fold ratio; Figure 4.3A), where both the Affymetrix and Agilent platforms showed a marked collapse (Figure 4.3B, 4.3C, and 4.3D), perhaps due to these saturation effects. Interestingly, the

Agilent data output automatically flags probes that show signal saturation. Out of the 12 probes corresponding to spike RNAs and represented on the Agilent array, 10 of them were flagged as saturated in both channels when hybridized with spikes at 10,000 cpc. None were flagged at lower concentrations. Notably, 27 additional Agilent probes, sharing no homology with the spike RNAs, were also flagged for saturation (see Table 4.4), ranging between 6 and 18 genes for the Cy3 and between 16 and 25 genes for the Cy5 channel. This confirms that such high concentrations of up to 10,000 cpc can also be relevant in biological processes. Most of them represent nuclear genes involved in chloroplast function.

The low concentrations (ratios below 1) inform on the sensitivity of each platform, as it shows how signals of the lower target concentrations get confounded with background noise. Overall, for all three platforms, linearity of the dynamic range ends around 10 cpc and the signal reaches a bottom plateau marking the limit of sensitivity around 1 cpc. Although the position of the plateaus for some spikes may in fact reflect a low level of transcription for the spike RNA cognate genes, they most probably indicate non specific background hybridization because the curves are not ranked in any conserved order across the platforms. Together, these observations suggest that the three platforms have similar sensitivity.

4.7 *In vivo* coverage

The percentage of the probes on an array that report a hybridization signal can also be interpreted as a measure of platform sensitivity. However, the comparative analysis of this parameter across the platforms is difficult because it depends on many factors, including scanner characteristics, data extraction software, and, subject to many different interpretations, the decision rule to declare that a signal is above background hybridization level. Aware of these differences, a summary of the results as they were exported by the particular data extraction software specific to each platform is presented. But due to all these factors mentioned above, these results can yield strikingly distinct results. Only genes transcribed in the base *Col* shoot sample were considered in this analysis, based on the three hybridization

Agilent Identifier	Gene description	Hybridizations with Cy5-saturated signal	Hybridizations with Cy3-saturated signal
AT1G08380.1	expressed protein	1, 2, 3, 4, 5, 6, 7, 1′, 2′, 5′, 6′, 7′	7, 3′, 6′
AT5G38410.1	ribulose bisphosphate carboxylase small chain 3B	All	1, 2, 3, 4, 5, 6, 7, 1′, 3′, 4′, 5′, 6′, 7′
AT1G61520.1	Chlorophyll A-B binding protein / LHCI type III	1, 2, 5, 6, 7, 1′, 2′, 5′, 6′, 7′	1′, 3′
AT1G55670.1	photosystem I reaction center subunit V	1, 2, 3, 4, 5, 6, 7, 1′, 2′, 4′, 5′, 6′, 7′	7, 1′, 3′, 5′, 6′
AT1G79040.1	photosystem II 10 kDa polypeptide	All	1, 2, 3, 4, 5, 6, 7, 1′, 3′, 4′, 5′, 6′, 7′
AT2G06520.1	membrane protein, putative, contains 2 transmembrane domains	7	All
AT1G29910.1	Chlorophyll A-B binding protein 2, chloroplast	All	All
AT1G31330.1	photosystem I reaction center subunit III family protein	All	1, 2, 3, 4, 5, 6, 7, 1′, 3′, 4′, 5′, 6′
AT2G34420.1	Chlorophyll A-B binding protein / LHCII type I	All	1, 2, 3, 4, 5, 6, 7, 1′, 3′, 4′, 5′, 6′, 7′
AT1G67090.1	ribulose bisphosphate carboxylase small chain 1A	All	All
AT5G26110.1	expressed protein	6′	6′
AT2G30570.1	Cytochrome P450 71A12, putative	All	All
AT1G29920.1	chlorophyll A-B binding protein 165/180, chloroplast	All	All
AT2G34430.1	Chlorophyll A-B binding protein / LHCII type I	All	All
AT1G06680.1	photosystem II oxygen-evolving complex 23	All	3,6
AT5G66570.1	oxygen-evolving enhancer protein 1-1, chloroplast	1, 2, 3, 5, 6, 7, 1′, 3′, 4′, 5′, 6′, 7′	1′,3′,5′,6′
AT3G47470.1	Chlorophyll A-B binding protein 4, chloroplast	1, 2, 4, 5, 6, 7, 1′, 2′, 5′, 6′, 7′	1′, 3′, 5′, 6′
AT4G10340.1	chlorophyll A-B binding protein CP26, chloroplast	All	1′, 3′, 4′, 6′
AT1G30380.1	photosystem I reaction center subunit psaK, chloroplast, putative	All	

Table 4.4: **Description of genes, flagged for saturation.** The table (continued on the following page) contains all genes that were flagged at least once for saturation in one of the Agilent data files, omitting the RNA spikes. Along with their Agilent identifier, the list also provides the numbers of the hybridizations, as defined in Figure 4.2, for which the signal was flagged for saturation.

Agilent Identifier	Gene description	Hybridizations with Cy5-saturated signal	Hybridizations with Cy3-saturated signal
AT2G39730.2	ribulose bisphosphate carboxylase/oxygenase activase	4, 5, 6, 7	
AT4G38970.1	fructose-bisphosphate aldolase, putative	1, 2, 4, 5, 7, 1′, 3′, 4′, 5′, 6′, 7′	5′
AT4G02770.1	photosystem I reaction center subunit II, chloroplast, putative	1, 2, 5, 7, 1′, 5′, 6′, 7′	
AT5G54270.1	chlorophyll A-B binding protein / LHCII type III	5′, 7′	
AT3G54890.1	Chlorophyll A-B binding protein / LHCI type I	1, 2, 3, 4, 5, 6, 7, 1′, 2′, 3′, 4′, 6′, 7′	13, 14, 1′,2′,3′, 4′,5′,6′,7′,
AT1G20340.1	Plastocyanin	5, 7, 5′, 7′	
AT1G15820.1	Chlorophyll A-B binding protein, chloroplast	All	1, 2, 4, 5, 6, 7, 1′, 3′, 4′, 5′, 6′, 7′
AT1G51400.1	photosystem II 5 kD protein	5′, 7′	

Table 4.4: ***Continued***

series (Figure 4.2). All spike probes and the various controls were omitted. For CATMA data, a signal was considered *above background* if it fulfilled the criterion as defined in Equation 3.2 for both channels, namely

$$\mathrm{Fg} > \mathrm{Bg} + 2\sqrt{\frac{\mathrm{var(Bg)}}{2} + \frac{\mathrm{var(Fg)}}{2}}. \tag{4.1}$$

The fraction of CATMA probe signals above this threshold ranged between 40.4% and 54.3% (average 50.6%). Separate experiments with leaf and shoot RNAs conducted with CATMA arrays also routinely showed that more than 50% of the probes yielded signal significantly above background according to the same criterion.

For Agilent, the information was extracted from the features, that were provided in the raw data files, *gIsWellAboveBG* and *rIsWellAboveBG* (Agilent (2003)). The vast majority of the probes were labeled with signal above background in both channels: between 93.6% and 99.6% (average 96.9%). Because it is highly unlikely that over 95% of the *Arabidopsis* genes are actually transcribed in *Col* shoots, we investigated the background and foreground values for the control features in the

complete Agilent data set. As expected, an average of 99.1% of the positive controls displayed signal above background, but oddly some 74% of the negative controls were also flagged as such. When we changed the feature extraction mode to *spatial detrending* instead of *background subtraction* (Feature Extraction Software version 7.5) we observed some improvement. With these settings, the percentage of flagged negative controls decreased from 74% to 25.9%, but on average still 91.9% of all *Arabidopsis* probes gave a 'significant' signal. We have not tried other alternative procedures for feature extraction, and we subsequently used the data obtained following standard *background subtraction* for all subsequent analyses presented below. Our observations, however, suggest that the raw data features gIsWellAboveBG and rIsWellAboveBG about signal significance have no absolute biological relevance. Applying the same decision rule, as defined in Equation 3.2, as for the CATMA data set resulted in an even larger percentage of probes with signal above threshold, above 99.85% for all hybridizations. By setting an alternative threshold defined as the median signal of the negative controls plus two standard deviations of the median signals, 63.1% of the *Arabidopsis* probes scored positive.

For Affymetrix data, we also relied on the flags that were provided in the data files (i.e., the number of probe sets labeled as *Present* by the *Detection Call* function in the MAS 5.0 software (Section 3.2.3)). Between 50.5% and 57.0% of all probe sets were assigned present calls (average of 53.9%).

As the *in vivo* coverage estimated for CATMA and Affymetrix is comparable, we made a more detailed comparison of the *in vivo* coverage. Based on the AGI codes, 14,844 genes had matching probes both on the CATMA v1 array and the Affymetrix chips. A gene was called *present*, if the signal was above background, according to the platform specific decision rules, in at least half of the hybridizations. The overlap between the present and absent calls was computed for the set of common genes and the results are shown in the Table 4.5. We observe that 83.7% of the genes detected on CATMA arrays are also detected by Affymetrix, and 79.4% vice versa.

Present on	Number of genes
Both CATMA and Affymetrix	7,167
Only CATMA	1,400
Only Affymetrix	1,855
None of them	4,422
Total	14,844

Table 4.5: **Comparison of *in vivo* coverage.** The numbers indicate genes represented on both CATMA v1 and Affymetrix ATH1 arrays and that yield a significant signal upon hybridization with *Col* shoot target. As discussed, the Agilent platform was not included in this comparison because of the apparent lack of biological significance of the Agilent present calls.

4.8 Specificity

Probe specificity was assessed by looking for cross-hybridization of spike RNAs to probes other than the cognate probes. For each of the 13 spike RNAs we focused on the three highest concentrations (10,000, 1,000, and 100 cpc) among the labeled targets and checked whether cross-hybridization could be detected. Therefore, we blasted the spike RNA sequences against the probe sequences of the three different platforms. We used an E-value of 10.0 as a threshold to declare a hit significant. Due to the short probe lengths of the Affymetrix platform, the corresponding E-values were high and usually only concerned one of the probes of a probe set (although all probes were tested). In the CATMA probe repertoire, some of the spikes gave fewer than three BLAST hits with an E-value below 10.0 (not counting the cognate probe), indicating very low levels of cross-homology. For each platform the normalized intensity ratios of the top three BLAST hits (i.e., provided there are at least three significant matches) were plotted, along with the intensity ratios of the cognate spike probe. None of the graphs gave an indication of the presence of cross-hybridization on any of the platforms. In Figure 4.5 the plots are shown for one of the spike RNAs (spike 1); all other plots are comparable. For spike 1, there was only one significant hit with the CATMA probes, namely CATMA5a02070 (E-value 0.55). The three

top-hits on Agilent were A_84_P22838, A_84_P168433, and A_84_P22774 (all with E-value 1.1) and on Affymetrix we found 252867_at (E-value 0.68), 263330_at (E-value 2.7), and 262647_at (E-value 2.7). In the graph the cellular copy number equivalent in spike mixes 1 to 7 is shown versus the resulting ratios relative to the reference mix, as in Figure 4.3. The cognate dose-response curves are in red; the non-cognate probe curves in black. Clearly, Figure 4.5 shows no consistent correspondence between the cognate probe signal associated with a consecutive series of (high) spike concentrations and the signals of non-cognate probe curves.
None of the three platforms showed evidence of cross-hybridization, since we could not detect hybridization patterns associated with any of these sets of spike RNAs, for any of the spike RNAs tested, in any of the microarray types. This is remarkable considering that a spike RNA is present at 10,000 cpc.

4.9 Signal reproducibility

Because the majority of the labeled target consisted of a single *Col* shoot RNA sample, transcript level measurements should theoretically be invariant across all hybridizations for all genes, except for those corresponding to spike RNAs. Therefore, the different hybridization series essentially consisted of a high number of replicates, eight (on Affymetrix chips) or fourteen repetitions (on the two-channel arrays) (Figure 4.2), that are valuable to assess characteristics of the platforms.
In particular, our data set was used to investigate whether the relationship between signal reproducibility and intensity across the transcript level range depends on the platform. Because the array signal is defined as platform-specific intensity, the $\log_2$ intensity values were first converted to a unique scale by Z-score transformation, so that the signal value distribution had a mean equal to zero and a standard deviation equal to 1 (Tan et al. (2003)). Furthermore, to compare similarly sized datasets, we calculated and plotted the Z-score curves for specific subsets of the data. We took the converted values from the seven Affymetrix hybridizations with RNA samples containing the spike RNAs in staggered concentration (1 to 7 in Figure 4.2B, excluding the reference sample). For two-channel

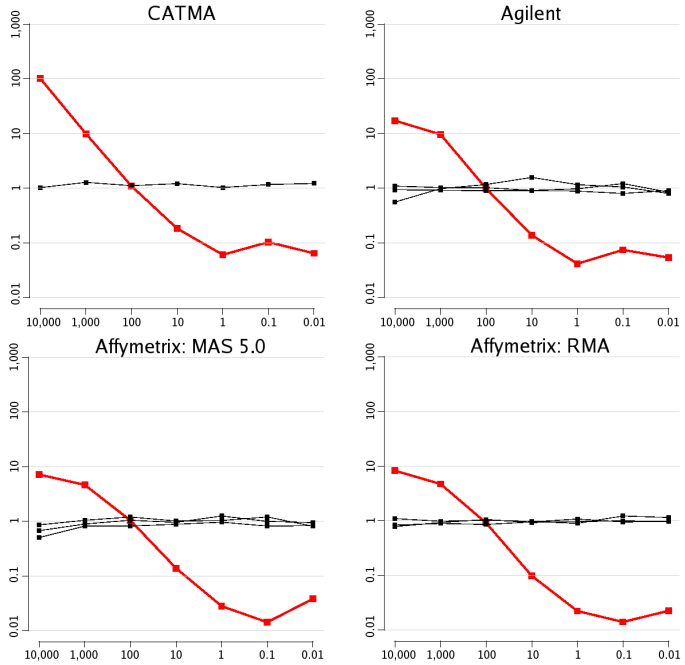


Figure 4.5: **Hybridization signals for highest-ranking homology probes.** Dose response curves for the probe matching each spike mRNA in each system and for the non-cognate probes most closely related to each spike mRNA sequence. In each graph the cell copy number equivalent in spike mix 1 to 7 (x-axis) versus the resulting normalized intensity ratios relative to the reference mix (y-axis) are shown as in Figure 4.3. The matching probe curves are in red; the three non-cognate probe curves in black. (A) CATMA. (B) Agilent. (C) Affymetrix with MAS 5.0 preprocessed data. (D) Affymetrix with RMA preprocessed data.

arrays, we used the seven pair-wise averages of the Cy5 and Cy3 intensities corresponding to the same RNA samples in the reciprocal dye-swaps (Cy5 from 1 to 7 and Cy3 from 1' to 7' in Figure 4.2A, excluding the signals from the reference channels). In doing so, we used a 7-slide data equivalent for all three platforms (two-color datasets typically include a

dye-swap hybridization) and compared the Affymetrix 11-probe set design (that actually measures each transcript 11 times, exporting an average) with the dye-swap design. Furthermore, only the set of 13,036 genes with cognate probes on all three arrays were considered, omitting, however, those matching the spike RNAs.

Figure 4.6 shows the corresponding Z-score frequency plots. Because these plots illustrate the distribution of the normalized data within- and across-platform, they allow a direct comparison of the hybridization characteristics of the different systems. The Z-score distributions of the individual arrays in any given group were all very similar, indicating that hybridizations were highly reproducible. The frequency distributions of CATMA, Agilent, and the Affymetrix RMA values had profiles suggestive of a Gaussian distribution, but sometimes with quite distinct 'shoulders'. For instance, the CATMA data displayed a significant broadening of the peak, and the Affymetrix MAS 5.0 values even showed a distinct bimodal distribution with an additional smaller peak at lower intensity. Affymetrix data analyzed with RMA had a Z-score distribution very similar to the distribution of CATMA data. The difference between MAS 5.0 and RMA indicates that at least part of the bimodality of the distributions resulted from data preprocessing.

To visualize the signal reproducibility in function of intensity, we plotted the Z-score standard deviation against the Z-score mean for each gene (Figure 4.7). The Loess lines representing the overall trend for each system are shown collectively in Figure 4.8.
CATMA values for background-subtracted data (CATMA BGS) showed variability independent of signal for high to medium intensity, but gradually increasing for low signal. In contrast, CATMA non-background subtracted data (CATMA non-BGS) resulted in a flatter Loess showing a somewhat decreased variability at low intensity. Agilent had overall higher variability increasing at both ends of the intensity spectrum. We presume that the variability at low intensity results from background subtraction, whereas higher intensity values may reflect saturation. Finally, MAS 5.0 variability was low for high to medium signal but with a sharp increase followed by a conspicuous drop for the lower intensity values. This profile

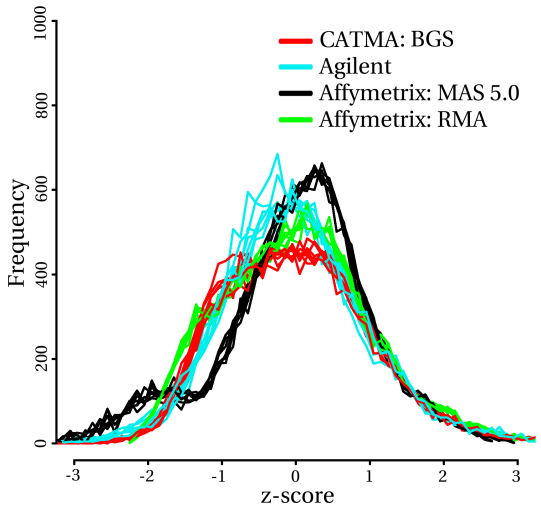


Figure 4.6: **Distribution of the Z-scores.** For each platform seven individual density functions are shown, each representing one particular hybridization.

was strikingly different for RMA-processed Affymetrix data, where the variability was overall very low and independent of intensity. This behavior is consistent with the statistical strategy behind RMA, which aims at reducing signal variance.

The same signal intensities (used to compute the Z-scores) were also used to assess the correlation between intensity values across platforms. We restricted the analysis to the genes that were present on all three platforms, and that displayed a significant signal indicative of detected expression, on both the CATMA and Affymetrix arrays (i.e., the gene was detected above the platform-specific threshold in at least four out of the eight hybridizations with Affymetrix chips and seven out of 14 hybridizations on CATMA arrays (similar to Section 4.7)). Both restrictions led to a subset of 6,473 common genes. The pairwise scatter plots of the $\log_2$ intensities are shown in Figure 4.9.

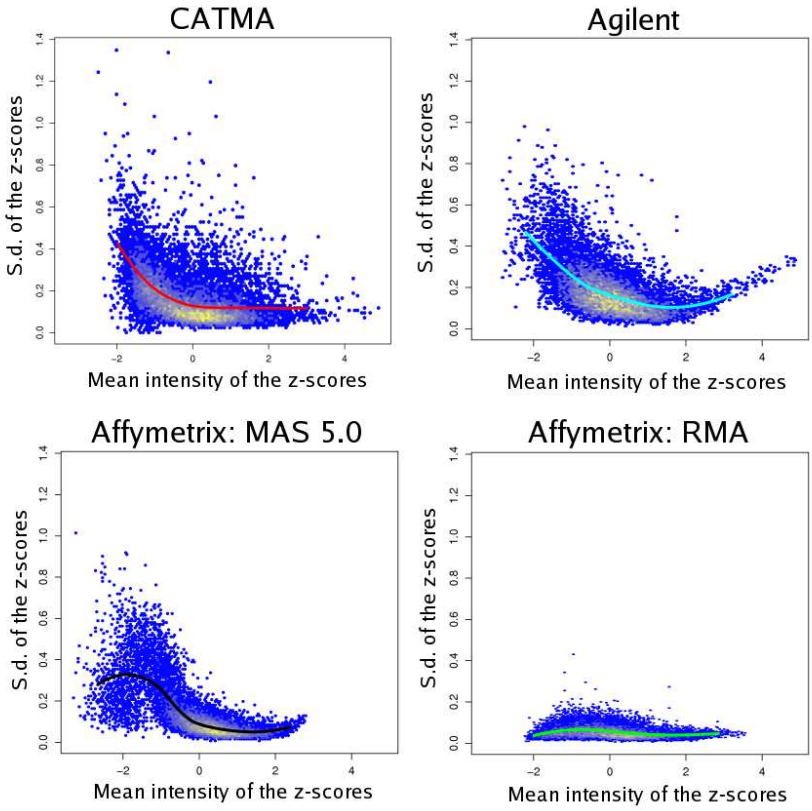


Figure 4.7: **Standard deviation versus mean intensity of the Z-scores.** The color scale of the plot reflects the density of the cloud (yellow to blue, highest to lowest density). The colored line through the data cloud represents the Loess line indicating the overall trend in the data. (A) CATMA; (B) Agilent; (C) Affymetrix with MAS 5.0 preprocessed data; (D) Affymetrix with RMA preprocessed data.

The resulting plots indicate a significant correlation between the individual signal values, and hence the hybridization characteristics of the probe elements. This is particularly satisfactory considering that the strategy for probe design was quite distinct for the three array types. The corre-

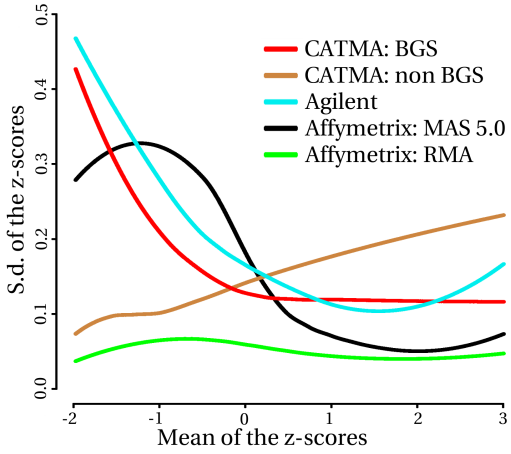


Figure 4.8: **Summary of signal reproducibility in function of intensity.** The Loess lines represent for each data set the overall trend of the Z-score standard deviation as a function of the Z-score mean for each gene.

lation coefficients for pair-wise comparisons are listed in Table 4.9. Not surprisingly, the highest correlation was measured between the MAS 5.0 and RMA expression values, both obtained from the same Affymetrix chips. Furthermore, there was a fair agreement of signal intensities when Affymetrix was compared to either CATMA or Agilent. The comparison of CATMA to Agilent yielded the lowest correlation.

4.10 False positives and FDR

One of the most important issues in microarray analysis is the reliability in the measurement of gene expression differences. On the one hand, poorly chosen boundaries to define meaningful fold changes may include too many false positives or false negatives. On the other hand, microarray statistics must cope with genome-wide datasets and minimize the number of false positives that may result from the multiple-testing problem (Benjamini and Hochberg (1995); Storey and Tibshirani (2003)). However, it is now generally accepted that the Bonferroni correction is much too restrictive. We have investigated systematically the accuracy of the plat-

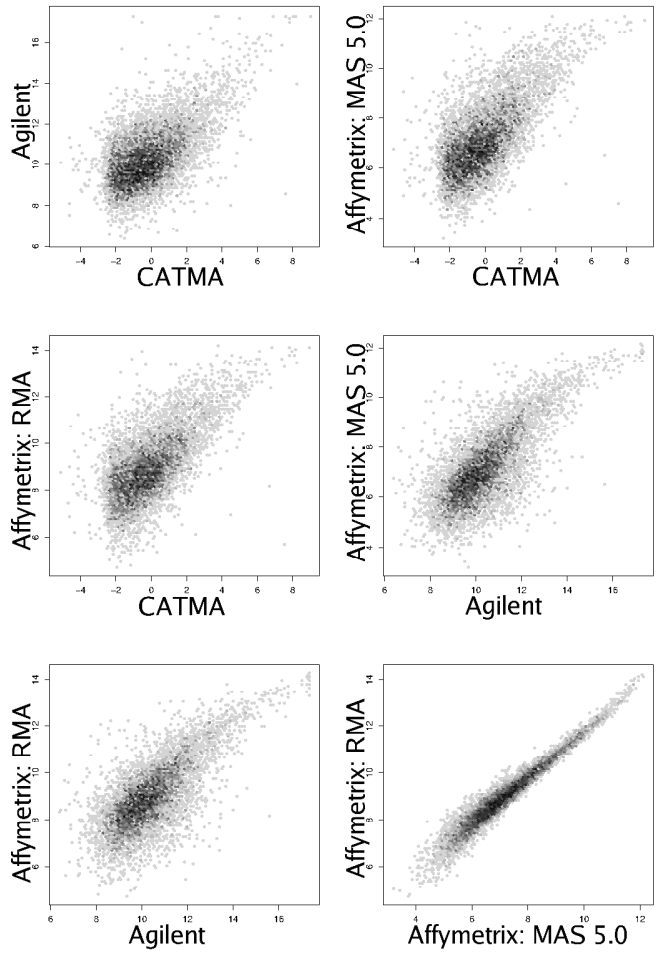


Figure 4.9: **Pairwise platform comparison of the intensities.** Comparison of the $\log_2$ intensities of the genes common to all three platforms and above detection threshold on Affymetrix and CATMA for at least half of the hybridizations.

forms in calling differentials using various statistical tools. Although our experimental design does not address the reliability of small fold-changes

Platforms	Correlation
CATMA/Agilent	0.5833
CATMA/Affymetrix MAS 5.0	0.6619
CATMA/Affymetrix RMA	0.6681
Agilent/Affymetrix MAS 5.0	0.7157
Agilent/Affymetrix RMA	0.7292
Affymetrix MAS 5.0/Affymetrix RMA	0.9728

Table 4.6: **Correlation between the platforms.** Correlation between the platforms was calculated for the $\log_2$ intensity signals of genes with probes on all three platforms. Only those genes were compared that were given a present call by the Affymetrix MAS 5.0 software in at least four out of the eight hybridizations and scored above background for at least seven out of the fourteen hybridizations on CATMA arrays.

(our lowest actual real fold-change is 10), it is useful because we have a large amount of measurements of fold-changes equal to 1. Again we benefit from the fact that all hybridizations rely on a single batch of *Col* shoot RNA, and the hybridizations series therefore essentially consist in eight or fourteen repetitions (Figure 4.2) and are valuable to assess in depth the robustness of the platforms. Taking advantage of our datasets and excluding the spike controls, we estimated the fraction of genes that are erroneously called differentially expressed using the statistical tool LIMMA (Smyth (2004); Section 3.3.1). A gene was called differentially expressed if the moderated t-test had a p-value, corrected to control for the false discovery rate (FDR), smaller than 0.05 (Benjamini and Hochberg (1995)).

This *Benjamini-Hochberg procedure* to control the false discovery rate works as follows. Suppose, one has m hypothesis tests $H_1, H_2, \ldots, H_m$ with their corresponding p-values $p_1, p_2, \ldots, p_m$. If one orders these p-values as $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$, one can compute k such that

$$k = \max_i \left(p_{(i)} \leq \frac{i}{m} q \right)$$

for a chosen value q. By rejecting those hypothesis tests $H_{(i)}$ with

$i = 1, 2, \ldots, k$, the false discovery rate is controlled at q (Benjamini and Hochberg (1995)).

To simulate a biological sample comparison for each platform, data from eight hybridizations were randomly assembled in two groups of four hybridizations. For Affymetrix, the expression measurements of these two subgroups were compared as two different samples each hybridized four times. For example, hybridization 1, 4, 5, and 8 (hybridization numbering as in Figure 4.2) measure an artificial sample A, while hybridization 2, 3, 6, and 7 measure sample B. As these sample names A and B are assigned at random in this artificial comparison and measure actually the same sample for all hybridizations, there should be no differentially expressed genes. For the two-channel arrays, one subgroup was used to calculate log-ratios of a two-sample comparison, while the second group was used to obtain dye-swap ratios. For example, hybridization 1, 3, $5'$, and $7'$ measure sample A in Cy5 versus sample B in Cy3; while hybridization 4, 6, 7, and $2'$ measure the dye-swap (i.e., sample A in Cy3 and sample B in Cy5).

We next used LIMMA to identify genes that appeared to be differentially expressed, based on these eight log-ratios. To get an average estimate of this false positive fraction, the procedure was repeated for all 70 possible different permutations of two sets of four arrays from the eight Affymetrix hybridizations, and for 70 different random assemblies of the two-channel platform array sets. The results are shown in Table 4.7. Because identical samples were compared, all differential genes are false positive observations. For each platform the minimum, average and maximum false positive rates are shown. The average false positive fraction was 2.16% for CATMA BGS, whereas it was 3.43% and 8.62% for Agilent and Affymetrix (MAS 5.0), respectively. The RMA-processed Affymetrix data yielded a smaller fraction of 7.71%, whereas the CATMA data analyzed without background subtraction (CATMA non-BGS) gave 0.73% false positives. These percentages would result in significant numbers of falsely identified differentially expressed genes, as indicated in the fourth column of Table 4.7. Interestingly, CATMA BGS gave the lowest range in the false positive fractions calculated in the 70 iterations,

Platforms	Low %	High %	Mean %	SD	Mean False Positives	Total Gene No.
CATMA BGS	1.60	2.77	2.16	0.19	410	18,967
CATMA non-BGS	0.30	1.98	0.73	0.33	138	18,967
Agilent	0.56	14.59	3.43	2.60	559	21,487
Affymetrix MAS 5.0	5.58	19.69	8.62	3.59	1,959	22,732
Affymetrix RMA	1.72	36.11	7.71	7.52	1,753	22,732

Table 4.7: **Detection of false positives.** For each platform, we selected all gene probes, omitting the spikes. For each platform and data preprocessing method, the percentages and numbers are given of genes flagged by the LIMMA procedure as differentially expressed. The results reflect 70 iterations of the LIMMA procedure, as described in the text.

with a standard deviation of 0.189. These results have to be treated with some caution as they not only reflect platform characteristics but also how well the LIMMA model fits the different datasets.

4.11 False negatives

Finally, we compared the accuracy of the platforms based on their ability to avoid false negative observations. Instead of investigating intensity values for invariant genes, we now focused on those corresponding to the thirteen spike RNAs and determined whether the data supported the correct statistical identification of 10-fold concentration increases. For that purpose, the LIMMA procedure was used to test whether spike genes were detected as differentially expressed when comparing consecutive spike mixes (1 vs. 2, 2 vs. 3, etc.; Table 4.3). The p-values obtained from the moderated t-test were corrected to control the FDR, according to the method of Benjamini and Hochberg (1995), with a significance threshold of 0.05. The results of the consecutive concentration comparisons are given in Table 4.8. For both CATMA and Agilent data LIMMA failed to distinguish correctly between a transcript absent and present at 0.1 cpc or between 0.1 and 1 cpc, confirming that the sensitivity threshold was between 1 and 10 cpc. In the CATMA data set this difference was correctly detected for 10 out of

13 cases, and for 6 out of 12 in the Agilent data. Additionally, for Agilent, four of the spikes were not accurately differentiated between 1,000 and 10,000 cpc, which can be explained by the saturation effect already observed in the dose-response curves (Figure 4.3). The number of false negatives from the Affymetrix data could not be estimated because of the insufficient numbers of replicates; each intensity in the range from 0.1 to 10,000 was only measured once for each spike, except for the intensity of 100.

4.12 Conclusion

Two technologies have dominated the microarray field: cDNA and oligonucleotide arrays. The main advantage of cDNA microarrays (Section 3.1.1) has been their relatively low cost. Affymetrix oligonucleotide arrays (Section 3.2.1), however, take advantage of the available genome sequence and are considered to offer higher reproducibility, but at a higher cost. More recently, long oligonucleotide platforms ($60 - 80$mers; Section 3.1.1) have emerged as a competing technology. Whereas the cost of these oligonucleotide-based technologies is slowly decreasing, multiple problems appear with the cDNA-based arrays, as the difficulty in obtaining full-genome coverage, lack of standardization among laboratories, higher levels of noise, and cross-hybridization between homologous transcripts (Section 3.1.1).
Here, we present the CATMA array for *Arabidopsis* that addresses these shortcomings. It is based on a standardized genome-scale PCR amplicon library, with minimal cross-hybridization and high quality control. The library is available at low cost for the production of spotted arrays.

To assess the quality of the data obtained with CATMA arrays, two commercial platforms, Affymetrix and Agilent arrays, were included in the performance study. All datasets were produced independently by laboratories best positioned to provide service with their particular platform. The differences observed resulted from a combination of factors: the arrays themselves but also all the equipment necessary for their processing, including the hybridization and washing station, the slide

Spike RNA	0 vs 0.1	0.1 vs 1	1 vs 10	10 vs 100	100 vs 1,000	1,000 vs 10,000
	CATMA	CATMA	CATMA	CATMA	CATMA	CATMA
1	0	0	-1	-1	-1	-1
2	0	0	0	-1	-1	-1
3	-1	+1	-1	-1	-1	-1
4	0	0	-1	-1	-1	-1
5	-1	0	-1	-1	-1	-1
6	+1	-1	-1	-1	-1	-1
7	0	0	-1	-1	-1	-1
8	0	+1	-1	-1	-1	-1
9	0	+1	-1	-1	-1	-1
10	0	0	-1	-1	-1	-1
11	0	0	0	-1	-1	-1
12	0	0	-1	-1	-1	-1
14	0	0	0	-1	-1	-1
	Agilent	Agilent	Agilent	Agilent	Agilent	Agilent
1	0	0	-1	-1	-1	0
2	0	0	-1	-1	-1	-1
3	0	0	0	-1	-1	-1
4	0	0	0	-1	-1	-1
5	0	0	-1	-1	-1	0
7	0	0	0	-1	-1	-1
8	-1	0	-1	-1	-1	-1
9	0	+1	0	0	-1	-1
10	0	-1	0	-1	-1	0
11	-1	0	-1	-1	-1	-1
12	0	0	-1	-1	-1	0
14	0	0	0	-1	-1	-1

Table 4.8: **Detection of false negatives.** The LIMMA procedure was used to compare consecutive sets of concentrations (0.1 cpc against 0 cpc, 1 cpc against 0.1 cpc, etc.). '-1' and '+1' indicate that the gene is flagged by LIMMA as down-regulated or up-regulated, respectively, whereas '0' is used for genes that do not appear to be differentially expressed. All pair-wise comparisons should theoretically be assigned '-1'.

scanner, and the software application producing the raw microarray data

file.

The comparison was based on a single, large shoot RNA sample spiked with synthetic poly(A) RNAs in various quantities. These were added to evaluate signal detection over a range of biologically meaningful abundance classes. The spike concentrations spanned a wide range of subsequent 10-fold dilutions, covering both the high, intermediate, and scarce abundance classes, allowing us to establish the detection dynamic range. We chose to use a significant number of spikes (14) to guarantee the robustness of the study and to attempt to address more extensively than most studies the potential for illegitimate hybridization. Except for the faulty Spike 13, all spike RNAs showed extremely similar hybridization characteristics, and the hybridization results, combining the spike genes and the genes transcribed in *Arabidopsis* shoots, constituted an extensive data set for a detailed comparison of the different platforms. The CATMA array performed very well when compared to the commercial oligonucleotide systems. Even at the highest concentrations (10,000 copies per cell), it showed no sign of saturation or signal decrease, whereas Agilent and Affymetrix arrays conspicuously lacked signal linearity in that range. For Affymetrix, RMA-processed data were slightly less saturated compared to MAS 5.0. In the Agilent data output file, some of the spike probes at the highest concentrations were flagged as saturated, together with 27 other probes, almost all corresponding to nuclear genes with chloroplast function (see Table 4.4), suggesting they still represented biologically relevant transcript levels. Although we could have performed multiple scans at different laser powers or detector gains, we chose to use a single setting because that is how microarray data are produced routinely by service providers. Also, integration of data resulting from multiple scans is cumbersome. Our results indicated that for abundant mRNAs, the CATMA array performed substantially better than both the short and long oligonucleotide arrays and will yield more accurate ratio-fold changes for such transcripts.

Overall, the three platforms were comparable in sensitivity, although results varied somewhat according to spikes. For some, the signal was still above background level at a concentration of 1 copy per cell, equivalent

to scarce RNAs. Because of the numerous replicates in the experimental design, the CATMA and Agilent platform sensitivity could be assessed with the LIMMA algorithm. The discrimination between subsequent spike RNA levels started to deteriorate between 1 and 10 copies per cell (Table 4.8), for which CATMA data yielded a correct call for 10 out of 13 spikes, whereas the Agilent data were accurate for 6 out of 12 spikes. Thus, we conclude that the sensitivity of CATMA arrays was at least equivalent to that of the Agilent arrays. A direct LIMMA comparison of Affymetrix with the other platforms was not possible because the Affymetrix experiment lacked sufficient replicates.

The analysis of CATMA data with background signal correction clearly produced the best dose response curves (for comparison, see Figure 4.4). However, background subtraction introduced a significant level of variance into the data, particularly for low signal. These somewhat contradictory findings illustrate the fact that there is still no single solution for data preprocessing: it remains prudent to test various alternatives even at the preprocessing level to thoroughly mine microarray datasets for information about gene expression levels. This is also evident from the differences observed between the Affymetrix results obtained with the MAS 5.0 or RMA packages. In our comparison, the RMA package outperformed MAS 5.0 for all studied parameters: dynamic range, reproducibility across the range of signal intensity, in particular for low or background signal, and FDR. The better performance of the RMA software clearly demonstrates that the mismatch features, not taken into consideration by RMA, are better discarded to measure gene expression. Interestingly, the datasets generated for this study, containing numerous repetitions and including three competing systems, may serve for the comparative evaluation of improved and future algorithms. The choice of preprocessing protocols is especially important to establish coherent repositories of data compendia, as such large databases will hold data from heterogeneous sources. A major challenge will be to effectively integrate data from different platforms for analysis and mining purposes, for example by using cross-platform normalization methods (Ferl et al. (2003))) or by taking p values, computed from the expression measurements of the different experiments (Rhodes et al. (2002)).

CATMA array probes were selected to exclude homology exceeding 70% identity. A similar design strategy was used for the probes of the two oligonucleotide arrays. Therefore, it came as no surprise that cross-hybridization could not be detected for any of the arrays, not even with spike RNAs at 10,000 copies per cell representing up to 3.3% of the poly(A) RNA pool. The ability of the tested platforms to exclude cross-hybridization problems because of sequence homology is a big advantage over cDNA-based arrays.

The coverage of the three arrays was matched against the latest TIGR annotation (release 5.0) of the *Arabidopsis* genome. The CATMA v2 array is comparable with the oligonucleotide arrays. Yet, microarray probe design has a moving target and all platforms will further evolve with advances in genome annotation because experimental transcription data are constantly growing, gene prediction algorithms are continuously improving, and new genome sequences are becoming available. The design of CATMA v3 yields an additional 6,000 probes, taking advantage of both the TIGR 5.0 annotation and the gene models obtained with recent improvements of the EuGène gene finder (`http://bioinformatics.psb.ugent.be/genomes_ath_index.php`). Likewise, Affymetrix is working on a new version of the ATH array, and Agilent has introduced the Arabidopsis 3 oligonucleotide array with close to 40,000 features. It will take a few more years before the Arabidopsis gene repertoire becomes completely stable, and additional updates of the array feature sets will be necessary.

Hence, the sensitivity, specificity, and coverage of the CATMA array make CATMA a strong competitor for other microarrays currently available for genome-scale transcript profiling. Because its probes are designed from the complete genome sequence rather than selected from available cDNA or EST collections, it minimizes homologies between probes, and maximizes the genome coverage. The up-front investment in the clone library has thus resulted in an ideal low-cost alternative for in-house spotting.
These series of synthetic RNAs provided detailed information about the dynamic response of the microarrays. Our results indicate that CATMA

arrays perform equally well as Agilent or Affymetrix arrays in terms of sensitivity, specificity, and the ability to prevent detection of false negative and false positive genes in differential expression studies. However, both the long and short oligonucleotide platforms suffer from signal saturation at high target concentrations, whereas the CATMA array does not. The solid performance of the CATMA array makes it a valid platform for functional genomics studies, and a well-managed core facility may be able to offer CATMA array service at a cost highly competitive with commercial alternatives.

Material and Methods

Plant material and RNA extraction

Arabidopsis (*Arabidopsis thaliana L.*) Heynh. *Col*-1 seeds were sown, cold stratified (at 4°C for 7 days), and grown at long-day conditions ($22°C$, $16h$ light/ $8h$ dark, with cool-white light [tube code: 840] 65 $mEm^{-2}s^{-1}$ photosynthetically active radiation) on agar-solidified culture medium (1 × MS [Duchefa, Haarlem, The Netherlands], 0.5 gL^{-1} MES, pH 6.0, 1 gL^{-1} sucrose and 0.6% plant tissue culture agar [LabM, Bury, UK]). Whole shoots were harvested at growth stage 1.04 corresponding to a fourth leaf length of approximately 1 mm (Boyes et al. (2001); developmental stage equivalent to TAIR development term 0000399), $6h$ after dawn, and immediately frozen in liquid nitrogen. Total RNA was extracted from pooled plant material using the TRIzol reagent (Invitrogen, Carlsbad, CA).

Preparation of spiked RNA samples

Spike poly(A) RNAs were synthesized from selected cDNA clones (Table 4.2; EMBL accession nos. AI997299, AI996580, AI998315, AI999518, AI995329, AW004197, AI995484, AI993419, AI994579, AI994777, AI992430, AI995003, AI995254, and AI994049) from a $6K$ cDNA collection distributed originally by Incyte, now available through Open Biosystems (Huntsville, AL; see `http://www.microarray.be/servicemainframe.htm`) and constructed by *Not*I-*Sal*I directional cloning in either Lambda ZipLox (Invitrogen) or pSPORT1. All clones were validated for this particular study by sequencing. Plasmid DNA was linearized by *Not*I digestion, the restriction site being positioned immediately after the poly(A) tail sequence; 1 μg of linearized plasmid was used as template for the in vitro synthesis of sense transcripts with the T7 RNA polymerase (AmpliScribe T7 High Yield transcription kit; Epicentre, Madison, WI). Following DNAseI treatment, the transcribed RNAs were purified by ammonium-acetate precipitation and resuspended in diethyl pyrocarbonate-treated water. The quality and quantity of all RNA samples (spikes and Col shoot total RNA) were assessed with the RNA LabChip (Bioanalyzer

2100; Agilent Technologies) and classical spectrophotometry. Despite our efforts to carefully quality control all spike RNAs, we originally overestimated Spike 13 RNA concentration and integrity and could not draw meaningful conclusions from it in the analysis of the hybridization data. We therefore omitted this spike from all subsequent analyses.

A large batch (500 μg) of *Arabidopsis* (*Columbia*) shoot RNA was diluted to 1 $\mu g \mu L^1$ and used to prepare 7 test samples at a final concentration of 0.5 $\mu g \mu L^1$, each containing a full range of spike RNAs at concentrations ranging from 0.1 to 10,000 cpc. Care was taken to use water containing total RNA at all dilution steps, to prevent the loss of spike RNAs at low concentrations through adsorption on plastic surfaces. An eighth RNA sample was constructed containing all RNA spikes at a concentration corresponding to 100 cpc. The eight RNA samples were constructed each in a single separate tube, aliquoted, and processed according to the protocols specific to each platform. All RNA samples were again checked for quality and quantity with the RNA LabChip at the end of the dilution procedure.

CATMA GST microarray

Design and synthesis of primary and secondary GST amplicons were described elsewhere (Thareau et al. (2003); Hilson et al. (2004)). As described, the GSTs primarily match ($3'$) exons or $3'$ untranslated region (UTR) sequences and occasionally (2.9%) contain intron sequences. The CATMA v1 array used in this study consisted of 19,992 features, including 18,981 unique GSTs, 768 positive/negative controls (Amersham BioSciences), and 243 blanks. GST PCR products were purified with MinElute UF plates (Qiagen, Hilden, Germany) and arrayed in 50% dimethyl sulfoxide on Type VIIstar reflective slides (Amersham BioSciences) using a Lucidea Array spotter (Amersham BioSciences). The spots had a diameter of approximately 100 microns and were 173 $\times$ 173 microns apart. The array design can be accessed via the ArrayExpress database as accession number A-MEXP-10 (`http://www.ebi.ac.uk/arrayexpress`) or via the VIB MicroArray Facility Web site (`http://www.microarrays.be`). Prior to hybridization, the slides were washed in $2\times$ saline-sodium phosphate-EDTA buffer, 0.2% SDS for 30 min at $25°C$.

RNA was amplified using a modified protocol of in vitro transcription as described previously (Puskás et al. (2002)). Briefly, $5\mu g$ of total RNA was reverse transcribed to double-stranded cDNA using an anchored oligo(dT) + T7 promoter [5′-GGCCAG-TGAATTGTAATACGACTCACTATAGGGAGGCGG-T24(ACG)-3′ (Eurogentec, Seraing, Belgium)]. From this cDNA, RNA was produced via T7-in vitro transcriptase until an average yield of 10 to $30\mu g$ of amplified RNA. The amplified RNA ($5\mu g$) was labeled with dCTP-Cy3 or Cy5 (Amersham BioSciences), by reverse transcription using random nonamer primers (Genset, Paris). The resulting probes were purified with Qiaquick (Qiagen) and analyzed for amplification yield and incorporation efficiency by measuring the DNA concentration at 280 nm, Cy3 incorporation at 550, and Cy5 incorporation at 650 using a Nanodrop spectrophotometer (NanoDrop Technologies,

Rockland, DE). A good target had a labeling efficiency of 1 fluorochrome every 30 to 80 bases. For each target, 40 pmol of incorporated Cy5 or Cy3 were mixed in $210\mu L$ of hybridization solution containing 50% formamide, $1\times$ hybridization buffer (Amersham BioSciences), 0.1% SDS. Each spike mix was hybridized against the reference RNA (spikes at 100 cpc) and repeated with dye swap to make up 14 hybridizations in total (Figure 4.2).

Hybridization and posthybridization washing were performed at $45°C$ with an Automated Slide Processor (Amersham BioSciences). Posthybridization washing was done in 1x sodium chloride/sodium citrate buffer (SSC), 0.1% SDS, followed by $0.1\times$ SSC, 0.1% SDS and $0.1\times$ SSC. Arrays were scanned at 532 nm and 635 nm using a Generation III scanner (Amersham BioSciences). Images were analyzed with ArrayVision (Imaging Research, St. Catharines, Canada).

All protocols are available at the VIB MicroArray Facility Web site (`http://www.microarrays.be`) and at ArrayExpress under accession numbers P-MEXP-578, P-MEXP-579, P-MEXP-581, P-MEXP-582 for Cy3 labeling, Cy5 labeling, hybridization, and scanning, respectively. The CATMA transcript profiling data have been submitted to ArrayExpress under accession number E-MEXP-30.

Agilent and Affymetrix Microarrays

The protocols used by ServiceXS for Agilent data production were published by Agilent Technologies, in particular the manuals Low RNA Input Fluorescent Linear Amplification Kit (version 1.0, February 2003) and Agilent 60-mer Oligo Microarray Processing Protocol (version 7.0, April 2004). Arrays were scanned with maximum (100%) laser intensity in both channels (default settings) to obtain maximum sensitivity. Lower intensity scanning may correct for saturated features. Features were extracted with background subtraction or with spatial detrending (Feature Extraction Software version 7.5). Spatial detrending estimates the background signal by fitting a surface over the lowest 1% to 2% of the intensities. By subtracting this surface fit, a systematic intensity gradient on the microarray is removed, thereby correcting for a background trend rather than local background measurements that may be biased. Apart from a slight decrease in the percentage of spots above background, spatial detrending gave essentially the same result as the background-subtraction method.

The procedures used for Affymetrix data production are described in the documentation provided by NASC (`http://nasc.nott.ac.uk/`; Craigon et al. (2004)), available together with the data from the ArrayExpress database (`http://www.ebi.ac.uk/arrayexpress`, accession no. E-NASC-32). For Affymetrix data, the hybridization characteristics of the internal RNA controls (Section 3.2.3) were monitored as an additional quality control: (1) the $3'$:$5'$ ratios for GAPDH and β-actin ranged from 1.0287 to 1.2408 and from 1.8012 to 2.1705, respectively, and are all indicative of successful hybridizations; (2) the spike controls (BioB, BioC, BioD, BioM, and CreX) were present on

all chips, except for BioB 5′ and BioB 3′ called 'Marginal' for chips 1 and 3, respectively; (3) when scaled to a target intensity of 100 (using Affymetrix MAS 5.0 software), scaling factors for all arrays were within acceptable limits (ranging between 0.311 and 0.518), as were background and mean intensity values. For all hybridizations, quality and quantity of starting RNA were verified by agarose gel electrophoresis and RNA LabChip analysis. The Agilent and Affymetrix transcript profiling data have been submitted to ArrayExpress under accession numbers E-MEXP-197 and E-NASC-43, respectively.

In Silico Coverage

The coverage of the three platforms was compared by BLAST analysis of their probe sequences against TIGR 5.0 gene models. The sequences of these gene models, including pseudogenes and transposable elements, were extracted from the XML files describing the chromosomes (at `ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/PSEUDOCHROMOSOMES`). The probes of Affymetrix and Agilent were designed based on TIGR annotation releases 2 and 3, respectively (available in the archives at `http://www.tigr.org`). The probes of CATMA were designed on gene models predicted by the EuGène software (Schiex et al. (2001)), supplemented with gene models uniquely described in the TIGR 3 release. For the analysis of Affymetrix and Agilent probes, we used only exonic sequences to correctly position probes that span exon boundaries. In line with the original design criteria employed for the GSTs, we used complete gene models including 3′ UTRs, to be able to correctly locate probes that were designed to span intron-exon boundaries or exon-3′ UTR boundaries. The set of sequences extracted from the TIGR files for the comparison against Affymetrix and Agilent contained the complete gene structure (exons, introns, and 3′ UTR sequences) of all protein-encoding genes, including their splice variants, and the pseudogenes. For CATMA, we extracted exon and intron sequences of all protein-encoding genes, and the pseudogene sequences. For both databanks, we added either the full 3′ UTR sequence or arbitrarily the 150 bases following the stop codon (when the 3′ UTR was shorter than 150 bases or if no 3′ UTR was available).

The sequences of the Affymetrix probe sets were retrieved from the company's Web site (`http://www.affymetrix.com/`), the sequences of the Agilent probes were retrieved from the company Web site (`http://www.agilent.com`; restricted pages requiring transfer agreement for access), and CATMA v2 were derived from the Array Design File accession number A-MEXP-58, publicly available at ArrayExpress (`http://www.ebi.ac.uk/arrayexpress/`). Perl scripts were used to extract the genes from XML files, to reconstitute exonic gene sequences, to adjust 3′ UTR sequences, and to automate the BLAST and extraction of data from the BLAST output files.

CATMA sequences ($150 - 500$ bp) matched TIGR 5.0 when aligned over at least 150 bases allowing for at most two discrepancies (base mismatch or gap); Agilent sequences (60mer) when aligned over the whole probe length allowing at most one base mismatch or gap; and Affymetrix probe sets (11 probes of 25 bases each) when at least eight probes from a set aligned perfectly. Splice variants were merged to allow comparison of CATMA

hits (BLAST against gene) with Agilent and Affymetrix hits (BLAST against all possible splice variants). TIGR 5.0 genes represented by features in the different arrays were simply counted based on these criteria.

Chapter 5

The CAGE project

In this chapter a European project, the Compendium of Arabidopsis Gene Expression or CAGE project, will be presented. This large project aimed at delivering a compendium of gene expression profiles in Arabidopsis thaliana. In a first section, the project will be presented. The following sections will elaborate on the deliverables of the CAGE project, for which our group had an active role. First of all, this includes a data preprocessing pipeline. All data communication between the partners, involved in the project, was done via MAGE-ML. The pipeline that we will present here starts therefore from MAGE-ML files, extracts all necessary information, performs a quality assessment and applies a within-slide normalization. The experimental design, imposed by the CAGE consortium, requested an alternative normalization strategy. Therefore, this design and its implications towards data normalization will be discussed. From the preprocessed data, users can then start to analyze experiments in depth.

As the data production within the CAGE project was slower than anticipated, the analysis of the compendium was not anymore within the scope of this work. But, we will present a preview of a smaller analysis, that compares two time series on leaf development, produced by two different partners.
Within the CAGE project, analyses of smaller subsets of the CAGE data set have been performed in close collaboration with some of the partners. This chapter will conclude with an example of such an analysis.

5.1 Compendium of *Arabidopsis* Gene Expression or CAGE

The *Compendium of Arabidopsis Gene Expression* or CAGE project is a European Demonstration project (project no. QLK3CT200202035) that aimed at producing an atlas of gene expression of *Arabidopsis thaliana* throughout its life cycle and under a variety of stress conditions. The project involved eleven partners (see Table 5.1) and started on November 1, 2002. Initially, it ended after three year, on October 31, 2005, but this period was extended with an additional 6 months, until April 30, 2006. Most partners already collaborated within the CATMA project (partners 1 – 9, Section 4.1).

Within the project, a total of 2,000 RNA samples were planned. For each sample, a biological replicate was planned, hence the sample list consists out of 1,000 different biological conditions.
The sample list contains three ecotypes *Columbia* (*Col*),*Landsberg erecta* (*Ler*), and *Wassilewskija* (*Ws*), which are three commonly used ecotypes as genetic background for mutants.
The samples can be divided in three groups:

- the *wild type samples*:
  this sample list contains amongst others:
  - time series of whole plant, leaf (1+2), and root
    The time series covered the major developmental stages. As a definition of the growth stages, the different stages described in Boyes et al. (2001) were used (see Table 5.2).

	Abbreviation	Partner
1	VIB-PSB	VIB, Department of Plant Genetics, Gent, Belgium
2	VIB-MAF	VIB, MicroArray Facility, Leuven, Belgium
3	URGV	Unité de Recherche en Génomique Végétale, INRA, Evry, France
4	RUU	Department of Molecular Genetics, University of Utrecht, Utrecht, The Netherlands
5	HRI	Horticulture Research International, Wellesbourne, UK
6	UNIL	Gene Expression Laboratory, Institute of Ecology, University of Lausanne, Lausanne, Switzerland
7	MPI-MG	Max-Planck Institut of Molecular Genetics, Department of Lehrach, Berlin, Germany
8	CSIC	Centro Nacional de Biotecnología, Madrid, Spain
9	SLU	Department of Forest Genetics and Plant Physiology, Umea, Sweden
10	ESAT-SCD	Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium
11	EBI	European Bioinformatics Institute, Hinxton, UK

Table 5.1: Partners in the CAGE project: This table shows the list of the partners involved in the CAGE project. Throughout the text, the numbering and abbreviations as mentioned in the table will be used to indicate specific partners.

  - flower and silique samples at single time points
  - stress conditions, as for example salt stress, drought, day length, or temperature on *Col*

- *mutant samples*:
  samples, in which specific genes are knocked out or have an upregulated gene activity

- *research samples*:
  For each partner, about half of the samples were defined in function of their own internal research projects.

Partners 1 and 4 are the big data producers and aimed at producing each 400 samples. The remaining data producing partners $(3,\ 5-9)$ planned to deliver each 200 samples.

The compendium data is being made available to the community as a publicly accessible database by the European Bioinformatics Institute (Partner 11), enabling query and upload, so that the compendium can even be expanded after the termination of the CAGE project.

The project not only aimed to supply a compendium of gene expression data, but it had also other related and important aspirations.
First of all, as a microarray platform the CATMA array was chosen and this project aimed to demonstrate the utility of the CATMA array. This CATMA array was a novel platform at the start of the project and its capacity had not yet been demonstrated. The benchmarking of the CATMA array against two commercial platforms, as presented in Chapter 4, was therefore also done within the framework of the CAGE project.
As all data were produced by different partners, it was vital to agree on a set of guidelines for the data production, as, for example, growth and sampling conditions, RNA extraction and hybridization protocols. These standardizations had as purpose to reduce the noise of the data and to increase the quality. These guidelines will also be at the disposal of future microarray users.
To further increase the standardization, all microarrays would also be produced by one single partner, the VIB MicroArray Facility (partner 2).

Stage Number	Approx. number of days after sowing	Description
0.0		Seed germination
0.1	3.0 (on plates)	Seed imbibition
0.5	4.3 (on plates)	Radicle emerges from seed coat
0.7	5.5 (on plates)	Hypocotyl and cotyledon emerge from seed coat
1		Leaf development
1.0	6.0 (on plates)	Cotyledons fully open
1.02	10.3 (on plates), 12.5	2 rosette leaves $>$ 1 mm
1.03	14.4 (on plates), 15.9	3 rosette leaves $>$ 1 mm
1.04	16.5	4 rosette leaves $>$ 1 mm
1.05	17.7	5 rosette leaves $>$ 1 mm
1.06	18.4	6 rosette leaves $>$ 1 mm
1.08	20.0	8 rosette leaves $>$ 1 mm
1.10	21.6	10 rosette leaves $>$ 1 mm
1.12	23.3	12 rosette leaves $>$ 1 mm
1.14	25.5	14 rosette leaves $>$ 1 mm
3		Rosette growth
3.20	18.9	Rosette is 20% of final size
3.50	24.0	Rosette is 50% of final size
3.70	27.4	Rosette is 70% of final size
3.90	29.3	Rosette growth is complete
5		Inflorescence emergence
5.10	26.0	First flower buds are visible in the rosette, plant has not yet bolted
6		Flower production
6.00	31.8	First flower is open
6.10	35.9	10% flowers to be produced are open
6.50	43.5	50% flowers to be produced are open
6.90	49.4	Flowering complete
8		Silique or fruit ripening. Seed pods become brown and then shatter.
8.00	48.0	First silique or seed pod shatters.
9		Whole plant senescence begins. Plant starts to lose, pigment becoming brownish.
9.70		Senescense complete

Table 5.2: **Developmental stages:** This table lists the developmental stages at which samples were taken within the CAGE project. This is a subset of the stages as defined in Boyes et al. (2001).

As a design, a reference design was used and each sample had a technical replicate, hence a total of 4,000 microarrays were planned.
The project also aimed to establish a data submission and processing pipeline, and analytical tools for the *Arabidopsis* research community. In this part, our group was involved.
After this brief overview of the different aspects of the CAGE project, we will discuss more into detail the design of the experiment.

5.2 The design of the experiments within CAGE

The ultimate goal of the CAGE project was to set up a compendium and, therefore, it is vital that samples produced under a variety of conditions or from different ecotypes can be combined and compared. Hence, within the project, there was decided to use a reference design (i.e., all samples are compared against a common reference sample, see Figure 5.1(a)). For the CAGE project an artificial reference sample was used.
The choice for a reference design is not a straightforward decision, while other, more advanced designs are available, as a loop or factorial design (see Figure 5.1(b) and (c), respectively). The main advantage of a loop design or a factorial design is that they are more powerful and provide a higher informativeness for specific questions. However, to obtain these designs, well-defined questions have to be defined. Such complex designs are also harder to carry out. Therefore, at the start of the project, it was decided to use a reference design. However, by now, more advanced designs are probably a viable option.
For the research samples, the different partners had the choice to follow that decision and to stick to the reference design, or to set up more advanced designs to answer their specific research questions of interest.

5.2.1 The oligo reference design

As described in Section 3.1.1, to all GST probes a primer pair was added (see Figure 3.2). This primer pair corresponds to the location of the GST on the well plate and was a combination of 16 and 24 primers, that we will denote as $r1 - r16$ and $c1 - c24$, respectively. The reference mix used within the CAGE project is a mix of the primers $r1$, ..., and $r16$. We will

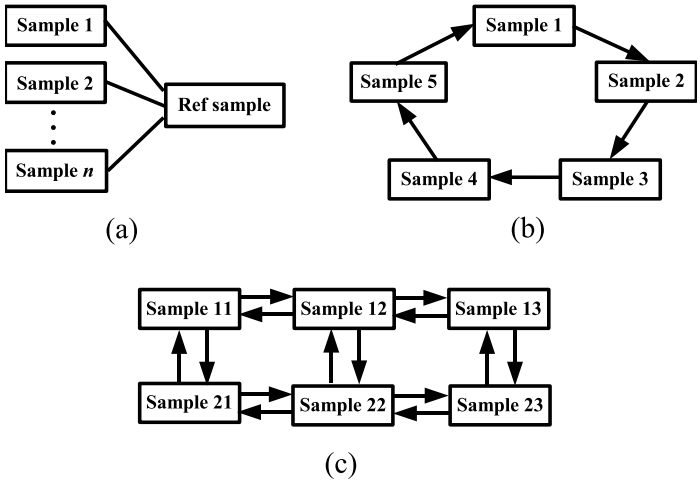


Figure 5.1: **Design of experiments.** The figure shows three possible designs. (a) A reference design: all samples are compared to a common reference sample. (b) A loop design: direct comparisons among the samples, arranged as a loop. (c) A 3×2 factorial design: direct dye-swap comparisons are made between the levels of the factors (see Section 3.3.2).

call this sample the *oligo reference* sample. As all probes on the CATMA array contain one of these 16 primers, this oligo reference sample will bind everywhere and produce a constant signal for all spots.
For the technical replicates, no dye-swap was performed. The oligo reference sample was each time labeled in Cy3 and the actual samples of interest in Cy5. Support for this choice can be found in Sterrenburg et al. (2002):

> "*When using a common reference, the experimental target can always be labeled with the same dye without having different gene-label interactions and thus it is not necessary to perform a dye-swap. Therefore, under our experimental conditions, the data suggest labeling the target with Cy5 and the reference with Cy3.*"

A small experiment, which supported this choice, was the following. Partner 3 performed six hybridizations in which RNA samples of buds and

Hybridization	Cy3	Cy5
1	buds	leaves
2	leaves	buds
3	oligo ref	buds
4	oligo ref	leaves
5	buds	oligo ref
6	leaves	oligo ref

Table 5.3: **Small experiment to test the effect of the dye choice for the oligo reference sample.** Six hybridizations allow to get an idea of the optimal choice for the channel for the oligo reference sample: the Cy3, the Cy5 channel, or a dye-swap.

leaves are compared. Two hybridizations compare buds and leaves directly on a single slide, with a dye-swap. Another set of two hybridizations compares the leaves and buds sample in the Cy5 channel to the oligo reference sample in the Cy3 channel. The last two hybridized the leaves and buds sample in the Cy3 channel and the oligo reference sample in the Cy5 channel (Table 5.3). Data were appropriately normalized with Loess normalization for hybridization 1 and 2 and General Linear Models for hybridization 3 – 6 (Section 5.2.2). Log-ratios between the two samples, leaves and buds, can be computed directly or via the oligo reference sample with oligo reference sample in Cy3, Cy5, or in both. In the case of the dye-swap (using hybridizations 1 and 2 in Table 5.3), the log-ratio of leaves versus buds is computed as

$$\frac{1}{2}\left[\log\left(\frac{\text{leaves}_{\text{Cy5}}}{\text{buds}_{\text{Cy3}}}\right)_1 - \log\left(\frac{\text{buds}_{\text{Cy5}}}{\text{leaves}_{\text{Cy3}}}\right)_2\right]. \tag{5.1}$$

The numbers of the hybridizations (as defined in Table 5.3) used for the computations are indicated in the subscripts.
Via the oligo reference hybridizations (Hybridizations 3 and 4) with the reference sample in the Cy3 channel, a similar log-ratio can be computed

Constructed log-ratio	Correlation
with hybridizations 3 and 4	0.928
with hybridizations 5 and 6	0.788
with hybridizations 3, 4, 5, and 6	0.912

Table 5.4: **Correlation between direct and indirect log-ratios.** The table shows the correlation between the log-ratio computed as in Formula 5.1 and the reconstructed log-ratios, computed via the oligo reference (i.e., with oligo reference in Cy3, in Cy5, or in both). As a reference, the correlation between the dye-swap hybridizations (i.e., hybridizations 1 and 2) equals 0.963.

as

$$\log\left(\frac{\text{Leaves}_{\text{Cy5}}}{\text{Oligo Ref}_{\text{Cy3}}}\right)_4 - \log\left(\frac{\text{Buds}_{\text{Cy5}}}{\text{Oligo Ref}_{\text{Cy3}}}\right)_3. \tag{5.2}$$

Analogously, the log-ratio can also be computed with the oligo reference in the Cy5 channel, by using Hybridizations 5 and 6. Or, one can take a dye-swap into account and combine all four hybridizations 3 – 6 by taking the average. The correlations between these three constructed log-ratios and the direct log-ratio (Formula 5.1) can be found in Table 5.4. From this table, it is clear that, based on the correlations, the log-ratios obtained via the oligo reference in the Cy3 channel show the highest agreement with the direct log-ratios. This result suggests that the choice of labeling the oligo reference sample in Cy3, without a dye-swap, is a sensible decision.

5.2.2 Implications of oligo reference design on normalization

As the oligo reference channel produces a constant signal in the Cy3 channel, the MA-plot has a completely different shape and the commonly used Loess normalization (Section 3.4) will fail in this case. In Figure 5.2(a), an MA-plot of Spike mix 1, as used also in the benchmarking of the CATMA array (Section 4.4), versus the oligo reference channel is shown. The typical shape of an MA-plot as presented in Figure 3.5(b) has disappeared; instead, a rather straight cloud along a diagonal axis is visible. This is a

side effect of the fact that the oligo reference channel has a more or less constant signal. An MA-plot therefore represents

$$M = \log_2\left(\frac{R}{G}\right) \approx \log_2(R) - \text{constant}$$

versus

$$A = \log_2\left(\sqrt{RG}\right) \approx \frac{1}{2}\log_2(R) + \text{constant}.$$

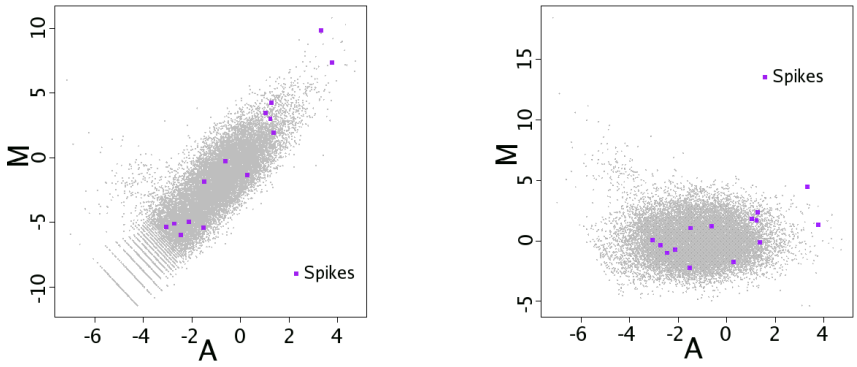
The classical normalization strategies as Loess correction assume that the majority of the genes is not differentially expressed between two samples in the Cy3 and Cy5 channel, which comes down to log-ratios M that are centered around 0. As we have in this case an artificial sample, producing a constant signal, this assumption is not valid anymore. Thus, it makes no sense to normalize the log-ratios to zero by performing a Loess correction (Figure 5.2(b)) and an alternative normalization procedure has to be applied.

A first normalization step that seemed to be opportune in the case of using an oligo reference design, was to correct for the *primer effect*. The intensities of the oligo reference channel are less constant than expected and a significant difference between the 16 primers has been observed. For example, a small self-self experiment was performed in which the oligo reference was hybridized in both channels. By making an MA-plot, the different primers could be clearly distinguished (see Figure 5.3). For all data in the CAGE project, this effect is removed by fitting an General Linear model through the intensities of the oligo reference channel (i.e., the Cy3 channel). This one-factor model can be written as

$$y_{ij} = \mu + \pi_i + r_{ij}, \tag{5.3}$$

where y_{ij} are the background corrected $\log_2$ Cy3 intensities, μ is an overall average, π is the primer effect with 16 levels ($i = r1, \ldots,$ and $r16$) and r_{ij} is the random normal error term for intensity j with primer i. These corrected intensities r_{ij} will be retained.

The second normalization that can be applied is a normalization for the print-tip effect, which is normally removed with a Loess normalization



(a) MA-plot prior to normalization (b) MA-plot after Loess normalization

Figure 5.2: **MA-plot in case of oligo reference design.** In the MA-plots, a hybridization of spike mix 1 (as defined in Section 4.4 and Table 4.3) versus the oligo reference channel is shown. The spikes in the data set are indicated with small squares. (a) Prior to Loess normalization, the MA-plot has an atypical form. (b) Performing a Loess normalization, causes the MA-plot to shrink into one dense cloud.

(Section 3.4). But as this is not appropriate in this case, this effect will also be corrected by fitting an one-factor General Linear model. The model can be defined as

$$y_{ij} = \mu + \tau_i + e_{ij},$$

where y_{ij} are the background corrected $\log_2$ Cy5 intensities or the for primer effect corrected $\log_2$ Cy3 intensities (i.e., r_{ij} from Equation 5.3), μ is the overall average, τ is the print-tip effect and e_{ij} is the random error term for intensity j and print tip i. These corrected intensities e_{ij} are then the corrected log intensities.

These within-slide normalizations were incorporated in a data preprocessing pipeline, developed specifically for the CAGE project.

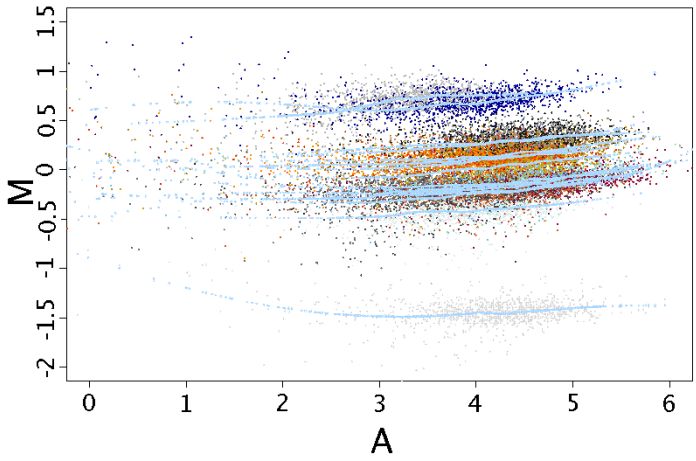


Figure 5.3: **MA-plot of oligo reference versus oligo reference sample.** In the MA-plot of this self-self experiment, the data points are colored according to its corresponding primer ($r1$, ..., or $r16$). Through each of the 16 data clouds a Loess line is fitted. The different clouds can clearly be distinguished, indicating that the Cy3 and Cy5 intensities behave differently for each primer.

5.3 Data preprocessing pipeline

Our main contribution to the CAGE project was to set up a data preprocessing pipeline. All data producing partners submit their data to MIAMExpress at EBI. With a few web forms, they can describe their experiments in detail and upload the corresponding data files. The data uploaded to MIAMExpress is then processed and MAGE-ML files are created (Parkinson et al. (2005), Section 2.3.3). The data preprocessing pipeline downloads these MAGE-ML files; it performs a quality assessment of the data and preprocesses the data with a within-slide normalization. These corrected values are added to the MAGE-ML file and an updated version is sent back to EBI. If the data is curated successfully, they are stored in ArrayExpress. Each experiment gets an accession number and a password and data be-

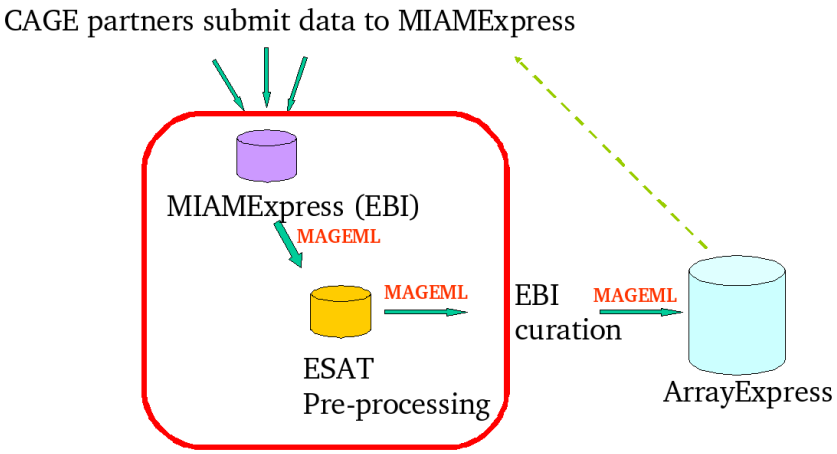


Figure 5.4: **Data flow in the CAGE project.** All partners submit their data to MIAMExpress. The MAGE-ML output is downloaded to ESAT for a quality assessment and normalization. An updated version of the MAGE-ML, completed with the preprocessed data values, is sent back to EBI. At EBI this data will be uploaded to ArrayExpress and made available to the partners.

comes accessible to the partners. About six months after the completion of the project, the data will be made publicly available. A schematic representation of the data flow within the CAGE project is shown in Figure 5.4. In the following section we will describe a few aspects of the data preprocessing pipeline into more detail.

5.3.1 RMAGEML

The data preprocessing and quality assessment part of the pipeline is written in R, calling the packages of Bioconductor. Therefore an import function for MAGE-ML files into R was required and as such function was lacking, a first important task was to provide these import tools. A Bioconductor package `RMAGEML` was written (Durinck et al. (2004)). This package extracts the information from MAGE-ML documents for two-channel

microarray experiments and maps this information to the required Bioconductor objects for the analysis of two-channel microarrays. One can choose between the `marrayRaw` object as defined within the `marray` package or the `RGlist` objects as required within the `limma` package. The information, that is extracted from the MAGE-ML files, is

- The layout of the slides (i.e., the number of grids and spots within a grid)
- Information about the probes, as for example identifiers for the corresponding genes
- The samples that were labeled and hybridized on the slides
- The available *quantitation types* (i.e., an indication of the kind of data that can be found in tab delimited data files)

An experiment of 18 hybridizations and 5184 spots takes 39 s to import on a 1.9 GHz system with 256 MB RAM. The time to load in an experiment is mainly slowed down by reading in the xml file describing the ArrayDesign. But, as within CAGE a limited number of versions of the CATMA array was used, we could speed up the data preprocessing significantly by reading in the ArrayDesign file only once. Once the necessary information was extracted from this ArrayDesign file, this information was stored in R objects and re-used for each slide with this array design.

5.3.2 Data extraction

In the specific case of our data preprocessing pipeline, the pipeline checks whether a new experiment is available at EBI and, if there is an experiment, the data (i.e., the MAGE-ML file describing the experiment and the tab delimited text files with the intensity data) are transferred to ESAT and placed into a directory. Based on the contents of this directory, the `RMAGEML` package creates an object `mageom`, a reference to the MAGE Object Model comprised in the MAGE-ML file. This is done with the function `importMAGEOM`.
From the `mageom` object, we extract the array type(s), on which the experiments are performed (i.e., different versions of the CATMA array)

with the RMAGEML functions getArrayID. And secondly, the partner to whom the experiment belongs, is extracted with the getOrganization function. Based on the array types, the different objects, describing the layout and probes on the slide, are loaded. Each partner has also its own quantitation types and based on the extracted partner information, the correct ones will be used. All the extracted information (i.e., the layout, probes description, quantitation types, and tab delimited text files with the intensities) allow to create an marrayRaw object with the RMAGEML function makeMarrayRaw. This object contains only the foreground and background intensities of each spot for both channels, but no indication on the reliability of the measurements. Therefore, also the columns with the local standard deviations of the foreground and background will be extracted separately.

5.3.3 Data preprocessing

The majority of the experiments consists out of hybridizations against the oligo reference, but for the research samples, the partners were free to opt for the classical dye-swap approach. Based on the names of the samples, the pipeline detects the design that was used and hence, which normalization is appropriate. In the case of a classical dye-swap experiment, the log-ratios were normalized with a Loess correction per print tip (Section 3.4). Alternatively, if an oligo reference was used in one of the channels, the General Linear Model normalization, as described in Section 5.2.2, was applied. So, first the oligo reference channel is corrected for primer effects and secondly, both channels are corrected separately for print-tip effects.
In both cases, these normalizations are within slide. This was the best option, as global normalization depends on the samples that are hybridized within the experiments. The sizes of the experiments that were uploaded differ a lot. For some partners, a complete biological experiment was uploaded at once (with up to 100 hybridizations); others uploaded their hybridizations two by two. Afterwards, these corrected intensities can be combined according to the analysis that is planned and data can be normalized and analyzed in depth, corresponding to the specific needs.

Each time a hybridization is preprocessed, an HTML file is generated, allowing to assess the quality of the hybridization.

5.3.4 Quality assessment

The main criteria applied to assess the quality of the data are already discussed in Section 3.1.2. As first criterion, the HTML files display for each hybridization the number and the percentage of spots above background in both channels. These numbers are computed in three different ways: (1) $\text{Fg} > \text{Bg} + 2\text{sd}(\text{Bg})$, (2) $\text{Fg} > \text{Bg} + 2\sqrt{\frac{\text{var}(\text{Bg})}{2} + \frac{\text{var}(\text{Fg})}{2}}$, and (3) $\text{Fg} > \text{Bg} + 2\text{sd}(\text{Fg})$. Typically, we expect that about $40 - 50\%$ of the spots are above background when applying the first criterion. The remaining criteria that involve the standard deviation of the foreground intensity are more stringent and less comparable between the partners.
For the hybridizations that use the oligo reference sample in one of the channels, the percentage of present calls is also computed for that channel separately. As this sample should hybridize for all probes, the signal should be above background in almost all spots. A percentage above 95% is expected.

The typical images as an image of the foreground and background intensities (e.g., Figure 3.3) and MA-plot (e.g., Figure 3.5(b) and Figure 5.2(a)) are made and displayed in the HTML files. For the hybridizations against the oligo reference, also the primer effects before and after normalization are shown, by making a boxplot of the oligo reference intensities per primer. Similarly, the print-tip effects are shown prior to normalization, by plotting the Cy3 and Cy5 $\log_2$-intensities per print tip.

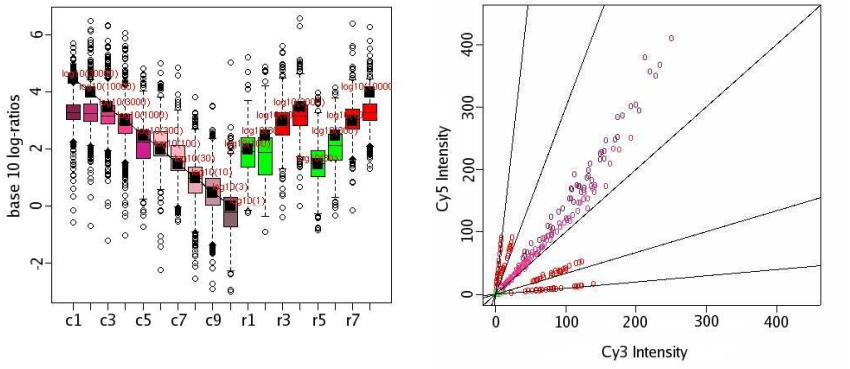
To the samples, there were also external control spikes added. These can be divided in mainly two classes: the *calibration spikes*, that measure a range of intensities in both the Cy3 and Cy5 channel, and the *ratio spikes*, that measure a ratio between the Cy3 and Cy5 channel. An overview of the different spikes is shown in Table 5.5. For the hybridizations against the oligo reference the spikes are only present in the Cy5 channel and the base 10 log-ratios were plotted along with their expected values. For the

Spike name	Cy5:Cy3 ratio	amount Cy5 pg/2μl	amount Cy3 pg/2μl
cYIR01	1:1	30,000	30,000
cYIR02	1:1	10,000	10,000
cYIR03	1:1	3,000	3,000
cYIR04	1:1	1,000	1,000
cYIR05	1:1	300	300
cYIR06	1:1	100	100
cYIR07	1:1	30	30
cYIR08	1:1	10	10
cYIR09	1:1	3	3
cYIR10	1:1	1	1
nYIR1	0	0	0
rYIR1	1:3	100	300
rYIR2	3:1	300	100
rYIR3	1:3	1,000	3,000
rYIR4	3:1	3,000	1,000
rYIR5	1:10	30	300
rYIR6	10:1	300	30
rYIR7	1:10	1,000	10,000
rYIR8	10:1	10,000	1,000

Table 5.5: **Spike controls.** This table shows the added amounts for the calibration spikes (cYIR01-cYIR10) and the ratio spikes (rYIR1-rYIR8).

remaining hybridizations, the Cy3 versus the Cy5 intensities were plotted. Examples are shown in Figure 5.5.

Next to these quality assessments on the hybridization level, also an HTML file per experiment was generated. These HTML files give an overview of the percentages of spots above background for each hybridization in the experiment and, more importantly, they provide the correlation between the technical replicates and between all hybridizations in the experiment. Between the technical replicates the correlation is expected to be around 0.90. Correlations were visualized in a *heatmap* (i.e., a color image of



(a) (b)

Figure 5.5: **The control spikes.** (a) In case of the oligo reference design, the base 10 log-ratios are plotted for all hybridizations in the experiment as box plots along with their expected values (black box). For the calibration spikes (cYIR01-cYIR10, denoted as c1, . . ., c10, respectively) a decreasing behavior is expected. The ratio spikes are denoted as r1, . . ., r8. (b) In the case of dye-swap hybridizations, the Cy3 intensities are plotted against the Cy5 intensities. We expect that the ratios for calibration spikes (with a color ranging from pink (low amount) to violet-red (high amount)) lie along the diagonal axis with slope 1 and the ratio spikes (colored green (low amount) or red (high amount)) have log ratios lying around the axes with slopes $\frac{1}{10}$, $\frac{1}{3}$, 3, or 10. The data shown in the examples are chosen from HTML files generated within the CAGE project.

the correlation matrix). This allows to detect quickly if a hybridization deviates from the other hybridizations in the experiment. An example was shown in Figure 3.4.

For all experiments, these HTML files are carefully inspected and deviating experiments or hybridizations were reported to the partner, responsible for the experiment. This allows the partner to correct and re-upload the faulty hybridizations or experiments.

5.3.5 The CAGE pipeline architecture

The technical aspects behind the CAGE preprocessing pipeline are described here. As mentioned before, all data analysis was done in *R* with the Bioconductor packages. The data preprocessing pipeline keeps track of the processed experiment, the location of the files and created HTML pages, and the hybridizations samples via a local MySQL data base.
The pipeline itself ran via a Perl script. When the Perl script was executed, the FTP site of EBI (Partner 11) was checked for new experiments, by comparing the list of processed experiments in our MySQL database with the experiments available at their FTP site. In the case of a new experiment, the Perl script downloads the data of the experiment (i.e., the MAGE-ML file describing the experiments and the tab delimited files with the intensity data). Once the data is downloaded and extracted, an *R* script is called. This script detects the new experiment and processes it, as described above. Hence, the data are normalized and we create the HTML files for quality assessment and an updated MAGE-ML file. Once this is finished, the *R* session is closed. The Perl script compresses the exported MAGE-ML, puts it back at the FTP site of the EBI, and checks for new experiments.

5.4 The CAGE data production

The data production within the CAGE project ran less smoothly than anticipated. The majority of the data arrived at ESAT during the extension period. The first MAGE-ML data were processed in the 31st month (June 2005) of the project. Half of the data reached ESAT in the last month of the extension of the project. An overview of the data production is shown in Figure 5.6. Various reasons caused this delay.
The VIB-MAF (Partner 2), which committed to print the 4,000 slides, experienced problems with their microarray printer, which forced them to suspend their array production and to change to a new printing robot. This delayed the data production for most partners, as they had to wait for their slides.
Due to the move of VIB-PSB (Partner 1) to new buildings, the plant growth rooms were too unstable to grow the plant material. The actual

hybridizations of the samples, produced by VIB-PSB, was done by VIB-MAF and they also suffered from hybridization problems. Since early 2005, the signal intensities were too low. This problem was eventually solved by switching to a new slide type.
Partner 4, RUU, together with VIB-PSB the only two big data producers in the project, only worked on the project during November 2003 until June 2005. Their contributions to the CAGE project stopped, as their post-doc resigned. From the 800 hybridizations, only 106 passed through the CAGE pipeline, and we are not sure yet whether the quality allows us to integrate this data into the compendium.
These specific problems, and perhaps a global underestimation of the work required within the CAGE project, led to a significant delay in the data production. This demanded an extension of the project with half a year, and work has to be continued even after this period.

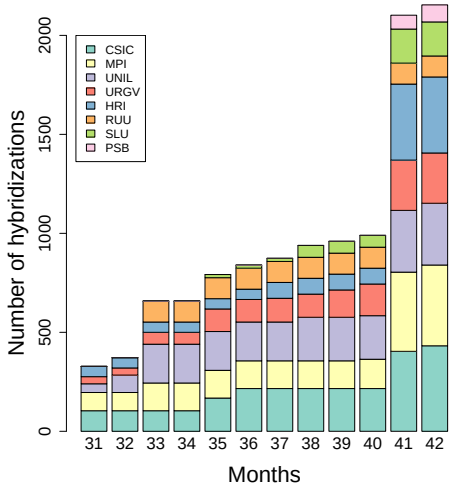
Now (July 2006), there is a significant amount of data (i.e., 2,154 hybridizations) available, which is already a nice deliverable. This data set contains a wide range of samples, which is not easy to summarize in a few sentences. Details on the produced samples can be found at `http://www.cagecompendium.org/SampleList/view_public.php`.

This data production delay prohibited us from a thorough analysis of the data, within the framework of this Ph.D. We were mainly restricted to the preprocessing of the hybridizations. An analysis of the compendium data could only start after the project was finished.

In the following section, we will give a small example that compares a time course experiment, performed by two partners in the CAGE project.

5.5 Time-course experiments on leaf development

Within the CAGE project, two partners (UNIL and MPI-MG, Partner 6 and 7) performed the same time course experiment. They produced samples of leaf 1+2 of *Columbia Arabidopsis thaliana* at the important growth stages, such that leaf development can be assessed from these data sets. The de-



Month	1 PSB	3 URGV	4 RUU	5 HRI	6 UNIL	7 MPI	8 CSIC	9 SLU	In total
31 — Jun 2005	0	36	0	52	44	92	104	0	328
32 — Jul 2005	0	36	0	52	88	92	104	0	372
33 — Aug 2005	0	60	106	52	196	140	104	0	658
34 — Sep 2005	0	60	106	52	196	140	104	0	658
35 — Oct 2005	0	114	106	52	196	140	168	16	792
36 — Nov 2005	0	114	106	52	196	140	216	16	840
37 — Dec 2005	0	120	106	80	196	140	216	16	874
38 — Jan 2006	0	117	106	80	220	140	216	60	939
39 — Feb 2006	0	138	106	80	220	140	216	60	960
40 — Mar 2006	0	160	106	80	220	148	216	60	990
41 — Apr 2006	70	254	106	384	312	400	404	172	2,102
42 — May 2006	86	254	106	384	312	408	432	172	2,154

Figure 5.6: **Data preprocessing status.** Each bar in the graph corresponds to the number of hybridizations, preprocessed by our data preprocessing pipeline, and this is plotted for each month, since the start of the CAGE project. The actual data started arriving in month 31 (June 2005). The colors correspond to the partner that produced the data. The numbers shown in the graph are listed below in the table.

velopmental stages (as defined in Boyes et al. (2001) or in Table 5.2) at which samples were harvested were

1.02 1.04 1.06 1.08 1.10 1.12 1.14 5.10 6.10.

At each time point, we have four hybridizations (i.e., two biological repeats, with each a technical replicate), except for time point 1.04. For this time point, we have 8 hybridizations: four biological replicates with each two technical replicates. All samples are hybridized against the oligo reference sample, as described in Section 5.2.1. The hybridizations were done on CATMA version 2.3 (A-MEXP-120 in ArrayExpress). The accession numbers of the experiments in ArrayExpress are E-CAGE-21 and E-CAGE-43 for MPI-MG and UNIL, respectively.
The purpose of the analysis presented in this section is to assess whether the gene expression patterns over time are conserved between the two partners. Therefore, both data sets will be analyzed separately and in a second step, the obtained expression profiles are then compared.

5.5.1 Data analysis steps

For both partners, the data were preprocessed by the CAGE preprocessing pipeline, as described in Section 5.3. In this way, data are background corrected and corrected for primer and print-tip effects.
On average, 33.6% of the spots were above background for MPI-MG, while for UNIL 44.5% of the spots were above background (according to Equation 3.1). Only those genes with more than two measurements above background over all hybridizations were retained for the analysis. This gives a set of 12,598 genes for MPI-MG and 16,221 for UNIL. The overlap between those two groups is a set of 11,347 genes. Hence, for MPI-MG, this cut-off seems to be more stringent.
The remaining values below background were imputed as follows. If we encounter such a missing value, we will search for the 10 genes that have an expression profile that is closest to the gene with the missing value(s) (according to Euclidean distance between the remaining, not missing values). We will replace the missing value by the average of these 10 not missing values. In the case that more than 50% of the values are missing, the overall average of the sample is inserted.

As explained in Section 3.3.2, the models of Wolfinger are, for computational reasons, split up into two parts, namely the *normalization model* and the *gene-specific model*. The normalization model is a global model that also normalizes the data, in this specific case, for array and sample effects. In the case of an oligo reference design, the signal in de Cy3 channel is more or less constant. We will fit the models on the log-ratio values, and can therefore omit the dye-effect. The model can be written as follows:

$$y_{ij} = \mu + \rho_i + a_j + r_{ij}, \tag{5.4}$$

where y_{ij} are the normalized $\log_2$-ratios, μ is a global average, the factor ρ_i fits the sample effects (for $i = 1, \ldots, 20$ samples), the second factor a_j is a random array effects factor ($j = 1, \ldots, 40$ arrays), and r_{ij} is the random error term. The residuals obtained from this model are then corrected for array and sample effects and on these residuals the gene-specific model will be fitted for each gene.
As gene-specific model, we model the time effect. The factor array is also added as a random factor to remove spot effects. Hence for each gene g, we have the following model:

$$r_{tjg} = \gamma_g + (\tau\gamma)_{tg} + (a\gamma)_{jg} + e_{tjg}, \tag{5.5}$$

where

- r_{tjg} corresponds to the residuals as obtained in Equation 5.4
- γ_g is the gene effect
- $(\tau\gamma)_{tg}$ is the time factor (fixed) with 9 levels $t = 1.02$, 1.04, ..., and 6.10
- $(a\gamma)_{jg}$ estimates the spot effect (random)
- e_{tjg} is the random error term

For both partners, we then select the group of genes with a significant time effect. We will use the Wald's F-test to compute significance values for this effect.

5.5.2 (Dis)agreement between CAGE partners

We will measure the agreement between the partners based on the correlations between the expression profiles. If we consider the group of 11,347 genes, that were shared by the two partners, the average correlation between the expression profiles is low (i.e., mean correlation of 0.196 and median of 0.217). To get an idea of the distribution of the correlations, a histogram of the correlations is shown in Figure 5.7(a). However, these genes are not all related to the leaf development process.
By restricting the data set to the set of genes that reach a certain significance level α for the time effect, the well-correlated genes are retained. If, for example, we retain those genes with a p-value smaller than 0.001 for both partners, we retain a set of 2,164 genes. Their correlation equals on average 0.549 (median correlation = 0.662). A histogram of the distribution of the correlations is shown in Figure 5.7(b). Applying a Benjamini-Hochberg (see Section 4.10) correction and selecting those genes that have a corrected p-value smaller than 0.001, further restricts the data set and increases the mean and median correlation to 0.592 and 0.701, respectively.
These results can also be seen by plotting histograms of the negative $\log_2$ p-values, according to the correlation, as in Figure 5.8. In this graph, both partners show an increasing trend (i.e., p-values become more significant) as the correlation increases. Hence, we demonstrate that expression profiles for the genes of interest (i.e., involved in the leaf development process) are conserved between two partners.

5.6 High-light stress on catalase-deficient plants

Only for a few specific experiments, a more detailed analysis could be done. In the following section, we will focus on one experiment, done in collaboration with VIB-PSB. The experiment aims to assess the influence of high light on catalase-deficient plants.

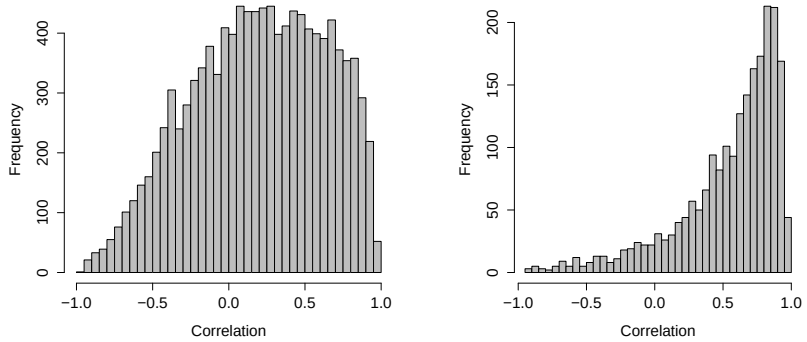


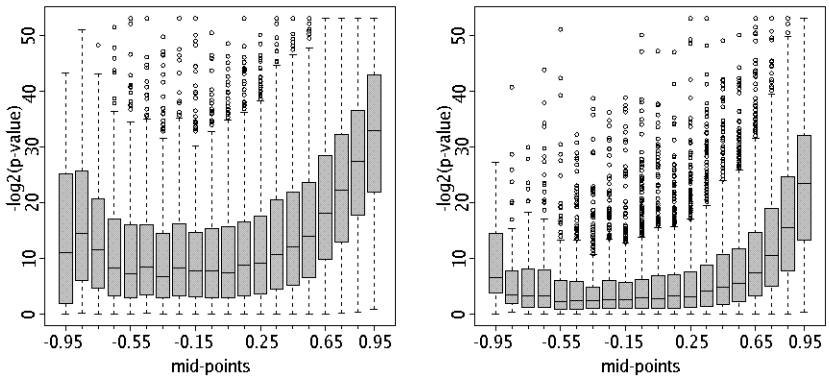
Figure 5.7: **Correlation between expression profiles for UNIL and MPI-MG.** (a) The histogram shows the distribution of the correlation between the expression profiles of UNIL and MPI-MG, for all genes in common (i.e., 11,347 genes). (b) If we restrict to those genes with a significant time effect, we retain the well-correlated genes.

5.6.1 The context of the experiment

Hydrogen peroxide (H_2O_2) is a chemical that is formed naturally by organisms through their metabolism. It is toxic and therefore, it is essential for plant survival that H_2O_2 is decomposed quickly into other, less dangerous, chemicals. Any abundant H_2O_2 production is mainly counteracted by the enzyme *catalase*. Catalase converts H_2O_2 into harmless water and oxygen:

$$2H_2O_2 \rightarrow 2H_2O + O_2.$$

One of the processes that can increase the H_2O_2 level in plants is *photorespiration*. During this process the plant takes up oxygen and it releases carbon dioxide (CO_2). Thereby it reduces the yield of photosynthesis—less O_2 is formed. Photorespiration takes place when the oxygen levels in the leafs are relatively high compared to carbon dioxide levels. Therefore, the process occurs on hot and dry days. In this situation, the *stomata* (i.e., minute breathing pores of leaves) are closed to prevent dehydration, but



(a) UNIL (b) MPI-MG

Figure 5.8: **Boxplot of the negative log p-values, per correlation between the expression profiles.** The expression profiles are grouped in 20 classes, according to their correlation. For each group the boxplot of their corresponding negative, base 2 log transformed p-values is plotted. (a) UNIL. (b) MPI-MG.

in the meantime, this also prevents CO_2 from entering. Hence, the O_2 concentration in the leafs will exceed the CO_2 concentration.
During high-light stress (HL), Noctor et al. (2002) showed that photorespiration is the main source of H_2O_2 produced in the plant. But thanks to catalase, this abundant H_2O_2 is removed and therefore the plant does not suffer from the high-light stress. However, in *catalase-deficient plants* (i.e., plants with a reduced catalase activity level) high-light stress can induce spontaneous cell death by insufficient removal of H_2O_2.

At VIB-PSB, several experiments were done to assess the influence of high-light stress on the catalase-deficient plants (Vandenabeele et al. (2003, 2004); Vanderauwera et al. (2005)). For example, in Vandenabeele et al. (2004) the influence of high-light stress was assessed for catalase-deficient plants after high-light irradiation for $0h$, $3h$, $8h$, and $23h$. For the catalase-deficient plants, cell death was already visible after

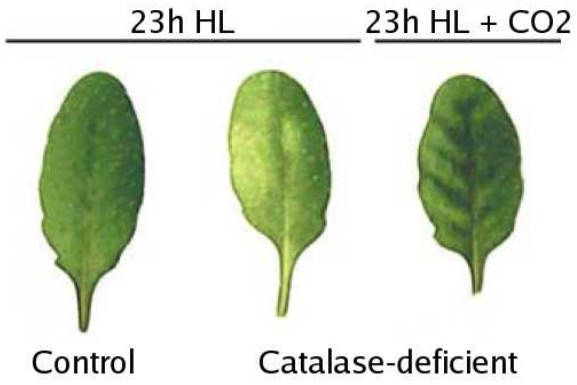


Figure 5.9: **Cell death in catalase-deficient plants after high-light stress.** The figure shows middle-aged leaves of *Arabidopsis* of control and catalase-deficient plants that were exposed to high-light stress for $23h$. When applying high levels of CO_2, the induction of cell death in the catalase-deficient plants is counteracted. This figure was obtained from Vandenabeele et al. (2004).

8 hours of high-light stress, while the control plants remained healthy. However, as one can expect, under high concentrations of CO_2 the induction of cell death in the catalase-deficient plants due to high-light stress was weakened (see Figure 5.9). In Vandenabeele et al. (2004), this effect was mentioned, but no microarray experiment was performed to assess the influence of this increased level of CO_2 at gene expression level.

The experiment described here will compare the effects of high-light stress between catalase-deficient *Arabidopsis* plants and control plants, and this with and without the addition of carbon dioxide.

5.6.2 Design of the experiment

In *Arabidopsis*, the catalase family consists of three genes (i.e., At1g20630, At4g35090, and At1g20620, or cat1, cat2, and cat3, respectively). As cat2 is usually highly expressed in leafs, transgenic *Arabidopsis* plants

were grown with decreased levels of cat2 (Vandenabeele et al. (2004)). In this particular experiment, catalase-deficient *Arabidopsis Col* plants (CAT2HP1) were grown with 20% of total residual catalase activity.
In the experiments mentioned above, the high-light stress was applied during $0h$, $3h$, $8h$, and $23h$, and in the catalase-deficient plants cell death appeared to be induced already after $8h$ of high light. Therefore, the last time point at $23h$ was excluded from this analysis. Also, to assess the early responders, instead of $3h$, samples were collected a first time at $1h$ of high-light stress. Hence, high-light irradiation was applied for $0h$, $1h$, and $8h$.
This comparison was not only done under the normal air conditions (i.e., CO_2 level of 400 ppm[1] and 21% O_2), but also under high levels of CO_2 (i.e., 1500 ppm and 21% O_2).
For each condition, leaf material of 6 week-old plants was harvested and RNA was extracted of pools of leaves of 20 up to 30 plants. Two biological replicates were taken under each condition and hybridized twice, such that we have for each condition two biological replicates with each two technical replicates. In total, the experiment comprises 48 hybridizations. The design of the experiment is summarized in Table 5.6. The hybridizations were done on the CATMA slide version 2.3 (A-MEXP-120 at ArrayExpress).

5.6.3 Data analysis steps

For this experiment, the percentage of spots above background (according to Equation 3.1) equals on average 39.54% (ranging between 28.6% and 59.1%). The correlation between the technical repeats of the raw data ranges between 0.854 and 0.940 with an average of 0.9082. Careful inspection of the HTML files, generated by the tools of the CAGE pipeline, gave no indication of severe problems with quality. Therefore, the quality of the slides is not brilliant, but fair enough.
The data analysis is similar to the data analysis steps, as described in Section 5.5.1. Again, we start from the data as preprocessed in the CAGE preprocessing pipeline (see Section 5.3). The base 2 log-ratios, computed from these residuals were reasonably normally distributed (see

[1] ppm= parts per million

HL Time	Control		CAT2HP1	
		high CO_2		high CO_2
0h HL	2 biol with 2 tech repl	2 biol with 2 tech repl	2 biol with 2 tech repl	2 biol with 2 tech repl
1h HL	2 biol with 2 tech repl	2 biol with 2 tech repl	2 biol with 2 tech repl	2 biol with 2 tech repl
8h HL	2 biol with 2 tech repl	2 biol with 2 tech repl	2 biol with 2 tech repl	2 biol with 2 tech repl

Table 5.6: **Design of the catalase experiment.** Catalase-deficient plants (CAT2HP1) are compared to normal control plants after high-light irradiation for $0h$, $1h$, and $8h$. This was done under normal levels and under high levels of CO_2 (i.e., 400 ppm and 1500 ppm, respectively). For each condition, two biological replicates with each two technical replicates were produced.

Figure 5.10) and on these log-ratios the Wolfinger models, as introduced in Section 3.3.2, will be fitted.

Prior to the fitting of the Wolfinger models, missing values were also imputed, as described in Section 5.5. Of course, the Wolfinger models deviate from the models in Section 5.5, as we have now a completely different design. The normalization model can be written as follows:

$$y_{ij} = \mu + \rho_i + a_j + r_{ij}, \tag{5.6}$$

where y_{ij} are the normalized $\log_2$-ratios, μ is a global average, the factor ρ_i fits the sample effects (for $i = 1, \ldots, 24$ samples), the second factor a_j is a random array effects factor ($j = 1, \ldots, 48$ arrays), and r_{ij} is the random error term. As a gene-specific model, we choose for a complete model including three factors for the effects of duration of the high-light irradiation, the genotype effect (i.e., control versus catalase-deficient plants), and the CO_2 factor to estimate the effect of high versus normal concentrations of carbon dioxide, and we include the interaction between these different factors. The factor array is also added as a random factor to remove spot

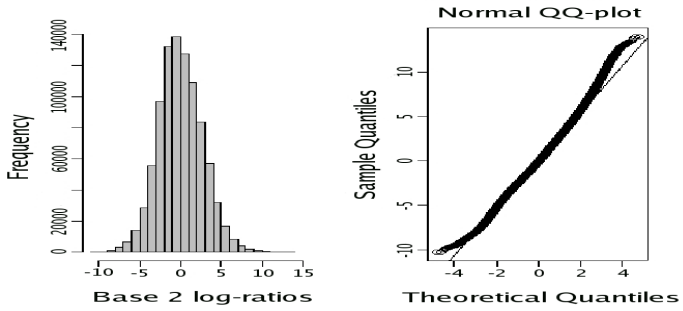


Figure 5.10: **Histogram and normal quantile plot of the log-ratios.** The histogram and normal quantile plot show the distribution of the $\log_2$-ratios as obtained with the CAGE preprocessing tools (i.e., background corrected and normalized for primer and print-tip effects).

effects. Hence for each gene i, we have the following model:

$$\begin{aligned} r_{itgcj} \quad = \quad & \gamma_i + (\tau\gamma)_{ti} + (\chi\gamma)_{gi} + (\omega\gamma)_{ci} + \\ & (\tau\chi\gamma)_{tgi} + (\chi\omega\gamma)_{gci} + (\tau\chi\gamma)_{tci} + \\ & (\tau\chi\omega\gamma)_{tgci} + (a\gamma)_{ji} + e_{itgcj}, \end{aligned} \tag{5.7}$$

where

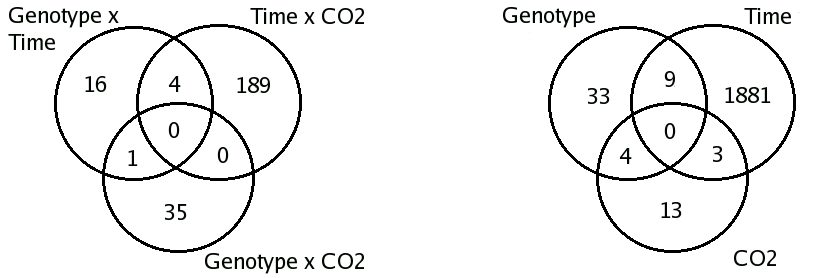
- r_{itgcj} corresponds to the residuals as obtained in Equation 5.6
- γ_i is the gene effect for gene i
- $(\tau\gamma)_{ti}$ is the Time factor (fixed) for gene i with 3 levels $t = 0h, 1h,$ and $8h$
- $(\chi\gamma)_{gi}$ estimates the Genotype effect (fixed) for gene i with 2 levels $g =$ control, CAT2HP1
- $(\omega\gamma)_{ci}$ measures the CO_2 effect (fixed) for gene i with 2 levels $c =$ Not, $+CO_2$
- $(a\gamma)_{ji}$ estimates the spot effect (random)
- e_{itgcj} is the random error term

5.6.4 Differentially expressed genes

A first group of genes that we select are genes with a significant interaction effect between all three factors (i.e., Time x Genotype x CO_2). For these genes, information on all three factors is required to make some predictions on their expression levels. We will use the Wald's F-test to decide whether this triple interaction effect is significant. Instead of applying corrections for multiple testing, we make the threshold for significantly expressed genes stringent by calling a gene differentially expressed if $p < 0.001$. In this way we select a set of 137 genes.
Similarly, for the remaining genes, we can also test for significant interactions between two factors (i.e., Time x Genotype, Time x CO_2, or Genotype x CO_2). In this way an additional set of 245 genes is selected. The numbers of genes for the different interactions are shown in Figure 5.11(a). For example, there are 189 genes with only a significant Time x CO_2 interaction, meaning that the effect of high-light irradiation depends on the CO_2 concentration.
Hence, for the genes with one or more interaction effects (382 genes), it does not make sense to investigate the main effects of Time, Genotype, or CO_2. Therefore, we will treat this group separately and search for significant main effects solely for the remaining group of genes. With the same tests, we find in total 1,943 genes with one or more significant main effects. The numbers per factor are shown in Figure 5.11(b).

5.6.5 The expression profiles

The expression profiles will be presented as clusters. Many clustering algorithms exist. In this case, we will employ *Adaptive quality-based clustering* or AQBC (De Smet et al. (2002)). This clustering technique has the nice advantage that no knowledge about unpredictable parameters, as, for example the number of clusters, is required. Instead, two more intuitive parameters have to be chosen, namely the minimal probability of a gene belonging to cluster (between 0.5 and 1) and the minimal number of genes in a cluster. As a required probability, 0.95 was chosen and the minimal number of genes in a cluster was chosen equal to two.
The first group of genes that was clustered, are the genes with one or more

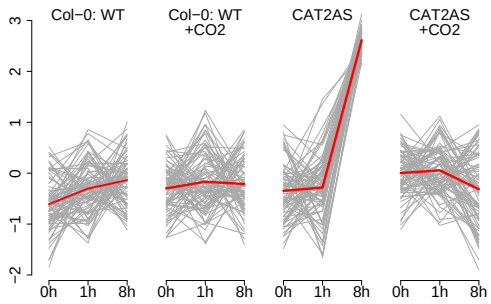


(a) Significant interaction terms (b) Significant main effects

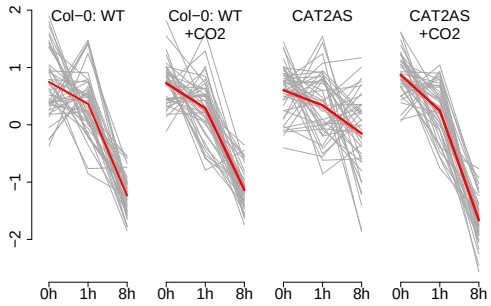
Figure 5.11: **Significant double interaction and main effects terms.** (a) The Venn diagram shows for each combination of two factors the number of significant interactions, after excluding those genes with a significant triple interaction. In total, this is a set of 245 genes. (b) The number of genes with significant main effects and their overlap are shown (after omitting those genes with one or more significant interaction terms). In total, this is a set of 1,943 genes.

significant interaction terms. Their expression patterns are standardized (i.e., with mean 0 and standard deviation 1). Clustering this set of 382 genes resulted in 34 clusters. The first three major clusters with 20 or more elements are shown in Figure 5.12. In the figure, the expression profiles are split according to their genotype and CO_2 concentration. For example, the first cluster consists of genes that are strongly up-regulated for catalase-deficient *Arabidopsis* plants. Under a high concentration of CO_2 however, the gene expression of this group of genes becomes again comparable with the expression pattern measured for the control plants. This indicates that this group of genes are induced specifically by photorespiratoric H_2O_2 (i.e., H_2O_2 that is produced by photorespiration).

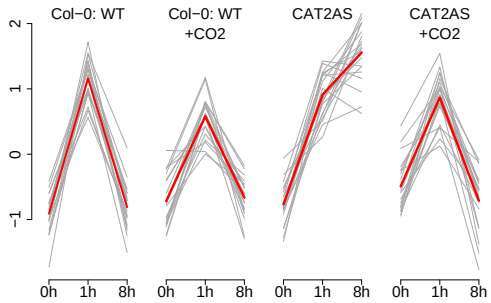
Similarly, the genes with a significant main time effect (1,893 genes) can be clustered. Under the same choice of parameters, this resulted in 24 clusters. The three largest clusters can be found in Figure 5.13. In these clusters, genes can be found with a similar gene expression over time, independent of the genotype and the CO_2 concentration.



(a) Cluster 1 (79 genes)



(b) Cluster 2 (42 genes)



(c) Cluster 3 (20 genes)

Figure 5.12: **Clustering genes with significant interaction terms.** The genes with a significant interaction term (382 genes in total) are clustered with AQBC clustering. The three largest clusters are displayed in the figure.

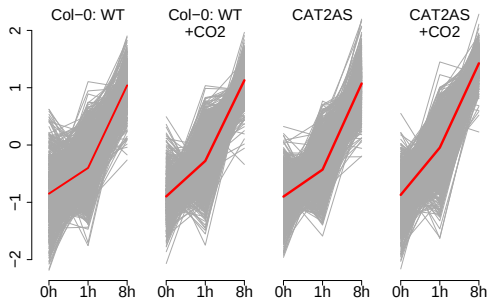
The biological validation is still ongoing. The effects of high light in the catalase-deficient plants were comparable to the results found in Vanderauwera et al. (2005), which describes a similar experiment done on Affymetrix chips. The effect of high CO_2 concentrations was however new in this study and the affected genes still have to be assessed.

5.7 Conclusion

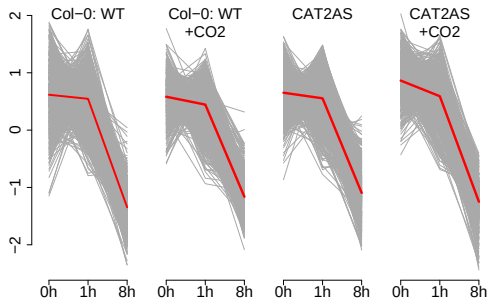
The *Compendium of Arabidopsis Gene Expression* or CAGE project was a European demonstration project, involving eight data producing and two bioinformatics partners, that aimed at building an atlas of gene expression of *Arabidopsis thaliana*, under a variety of conditions. Our main contributions to the project was the development of a data preprocessing pipeline. One part of the pipeline, namely the communication between data stored in MAGE-ML format and a statistical environment as R and its Bioconductor packages, resulted in an on its own standing Bioconductor package, called `RMAGEML`.

The use of the oligo reference as a design had implications towards data preprocessing and made the classical Loess normalization inappropriate. Therefore, next to Loess normalization, normalization with General Linear models, specific for this design, was implemented in the CAGE pipeline.

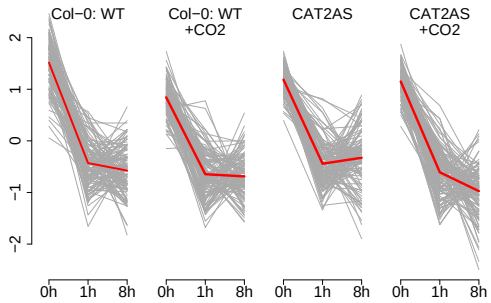
The CAGE pipeline is probably the first pipeline that incorporated the use of MAGE-ML files and therefore, it suffered from the inaccuracy of the MAGE-ML files. Although the pipeline was automatic, it broke down many times, due to inconsistencies in the MAGE-ML files, as, for example, changed quantitation types, parts of the MAGE-ML that 'duplicated', and wrong name of the oligo reference sample. These were probably caused by the fact that MAGE-ML is perhaps too complex and too flexible. And even, if data was preprocessed in the end, errors were detected, via the quality assessment of the HTML file generated by the pipeline, as, for example, Cy3 and Cy5 intensities that were swapped, hybridizations that got mixed up, or even the order of the intensities in the tab delimited files got



(a) Cluster 1 (785 genes)



(b) Cluster 2 (581 genes)



(c) Cluster 3 (120 genes)

Figure 5.13: **Clustering genes with significant interaction terms.** The genes with a significant Time term (1,893 genes in total) are clustered with AQBC clustering. The three largest clusters are shown in the figure.

switched and intensities did not correspond anymore to the correct genes. Or, data had to be rescanned, due to ill-chosen scanner setting, leading to saturation. We played an important role in flagging these problems and in ameliorating the data set as much as possible. Though manual intervention was often required, the pipeline was able to follow the data production.

As the data production within CAGE was a lot slower than anticipated (i.e., half of the data set (more than 1,000 hybridizations) have been preprocessed in the last month of the extension), an in depth data analysis on the data produced within the CAGE project became impossible within this time frame and data analysis of the compendium will still continue for some time after the project. However, a short preview of a comparison between two partners, that performed the same time course experiment of leaf development, looks promising.

The CAGE project was an excellent opportunity for setting up closer collaborations with the partners, involved in the CAGE project. The Bioconductor training session, organized by our group to make the different partner acquainted with the Bioconductor tools, were a stimulation for closer collaborations. A number of dedicated, research experiments have been analyzed together in collaboration with specific CAGE partners and are now waiting for biological validation. We have discussed here, as an example, an experiment in collaboration with VIB-PSB, on the influence of high-light stress on catalase-deficient plants, under a normal and a high concentrations of CO_2.

Chapter 6

Analysis of loop design experiments for Array CGH

As introduced in Section 2.5, Array Comparative Genomic Hybridization or array CGH allows us to detect for DNA copy number aberrations with a high resolution. In this chapter, we present a tool to analyze array CGH experiments in which three patients are compared in a loop design. This has important advantages, compared to the classical set-up in which a test patient is compared to a normal, reference patient on two slides (including a dye-swap). We will elaborate on the choice of this design and compare two analysis approaches. Based on the signal-to-noise ratio and false positive and false negative rates, we will decide which method is preferable. This method is then implemented in a web based tool.

6.1 Array CGH to detect chromosomal aberrations

Array Comparative Genomic Hybridization (array CGH) uses a genomic DNA microarray to detect copy-number aberrations and variations at high

resolution on a genomewide scale. Compared to classical karyotyping, it offers a resolution between 10kb and 1Mb, instead of about 5Mb.

6.1.1 The basic principle of array CGH experiments

The most frequent experimental setup for array CGH consists in comparing genomic DNA of a patient (test) with that of a normal individual (reference) using a two-channel microarray consisting of DNA segments spread across the whole genome. DNA from the test and reference samples is extracted, labeled with different fluorescent dyes (usually Cy3 and Cy5), hybridized to the microarray, and then scanned by two-channel fluorescence. Aneuploid chromosomal regions are detected as probes with a deviant log ratio of the intensities of the test against reference signal (approximately $\log_2(1/2)$ for a deletion and $\log_2(3/2)$ for a duplication). Usually the experiment is repeated in a dye-swap with the fluorescent labeling of test and reference exchanged. The signals are then averaged over the dye-swap replicates to reduce the signal-to-noise ratio.

The array CGH probes that are used in this work are PCR amplified *Bacterial Artificial Chromosomes* (BAC) clones. All analysis shown in this work has been done in close collaboration with Centrum Menselijke Erfelijkheid Leuven (CME-UZ). They use a 3k array, which guarantees a genomewide coverage with a $\sim$1Mb resolution, but they aim at switching to a 32k BAC clone array (Ishkanian et al. (2004)). This tiling set of clones will increase the resolution with 10 fold. All analysis shown here is performed on the 3k BAC clone array. However, the discussion applies equally to spotted long oligo platforms and, therefore, we will refer to our probes as *targets*.

6.1.2 Alternative technologies

Several laboratories have also used *cDNA arrays*, designed for expression profiling, as an alternative technique. Advantage is that deletions or duplications are then directly mapped to a gene instead of a position on the

genome. But this method cannot compete with the current platforms in terms of achievable resolution.
Another, more important technique are SNP genotyping platforms. A *single nucleotide polymorphism* (SNP) is a DNA sequence variation occurring when a single nucleotide (A, C, G, or T) in the genome differs between members of the species. SNPs make up 90% of all human genetic variations and is thereby the most common genetic variation in the chromosome. *Single nucleotide polymorphism comparative genomic hybridization* (SNP-CGH) places oligonucleotide SNPs on an array. They allow to simultaneously genotype thousands of SNP markers across the human genome, whereas array CGH methods are unable to query genomic DNA on an allele-specific basis.

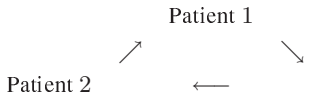
6.1.3 Applications

Array CGH allows the analysis of patients with constitutional and acquired genetic disorders. These analyses lead to the detection of disease causing genomic imbalances. Array CGH can be applied for prenatal screening by analysis of fetal DNA. Recent developments focus on the reducing the required amount of DNA, such that array CGH can detect chromosomal imbalances from a single cell. This can be important for the diagnosis in preimplantation embryos (Le Caignec et al. (2006)).
Also, in cancer, array CGH has important applications. The comparison of differentially labeled normal and pathological DNA from tumors against reference DNA identifies segmental genomic alterations. These DNA copy number gains and losses allow to detect regions associated with cancer and to validate candidate cancer-related genes. The high resolution of the technique also leads to the detection of small, novel alterations that may be important for the disease, but which are not detectable by lower-resolution techniques.

6.2 A loop design for array CGH

The alternative design proposed here is a loop design in which three hybridizations are carried out with three test patients that are compared with each other: Patient 1 versus Patient 2, Patient 2 versus Patient 3, and

Patient 3 versus Patient 1, as shown in Figure 6.1.
This design measures the intensities of three test samples in a statistically balanced way and requires no normal reference sample. Hence, only three arrays are used to analyze three patients and to obtain two measurements for each of them. For the classical dye-swap design, half of the resources are consumed to measure the reference sample of a normal individual and, therefore, six arrays would be necessary to obtain as many measurements from the test samples.
Extensive genomic variation (called copy number variation) is also present in normal individuals. So, in the classical dye-swap design, a deviant log ratio for one target in the test sample could just as well be associated with the reciprocal target in the reference sample. The difficulty in disambiguating deviations between the test and reference sample prevents us also from replacing the reference sample with a second test sample in the dye-swap design. The loop design, on the contrary, unambiguously associates a deviation to the correct sample by looking for a unique pattern of log ratios. For example, a duplication in Patient 1 will be associated approximately to a positive log ratio in the Patient 1 vs. Patient 2 hybridization, a negative log ratio in the Patient 3 vs. Patient 1 hybridization, and a null log ratio in the Patient 2 vs. Patient 3 hybridization. No deletion or duplication in another patient will display the same pattern, so the association is unambiguous.



	Cy 5	Cy3
Slide 1	Patient 1	Patient 3
Slide 2	Patient 2	Patient 1
Slide 3	Patient 3	Patient 2

Figure 6.1: **The loop design.** Schematic overview of the loop design, in which three patients are compared. The table shows the three hybridizations. This numbering of the slides will be used throughout the text.

6.3 The different analysis methods

To analyze such a loop design, a number of approaches are possible. In a first step, prior to fitting the loop design and to estimate the contrasts between the patients, we will preprocess the data to remove dye effects.
A first method that we will apply to detect aberrant targets, is the mixed model approach as proposed by Wolfinger (Section 3.3.2), on the absolute intensities. Secondly, we will fit the LIMMA models (Section 3.3.1) on the $\log_2$-ratio intensities. These two methods provide two different ways to estimate for each target the contrasts between the different patients in the loop design. We compare both methods on a test data set and we will implement the method with the best signal-to-noise-ratio as a user-friendly web application.

6.3.1 Preprocessing

Prior to all methods explained below, basic preprocessing steps were performed. The spot intensities were corrected for local background and only those spots with a signal above background, according to Equation 3.1, were retained for the analysis. In this way, only few spots are lost, as almost all spots are above background (on average 96.6%, computed over 19 loop designs or 57 hybridizations). The ratios of the Cy5 to the Cy3 intensities were computed for each target and base 2 log transformed. The log-ratios are normalized using a *spatial Loess normalization*, in which one applies a loess regression to fit the $\log_2$-ratios (M-values) on the coordinates on the slide as predictor variables.

6.3.2 Mixed model approach

Firstly, we will apply the mixed model approach as proposed by Wolfinger et al. (2001) (see Section 3.3.2). The models were described to analyze cDNA microarrays, but they are generally valid and can be applied for the analysis of CGH array in a straightforward manner. As the mixed models, proposed in Wolfinger et al. (2001), estimate effects on the absolute intensities, instead of the $\log_2$-ratios, we will extract the corrected Cy3 and Cy5 intensities from the spatial Loess corrected $\log_2$-ratios (M) and the average

$\log_2$ intensity values (A), as

$$\begin{cases} \text{Green} = (-M + 2A)/2, \text{and} \\ \text{Red} = (M + 2A)/2. \end{cases} \tag{6.1}$$

Again, the mixed models as proposed in Wolfinger et al. (2001) consist out of two submodels, the normalization model and the target-specific model.

The normalization model

This model will normalize for array, dye, and patient effects. The fitted model can be written as follows:

$$y_{cij} = \mu + \tau_i + a_j + (\tau a)_{ij} + r_{cij},$$

where y_{cij} are the Cy3 and Cy5 intensities, as computed in Equation 6.1, μ is the overall average, τ_i is the fixed patient effect with three levels ($i =$ patient $1, 2, 3$), a_j estimates the random array effect (also with three levels $j =$ array $1, 2,$ or 3), and $(\tau a)_{ij}$ fits the interaction effect between the patient and array effect, and in this way it also corrects for the dye effect.

The target-specific model

For each target, we extract the residuals r_{cij} from the normalization model and fit a target-specific model:

$$r_{cij} = \kappa_c + (\kappa\tau)_{ci} + (\kappa a)_{cj} + e_{cij},$$

where r_{cij} are the residuals obtained from the normalization model, κ_c is the overall average for Target or Clone c, $(\kappa\tau)_{ci}$ is the fixed patient effect for Target c with three levels ($i =$ patient $1, 2, 3$), $(\kappa a)_{cj}$ estimates the spot effect for Target c and Array j, and e_{cij} fits the random error effect.

Finding the duplicated or deleted targets

Our main interest is in the estimates of the $(\kappa\tau)_{ci}$ effects, which reflect the difference between the patients for Target c. Specifically, we assess whether the contrasts Patient 2 vs. Patient 1 $(= (\kappa\tau)_{c2} - (\kappa\tau)_{c1})$ and

Classification	Log-Ratio patient$_1$ / patient$_3$	Log-Ratio patient$_2$ / patient$_1$
Duplication for patient$_1$	positive	negative
Duplication for patient$_2$	0	positive
Duplication for patient$_3$	negative	0
Deletion for patient$_1$	negative	positive
Deletion for patient$_2$	0	negative
Deletion for patient$_3$	positive	0

Table 6.1: **Classification of the targets.** Each target is classified as upregulated (*positive*), downregulated (*negative*), or not differentially expressed (0) for all three contrasts (i.e., patient 1 versus 3, patient 2 versus 1, and patient 3 versus 2). Based on these results a target can be recognized as duplicated (deleted) for Patient 1, 2, or 3, according to this scheme.

Patient 1 vs. Patient 3 $(= (\kappa\tau)_{c1} - (\kappa\tau)_{c3})$ are equal to zero with a Wald's F-test. In the case where the contrast is significantly larger than zero for a chosen significance level α, we call this contrast *positive*. In the case where it is smaller than zero, it is called *negative*. Else we assign 0. Based on this hypothesis testing, the targets are classified as duplicated or deleted according to the classification shown in Table 6.1. For example, if the contrast Patient 1 vs. Patient 3 $(= (\kappa\tau)_{c1} - (\kappa\tau)_{c3})$ is positive and the contrast Patient 2 vs. Patient 1 $(= (\kappa\tau)_{c2} - (\kappa\tau)_{c1})$ is negative for a target, then this target is likely to be duplicated for Patient 1. In some rare cases, we obtain as result a target that has, for example, a negative value for both contrasts $(\kappa\tau)_{c1} - (\kappa\tau)_{c3}$ and $(\kappa\tau)_{c2} - (\kappa\tau)_{c1}$, which is none of the combinations in Table 6.1. In this case, we call the target *strange*.

6.3.3 The LIMMA approach

An alternative statistical tool is a linear model of the log ratios LIMMA (Section 3.3.1, Smyth (2004)). In contrast to the mixed models, this technique fits the 2D spatial Loess corrected $\log_2$-ratios directly. As a model,

we can apply the model as described in Equation 3.14. In this particular case, we can choose the following contrasts $C_{c1} = \log_2(P_{c1}/P_{c3})$ and $C_{c2} = \log_2(P_{c2}/P_{c1})$, where P_{ci} corresponds to the true underlying signal for Target c for Patient i. These contrasts correspond to the samples that were directly compared on the first two slides in Figure 6.1 and the observed log ratios on these slides should on average be equal to the contrast. The data of the third slide should then correspond on average to $C_{c3} = \log_2(P_{c3}/P_{c2}) = -C_{c1} - C_{c2}$. The linear model that fits the data can be written as

$$E\begin{pmatrix} y_{c1} \\ y_{c2} \\ y_{c3} \end{pmatrix} = X\mathrm{C}_c = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} C_{c1} \\ C_{c2} \end{pmatrix},$$

where E denotes the expectation of a random variable, X is the matrix of linear dependencies, C_c is the vector of contrasts for Target c, and y_{ci} denotes the $\log_2$ ratio for Target c measured on the i^{th} slide. For each target, the least squares estimates of the three contrasts are obtained. To classify the contrasts as significantly up-, downregulated, or not differentially expressed, we apply the moderated t-statistic as implemented in LIMMA (Section 3.3.1). The p-values from the moderated t-test were corrected to control the false discovery rate with Benjamini-Hochberg (Benjamini and Hochberg (1995) or Section 4.10). Similarly to the mixed models approach, we can detect targets that are duplicated or deleted for a patient, based on the p-values of the contrasts. For a chosen cut-off value α, we decide whether a target is not differentially expressed (0), upregulated (*positive*), or downregulated (*negative*) for a contrast. Based on the two contrasts, we can again classify a target as duplicated or deleted for a patient according to Table 6.1. Again, we can obtain *strange* targets.

6.4 Benchmarking the mixed models and LIMMA approach

Both methods, mixed models and LIMMA approach, provide two distinct ways to analyze the loop design experiments. To decide which method is preferable, we will first check which estimation method separates the

aberrant and the non-aberrant targets best. This will already give a hint to which method is preferred. Secondly, we will compare the false positive and false negative rates for a number of cut-off values α. Based on this information, we will decide which method to use.

6.4.1 The test data set

For the comparison of the analysis approaches, will consider a data set, consisting out of nine loop designs. In 15 out of these 27 patients, previously large deletion or duplication regions were detected by by the Center for Human Genetics in Leuven. The aberrations were detected using basic statistics in excel (Vermeesch et al. (2005)). Subsequent to the excel analysis, a region is scored aberrant, if one target passes the threshold of $4 \times SD$ and if two or more flanking targets were passing the threshold of $\log_2\left(\frac{3}{2}\right) - 2 \times SD$ as described in Vermeesch et al. (2005). If a deletion or a duplication larger than 3Mb was detected, FISH was performed to confirm the results of the array. In case of a duplication smaller than 3Mb we opted to perform a quantitative PCR (qPCR) experiment. In total, 15 out of 27 patients show one or multiple targets anomalies where as 12 patients are apparently normal due to the results of the array.
The array CGH slides that are used in these loop design experiments, contain two copies of each target. We will combine these six Cy3 and six Cy5 measurements.
A short summary of the data set is shown in Table 6.2. In total, this data set comprises 635 aberrant targets: 274 deleted and 361 duplicated targets. Two experiments (i.e., Experiments 1 and 9) include a sex mismatch. As for those experiments, the Y chromosome is absent for at least one of the patients, the measurements on the Y chromosome were excluded for both experiments from all computations. Because the X chromosome has regions with chromosome Y homology, the intensity ratios of chromosome X targets are also more variable for those patients, and hence also the X chromosome was excluded from the computations for experiment 1 and 9. In total, this reduced data set comprises 328 aberrant targets: 116 deleted and 212 duplicated targets. Over all 9 experiments, we have a set of 30,668 measurements for non-aberrant targets.

Experiment	Patient	Deletions/duplications	length
1	1	deletion on 13	25
	2	duplication of X	149
2	1	duplication of 18	102
3	1	deletion on 10	6
	2	duplication on 7	20
	3	duplication on 15	43
4	2	deletion on 4	15
5	1	deletion on 12	14
6	1	deletion on 9	6
	3	deletion on 12	7
7	2	duplication on 5	13
	2	deletion on 18	16
8	1	duplication on 13	15
	1	deletion on 13	12
9	1	duplication on 7	11
	1	deletion on 7	15
	2	deletion of X	158
	3	duplication on 21	8

Table 6.2: **Loop design test data set.** The nine loop design experiments, listed in the table, will be used as a test data set to compare the two methods, LIMMA and mixed models. For each aberration present in the experiment the number of targets on the slide lying in the deleted or duplicated region is indicated.

6.4.2 Signal-to-noise ratios

Assessing which method is best capable of distinguishing between the intensities of aberrant and non-aberrant targets can be done by computing *signal-to-noise-ratios* (SN) (for both deletions and duplications separately) as

$$SN_{\text{dupl/del}} = \frac{|\text{mean}_{\text{dupl/del}} - \text{mean}_{\text{non-aberrant}}|}{\sqrt{\frac{1}{2}\left(\text{var}_{\text{dupl/del}} + \text{var}_{\text{non-aberrant}}\right)}}.$$

	Mixed model	Linear model
$N_{\text{non-aberrant}}$	30,668	30,668
$\text{mean}_{\text{non-aberrant}}$	0.00701	-0.00064
$\text{s.d.}_{\text{non-aberrant}}$	0.11306	0.06914
N_{dupl}	212	212
$\text{mean}_{\text{dupl}}$	0.48515	0.48777
$\text{s.d.}_{\text{dupl}}$	0.12920	0.11967
$\mathbf{SN_{dupl}}$	**3.93861**	**4.99782**
N_{del}	116	116
mean_{del}	0.76779	0.78468
s.d._{del}	0.22275	0.18787
$\mathbf{SN_{del}}$	**4.30702**	**5.54778**

Table 6.3: **Signal-to-noise ratios.** For the three methods (i.e., taking the average over the measurements, the mixed models, and the LIMMA approach), the number of targets, average, and standard deviation (s.d.) of the $\log_2$-ratios are given for the non-aberrant, duplicated, and deleted targets. Based on these numbers the signal-to-noise ratios are computed.

As we have collected a data set with 212 duplicated targets, 116 deleted, and 30,668 non-aberrant targets, we can compute the SN values based on the absolute values of the contrasts Patient 1 vs. Patient 3 and Patient 2 vs. Patient 1, for both LIMMA and the mixed model. The results are shown in Table 6.3.
As could be expected, the signal-to-noise ratio is larger for the deleted targets than for the duplicated targets. The LIMMA approach, however, leads to a significant reduction in the noise, especially for the non-aberrant targets, and this results in a larger signal-to-noise ratio. Therefore, these statistics are favorable for the LIMMA approach.

6.4.3 True positive and false positive rate

In this section, we will assess the classification capability to call a target duplicated or deleted.

First, we compute for a number of significance levels α, the percentage of the duplicated and deleted targets that are correctly classified as duplicated and deleted, respectively, in our test data set, according to both methods. These *true positive (TP) rates* are shown for the mixed models and LIMMA method in Figure 6.2 in function of the significance level α. For the LIMMA method, the TP rate reaches a maximum of 0.954 for a significance level $\alpha = 0.009$. Afterwards, this TP rate drops a little, as some targets become *strange* targets for larger significance levels α.
For the classification with the mixed models, the TP rates grow slowly as the significance level increases. Within this range of significance levels α, it never reaches the maximum TP rate value obtained with the LIMMA approach. Perhaps it comes closer to the result obtained with LIMMA if we allow for even larger significance levels α, but this will increase the false positive rate and the number of *strange* targets.

Outside the duplicated and deleted regions, other targets were also classified as duplicated or deleted. These positives can be false positives, due to technical artifacts, or they can indicate true biological variations. At this point, we will not make the distinction between both kinds of aberrant targets, as it does not affect the method comparison, and we will refer to this set of positives as non-confirmed positives. At a later stage, this set of non-confirmed positives will be examined in depth, for one method and one significance level α.
The number of these non-confirmed positives is shown in Figure 6.3. The figure shows that there is no clear difference in the non-confirmed positive rates between both methods. For low significance levels α the linear model has a slightly smaller number of non-confirmed positives.

The combined results on the TP and FP rate lead to the conclusion that LIMMA is the preferable method.

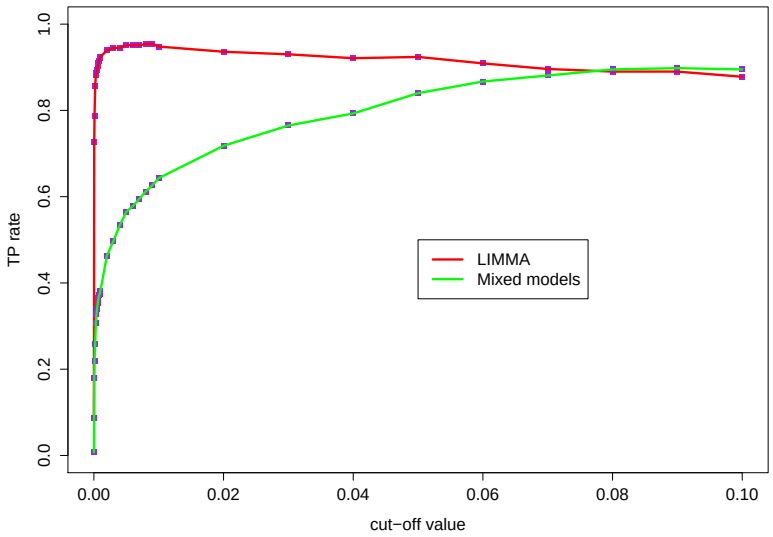


Figure 6.2: **True positive rate.** For the mixed models and the LIMMA approach, the true positive (TP) rates are plotted for a number of significance levels α with a green and red line, respectively.

6.5 Optimization of the LIMMA approach

In the previous section, we focussed on how well the different methods fit the measurements by assessing its capability to divide the non-aberrant targets from the deviating targets and by comparing the FP and TP rates. This indicated that the LIMMA approach was best suited to distinguish these groups of targets, although also the LIMMA approach has a fairly high FP rate. However, we did not yet benefit from all available information.

6.5.1 Completely and partially deleted targets

To further optimize the target classification, we will use the fact that if, for example, a target of Patient 1 is deleted or duplicated, its contrast Patient 2

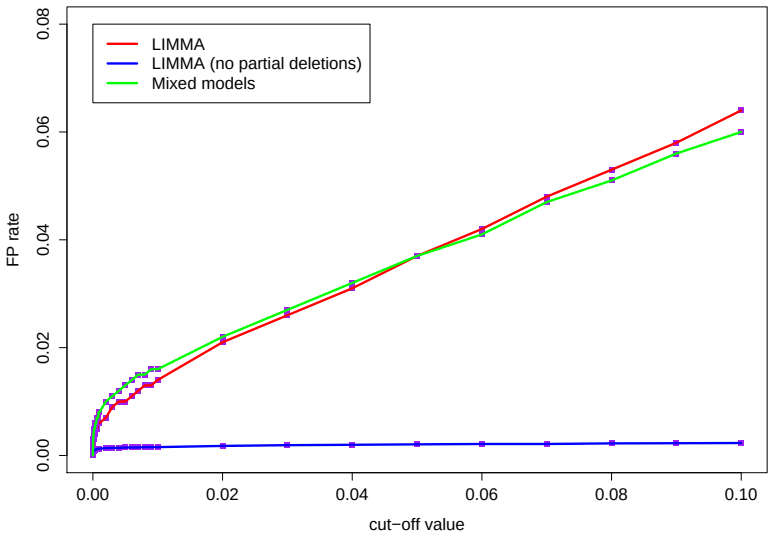


Figure 6.3: **False positive rate.** For the mixed models and the LIMMA approach, the false positive (FP) rates are plotted for the range of significance levels α with a green and red line, respectively. The FP rate for the complete deletions or duplications, obtained via LIMMA is indicated in blue.

vs. Patient 1 should theoretically be equal to

$$\log_2\left(\frac{2}{1}\right) = -1 \text{ or } \log_2\left(\frac{2}{3}\right) = 0.58,$$

respectively. However, nonlinear saturation effects in the signals cause a deviation from these values. Instead of taking the theoretically expected values (i.e., ± 1 and ± 0.58), we will estimate the expected values based on the LIMMA estimates of the contrasts for the group of confirmed deletions and duplications, after exclusion of the deletions and duplications on the X and Y chromosome. This results in an average for the absolute log-ratios of 0.86 for the deleted targets and 0.54 for the duplicated targets.
Therefore, if we detect with the LIMMA method a target that is likely to be

duplicated or deleted, we extract its contrasts C_{c1} and C_{c2}. If their absolute value is not larger than 0.54 or 0.86, respectively, we use an adapted version of the moderated t-test, as implemented in LIMMA. We use the same standard deviation of the contrasts, as computed within the previous LIMMA procedure, and test one-sidedly the hypothesis $H_0 : C = 0.54$ versus $H_a : |C| < 0.54$ for the duplicated targets and $H_0 : |C| = 0.86$ versus $H_a : |C| < 0.86$ for the deleted targets. If we cannot reject the hypothesis at a significance level α for both tests, we call the target *completely* deleted or duplicated, else we call the targets *partially* deviating. As a significance level, we choose $\alpha = 0.01$.
For this significance level, the non-confirmed positives restricted to the targets that are completely deleted is plotted as a blue line in Figure 6.3. This non-confirmed positives rate is smaller than 0.001.

6.5.2 The non-confirmed positives

The non-confirmed positives rate obtained in the previous sections is not a direct indication of the false positives rate, as they can *comprise* not only false positives, but also both true positives or polymorphic targets. A polymorphism is a naturally occurring variation in the DNA sequence, that occurs more frequently than can be accounted for by mutation alone. Often, a variant that occurs in more than 1 percent of the population is called a *polymorphism*.
To detect such polymorphisms, we compose a list of 46 normal patients and assess the targets that are deleted or duplicated twice or more for these patients. This resulted in a list of potential polymorphic clones, as shown in Table 6.4. If we ignore these clones in the computation of the FP rate, the FP rate further decreases.

6.6 Implementation in a web application

The LIMMA method is implemented as a web application and is publicly available at `www.esat.kuleuven.be/loop`.

Clone ID	# detected	Chromosome
SC1_1Mb_PACE9	9	1
NONSC1B4	9	15
NONSC24B9	7	5
SC9BACMbset_1F12	7	17
SC10BAC-1Mbset-1E12	6	10
NONSC16A2	5	5
NONSC32E3	5	16
nonsc43G1	5	17
NONSC24A3	5	-
SC6PAC1Mbset1F9	4	6
SC10BAC-1Mbset-1A3	4	10
NONSC31C3	4	17
SC13_1Mb_BAC1A5	4	Y
SC1_1Mb_BAC1G5	3	1
NONSC10G9	3	2
Cancer_1G8	3	5
NONSC10C7	3	7
Cancer_1A10	3	11
SC13_1Mb_BAC1E4	3	13
nonsc34A4	3	16
nonsc40E7	3	16
NONSC8F4	3	17
NONSC33A5	3	19
NONSC3C7	2	2
NONSC8D3	2	2
telomereG1	2	2
nonsc40D1	2	4
NONSC10B6	2	5
NONSC2C8	2	8
nonsc40A4	2	8
nonsc41E2	2	8
NONSC11G2	2	8
telomereB8	2	14
NONSC32B9	2	15
NONSC10C3	2	16
SC22_0.75_BACB8	2	22

Table 6.4: **Polymorphic clones.** The list of potential polymorphic clones. Next to their identifier, an indication of the number of times it was detected as deviating, along with its chromosome (if determined).

6.6.1 Data processing

On first use, the user is asked to create a username and password. On subsequent login the application offers three components: an upload wizard for GPR files, a slide view which provides an overview of all uploaded hybridizations, and a loop design view offering reports of all loop design

experiments.
The *upload wizard* allows to upload the hybridization data in GenePix GPR file format. You have to specify the unique identifiers for the Cy5 and Cy3 samples: these are used to verify loop designs as valid, and as sample references throughout the tool.
The *slide view* provides an overview of the basic information on uploaded hybridizations. Each hybridization record can be folded open to view technical information, timing, and lab information. In this view, three hybridizations that make up a single loop design can be checked, and these selected hybridizations are combined into a loop design.
After submitting three hybridizations as a new loop design, a new entry is added in the *loop design view*. A status comment indicates what phase of the analysis is active. When done, four reports can be viewed: an overview, and three individual patient reports.
The *overview report* will display the experimental design and quality assessment information, as, for example, the Cy5 and Cy3 slide background images (as in Section 3.3), and the MA-plots (see Section 3.5(b)) before and after spatial Loess normalization (see Section 6.3.1). Significantly aberrant targets are shown in an overview, highlighting targets previously marked as polymorphic (Table 6.4). If needed, a list of significantly aberrant targets and a list of all targets can be downloaded from this report as a tab delimited text file for further processing. A graphical overview (as for example in Figure 6.4) shows normal and aberrant targets ordered by chromosomal position for all hybridizations.
The *individual patient reports* show quality statistics and a graphical display of the significantly aberrant targets, for the specific patient. On this overview, the user can zoom in to individual chromosomes (see, for example, Figure 6.5). A table provides an overview of aberrant targets, and can again be downloaded as a tab delimited file.

6.6.2 Architecture

The web application consists of a set of Java Server Pages that are run in Apache Tomcat java application servers on two different machines. A single MySQL instance with daily back up is used as a back end data store and holds user and hybridization loop design data. GPR files are stored on

raid disks with daily backup, as are images and statistics results. Statistics and visualization scripts were written in R and make use of Bioconductor modules. Instances of RServe, a daemon interface to R which accepts and handles remote calls from Java, are installed on two Unix machines for increased availability.

6.7 Conclusion

The analysis of the loop designs on array CGH is a nice example of the improvement a well-chosen experimental design can bring. It improved the analysis in two ways: the resources better spent, and this loop design allows also to classify correctly the targets as duplicated or deleted.

The analysis of the loop design was done in two ways with the mixed model approach, as proposed by Wolfinger et al. (2001), and LIMMA (Smyth (2004)). Both the signal-to-noise ratio and the false positive and false negative rates gave indications that the LIMMA approach clearly outperformed the mixed model approach. We do not draw any conclusion with regard to the suitability of one class of method versus the other in general. In our setting, we hypothesize that the mixed model was less robust to deviation from the underlying normality assumptions or that the more compact (fewer parameters) estimation procedure of the linear model increased its robustness.
The results indicate that the experimental loop design, together with a statistical analysis by a linear model, provides an efficient procedure for the detection of chromosomal aberrations in congenital anomalies by array CGH. It is significantly superior to the classical setup by doubling the use of resources and unambiguously assigning variation to the correct patient.

This method was implemented in web-based application. The tool is used by two laboratories (i.e., Center for Human Genetics, Leuven, Belgium and Service de Génétique, Reims, France).

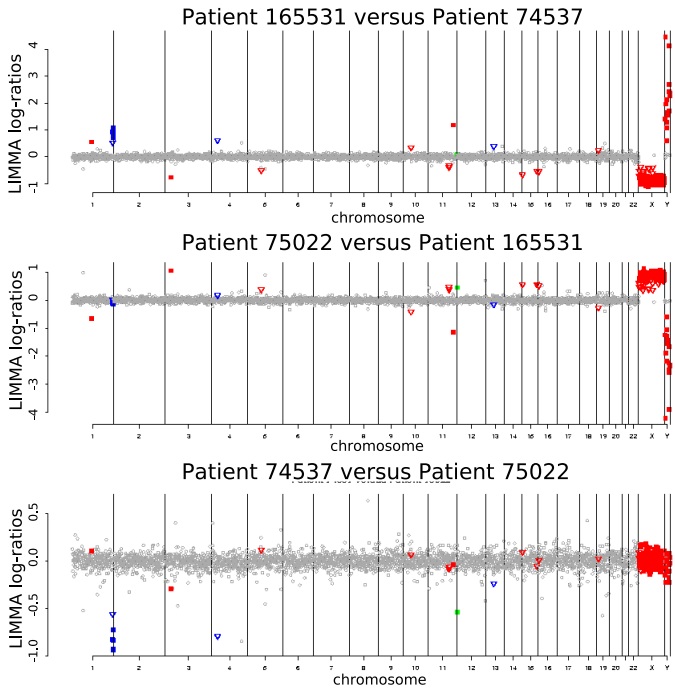


Figure 6.4: **Graphical overview of deletions and duplications.** The figure, as shown in the overview report, displays all duplications and deletions for each hybridization, ordered according to the location on the chromosome. The partial and the complete deletions or duplications are indicated as triangles and squares, respectively. Grey symbols show the non-aberrant targets, the blue symbols are aberrant for patient 74537, the red symbols are aberrant for patient 165531, and the green symbols for patient 75022. In this example, you can detect clearly that patient 74537 has a deletion on chromosome 1 and patient 165531 has a deletion of chromosome X and a duplication of chromosome Y. This indicates a clear sex mismatch.

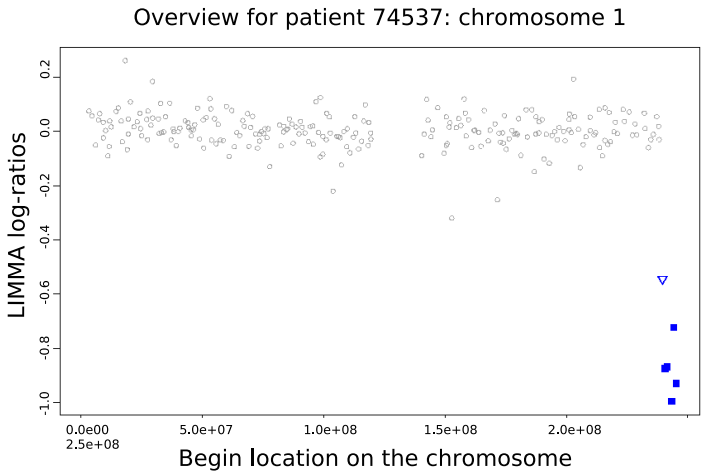


Figure 6.5: **Graphical overview of deletions and duplications per patient and per chromosome.** Per patient, figures are made by taking the average of the log-ratios. For example, for patient 1 the average $\frac{1}{2}\left(\log\left(\frac{\text{patient}_1}{\text{patient}_3}\right) - \log\left(\frac{\text{patient}_2}{\text{patient}_1}\right)\right)$ is shown. This figure shows the average log-ratios for patient 74537 on chromosome 1. Then we see clearly that the deletion is at the end of the chromosome.

Chapter 7

Conclusions and future directions

Finally, a discussion of the work, presented in this thesis, and some indications for related, future research...

7.1 The CAGE project

The majority of this work was done within the framework of the Compendium of *Arabidopsis* Gene Expression or CAGE project. The project aimed at producing 2,000 biological samples of *Arabidopsis thaliana*, to build an atlas of gene expression throughout its life cycle and under a variety of stress conditions.

As a platform the *Complete Transcriptome MicroArray* or CATMA array was chosen. People in the CATMA consortium had high expectations concerning the CATMA array, as the design of the probes was so carefully done, but its utility had not yet been proven. This was the motivation to set up a dedicated experiment that compares the CATMA array with

two alternative, commercial, well established platforms, namely the long oligonucleotide platform of Agilent and the short oligonucleotide platform of Affymetrix. Different aspects of the platforms were compared and the results for the CATMA array are promising. We can state that the CATMA platform performs at least as well as the two other platforms. We can now only hope that the CATMA array will continue to be used and that the data produced within the CAGE project can stimulate this. This work has been done under the guidance of Dr. M. Kuiper and Dr. P. Hilson (VIB-PSB, Gent). The results were published in Allemeersch et al. (2005).

Our main contribution to the CAGE project was the preprocessing of the CAGE data. Within the CAGE project, microarray data exchange and storage of the experiments was done in MAGE-ML format. In this way, CAGE is a well annotated, MIAME compliant compendium.
At the start of the project, there were no tools available to import MAGE-ML files into a microarray data analysis environment. Therefore a software package `RMAGEML` for the statistical environment R has been written, that enables to import MAGE-ML files and to export the files again, updated with preprocessed values. This work has been done in collaboration with my former colleague Dr. Steffen Durinck and Dr. V. Carey (Channing lab, Harvard). The package is now part of Bioconductor and the tool has been published in Durinck et al. (2004).
For the preprocessing and quality assessment of the 4,000 hybridizations, produced within the CAGE project, we developed a preprocessing pipeline, that preprocesses the data in an automatic way, starting from the raw data in MAGE-ML format. To our knowledge it is the first pipeline that incorporates data stored in MAGE-ML format and can export MAGE-ML files, updated with the preprocessed values.
For each hybridization, the preprocessed data, and figures and statistics for quality assessment were made available via a web application. All hybridizations were checked manually and the partner that generated the data was warned in case of serious quality problems or erroneous data.
Although the pipeline was built to work in an automatic way, many manual interventions were required, due to incorrect submission of the data by the partners or by errors in the MAGE-ML encoding (e.g., constantly changing quantitation types, missing labels for the samples etc.). Some of these

problems could have been avoided by simplifying the data submission and, for example, by not allowing for free text in the submission tool. In the end, to fasten up the data submission, EBI allowed to upload the data of large experiments in excel sheets, which simplified the submission and led to a significant decrease in errors.
The data production was slow, due to a number of problems, as, for example, problems with the growth chambers and the printing of the arrays. Now, at the end of the project, a sizable data set has been produced with more than 2,000 hybridizations and the actual analysis of the compendium data set can start. The analysis of some smaller experiments is in progress, in close collaboration with the specific partners who produced the experiment. An example, in collaboration with VIB-PSB, was shown.

Future directions

We showed a very short example of a comparison between two partners, but the actual analysis of the compendium data still has to start. This work should be done in close collaboration with biologists.
A first, important step will be to assess the quality and the comparability of the data produced by the different partners. This can be done by studies, similar to the comparison of the two partners as shown in this work, but also (and perhaps to a larger extent) by including biological knowledge as, for example, by checking the behavior of selected, well-known genes.
Another important step, also requiring biological knowledge, will be to group the experiments, such that the data can be divided in large blocks that describe comparable conditions. Then the actual analysis to assess agreement or disagreement between the partners can start.
Most likely, data will have to be analyzed for each partner separately and significance results then have to be combined afterwards, instead of combining the expression measurements of the different partners.
As an alternative, we should also incorporate the CAGE data as an important source of information to increase the power of gene prioritization applications, as for example Endeavour (Aerts et al. (2006)).
In any case, the analysis of the expression data produced in the CAGE projects can start now and will continue still for quite some time.

The CAGE project was also an ideal opportunity to bring the MAGE-ML format into practice. As this was one of the first cases in which MAGE-ML was used for data exchange, we suffered a lot from child diseases. MAGE-ML format is complex and, therefore, not many data analysis applications support MAGE-ML currently. We advocate to simplify MAGE-ML, by defining how MAIME should be coded into MAGE-ML and by reducing the number of free text inputs. The MGED community is working on the development of a second version of MAGE that eliminates ambiguities (Ball and Brazma (2006)). Once a simpler version of MAGE has been developed, our RMAGEML package should follow adaptations to the standard and keep up to date.

7.2 ArrayCGH

We described an analysis tool to analyze arrayCGH loop designs, in which three patients are placed in a loop design. This has advantages over the classical two-by-two comparison in which a patient is compared to a normal, reference patient. Not only are the resources better spent, this loop design allows also to classify correctly the clones as duplicated or deleted. In a two-by-two comparison, it is not clear whether an aberrant clone is duplicated for the patient of interest or deleted for the normal, reference patient and vice versa. By comparing a patient with two other patients, the vast majority of the clones can be classified correctly.
To analyze the loop design, two methods have been tested, namely the mixed model approach, as proposed by Wolfinger et al. (2001), and LIMMA. The signal-to-noise ratios and the false positive and false negative rates led to the conclusion that the LIMMA approach was preferable to the mixed model approach. This method was implemented in web-based application. The tool is now used by two laboratories (i.e., Center for Human Genetics, Leuven, Belgium and Service de Génétique, Reims, France). The publication of this tool is in progress.

Future directions

This analysis tool can be optimized in many ways. One important task,

that has to be done in any case, is to store all targets on the array in a database, such that the database can be updated regularly and, hence, that the tool will always refer to the correct annotation of the targets. By storing the information of the targets and the aberrations detected for specific patients in a database, it becomes also straightforward to link this information to phenotypical information and text mining tools.
A second modification could be to integrate clone-specific standard deviations. We have already observed that the different clones have a different standard deviation and by extracting these standard deviations over a large set of loop design experiments, clone-specific standard deviations can be estimated more accurately and can be implemented in the t-test. This can help to classify the targets on the X chromosome more correctly.
A third improvement would be to include the information of the neighboring clones. At the moment, each clone on itself is classified as deleted, duplicated or non-aberrant. For the aberrant clones, we also test whether this clone is likely to be completely or partially deleted or duplicated. But, both tests ignore the classification of the adjacent clones. Taking this information into account will lead to the identification of a region that is duplicated or deleted, instead of a list of individual clones, and make the tool more attractive.

Index

References

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006), "Gene prioritization through genomic data fusion," *Nature Biotechnology*, 24, 537–544.

Affymetrix (2001), "Microarray Suite User Guide, Version 5." `http://www.affymetrix.com/products/software/specific/mas.affx`.

— (2002), "Statistical Algorithms Description Document," `http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf`.

Agilent (2003), "Feature Extraction Software User Manual v. 7.1. In, Ed 6th," .

Allemeersch, J., Durinck, S., Vanderhaeghen, R., Alard, P., Maes, R., Seeuws, K., Bogaert, T., Coddens, K., Deschouwer, K., Van Hummelen, P., Vuylsteke, M., Moreau, Y., Kwekkeboom, J., Wijfjes, A. H. M., May, S., Beynon, J., Hilson, P., and Kuiper, M. (2005), "Benchmarking the CATMA microarray: a novel tool for Arabidopsis transcriptome analysis." *Plant Physiology*, 137, 588–601.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006), "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, 7, 55–65.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944), "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types." *Journal of Experimental Medicine*, 79, 137–158.

Bachem, C. W., van der Hoeven, R. S., de Bruijn, S. M., Vreugdenhil, D., Zabeau, M., and Visser, R. G. (1996), "Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development," *Plant J*, 9, 745–753.

Ball, C. A. and Brazma, A. (2006), "MGED standards: work in progress," *OMICS*, 10, 138–144.

Barczak, A., Rodriguez, M., Hanspers, K., Koth, L., Tai, Y., Bolstad, B., Speed, T., and Erle, D. (2003), "Spotted long oligonucleotide arrays for human gene expression analysis," *Genome Res*, 13, 1775–1785.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005), "NCBI GEO: mining millions of expression profiles–database and tools," *Nucleic Acids Res*, 33, 562–566.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc Ser B*, 57, 289–300.

Boyes, D., Zayed, A., Ascenzi, R., McCaskill, A., Hoffman, N., Davis, K., and Görlach, J. (2001), "Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants," *Plant Cell*, 13, 1499–1510.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001), "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nat Genet*, 29, 365–371.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S.-A. (2003), "ArrayExpress–a public repository for microarray gene expression data at the EBI," *Nucleic Acids Res*, 31, 68–71.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000), "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays," *Nat Biotechnol*, 18, 630–634.

Chudin, E., Walker, R., Kosaka, A., Wu, S., Rabert, D., Chang, T., and Kreder, D. (2002), "Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays," *Genome Biol*, 3.

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

Craigon, D., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004), "NASCArrays: a repository for microarray data generated by NASC's transcriptomics service," *Nucleic Acids Res*, 32, 575–577.

Crowe, M., Serizet, C., Thareau, V., Aubourg, S., Rouzé, P., Hilson, P., Beynon, J., Weisbeek, P., van Hummelen, P., Reymond, P., Paz-Ares, J., Nietfeld, W., and Trick, M. (2003), "CATMA: a complete Arabidopsis GST database," *Nucleic Acids Res*, 31, 156–158.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., and Moreau, Y. (2002), "Adaptive quality-based clustering of gene expression profiles," *Bioinformatics*, 18, 735–746.

DeFrancesco, L. (2002), "Journal trio embraces MIAME," *The Scientist*, 10.

Durinck, S., Allemeersch, J., Carey, V. J., Moreau, Y., and De Moor, B. (2004), "Importing MAGE-ML format microarray data into BioConductor," *Bioinformatics*, 20, 3641–3642.

Ferl, G., Timmerman, J., and Witte, O. (2003), "Extending the utility of gene profiling data by bridging microarray platforms," *Proc Natl Acad Sci U S A*, 100, 10585–10587.

Ford, C. and Hamerton, J. (1956), "The chromosomes of man," *Nature*, 178, 1020–1023.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004), "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol*, 5, R80.

Gress, T. M., Hoheisel, J. D., Lennon, G. G., Zehetner, G., and Lehrach, H. (1992), "Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues," *Mamm Genome*, 3, 609–619.

Hilson, P., Allemeersch, J., Altmann, T., Aubourg, S., Avon, A., Beynon, J., Bhalerao, R., Bitton, F., Caboche, M., Cannoot, B., Chardakov, V., Cognet-Holliger, C., Colot, V., Crowe, M., Darimont, C., Durinck, S., Eickhoff, H., de Longevialle, A., Farmer, E., Grant, M., Kuiper, M., Lehrach, H., Léon, C., Leyva, A., Lundeberg, J., Lurin, C., Moreau, Y., Nietfeld, W., Paz-Ares, J., Reymond, P., Rouzé, P., Sandberg, G., Segura, M., Serizet, C., Tabrett, A., Taconnat, L., Thareau, V., Van Hummelen, P., Vercruysse, S., Vuylsteke, M., Weingartner, M., Weisbeek, P., Wirta, V., Wittink, F., Zabeau, M., and Small, I. (2004), "Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications," *Genome Res*, 14, 2176–2189.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R.,

Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., and Linsley, P. S. (2001), "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer," *Nat Biotechnol*, 19, 342–347.

Ihaka, R. and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, 4, 249–264.

Ishkanian, A. S., Malloff, C. A., Watson, S. K., DeLeeuw, R. J., Chi, B., Coe, B. P., Snijders, A., Albertson, D. G., Pinkel, D., Marra, M. A., Ling, V., MacAulay, C., and Lam, W. L. (2004), "A tiling resolution DNA microarray with complete coverage of the human genome," *Nat Genet*, 36, 299–303.

Jordan, B. (2002), "Historical background and anticipated developments," *Ann N Y Acad Sci*, 975, 24–32, historical Article.

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., and Pinkel, D. (1992), "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors," *Science*, 258, 818–821.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000), "Analysis of variance for gene expression microarray data," *J Comput Biol*, 7, 819–837.

Knight, J. (2001), "When the chips are down," *Nature*, 410, 860–861, news.

Kuo, W., Jenssen, T., Butte, A., Ohno-Machado, L., and Kohane, I. (2002), "Analysis of matched mRNA measurements from two different microarray technologies," *Bioinformatics*, 18, 405–412.

Laird, N. and Ware, J. (1982), "Random-effects models for longitudinal data," *Biometrics*, 38, 963–974.

Le Caignec, C., Spits, C., Sermon, K., De Rycke, M., Thienpont, B., Debrock, S., Staessen, C., Moreau, Y., Fryns, J.-P., Van Steirteghem, A., Liebaers, I., and Vermeesch, J. R. (2006), "Single-cell chromosomal imbalances detection by array CGH," *Nucleic Acids Res*, 34, e68.

Lee, J., Bussey, K., Gwadry, F., Reinhold, W., Riddick, G., Pelletier, S., Nishizuka, S., Szakacs, G., Annereau, J., Shankavaram, U., Lababidi, S., Smith, L., Gottesman, M., and Weinstein, J. (2003), "Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells," *Genome Biol*, 4.

Lejeune, J., Gauthier, M., and Turpin, R. (1959), "Human chromosomes in tissue cultures." *C R Hebd Seances Acad Sci*, 248, 602–603.

Li, C. and Wong, W. (2001), "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection." *Proceedings of the National Academy of Science USA*, 98, 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, 1675–1680.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., and Dabrowski, M. (2003), "Comparison and meta-analysis of microarray data: from the bench to the computer desk," *Trends Genet*, 19, 570–577.

Nimgaonkar, A., Sanoudou, D., Butte, A., Haslett, J., Kunkel, L., Beggs, A., and Kohane, I. (2003), "Reproducibility of gene expression across generations of Affymetrix microarrays," *BMC Bioinformatics*, 4, 27–27.

Noctor, G., Veljovic-Jovanovic, S., Driscoll, S., Novitskaya, L., and Foyer, C. H. (2002), "Drought and oxidative load in the leaves of C3 plants: a predominant role for photorespiration?" *Ann Bot (Lond)*, 89 Spec No, 841–850.

Okubo, K. and Matsubara, K. (1997), "Complementary DNA sequence (EST) collections and the expression information of the human genome," *FEBS Lett*, 403, 225–229.

Oostlander, A., Meijer, G., and Ylstra, B. (2004), "Microarray-based comparative genomic hybridization and its applications in human genetics," *Clin Genet*, 66, 488–495.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S., and Brazma, A. (2005), "ArrayExpress–a public repository for microarray gene expression data at the EBI," *Nucleic Acids Res*, 33, 553–555.

Pinheiro, J. and Bates, D. (2000), *Mixed-effects models in S and S-PLUS*, Springer-Verlag.

Puskás, L., Zvara, A., Hackler, L., and Van Hummelen, P. (2002), "RNA amplification results in reproducible microarray data with slight ratio bias," *Biotechniques*, 32, 1330–1334.

Redman, J., Haas, B., Tanimoto, G., and Town, C. (2004), "Development and evaluation of an Arabidopsis whole genome Affymetrix probe array," *Plant J*, 38, 545–561.

Rhodes, D., Barrette, T., Rubin, M., Ghosh, D., and Chinnaiyan, A. (2002), "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Res*, 62, 4427–4433.

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002), "Using the transcriptome to annotate the genome," *Nat Biotechnol*, 20, 508–512.

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, 467–470.

Schiex, T., Moisan, A., and Rouzé, P. (2001), "EUGÈNE: an eukaryotic gene finder that combines several sources of evidence," *Lect. Notes Comput. Sci.*, 2066, 111–125.

Shaffer, L. G. and Bejjani, B. A. (2004), "A cytogeneticist's perspective on genomic microarrays," *Human Reproduction Update*, 10, 221–226.

Smeets, D. F. C. M. (2004), "Historical prospective of human cytogenetics: from microscope to microarray," *Clin Biochem*, 37, 439–446.

Smyth, G. (2004), "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology*, 3.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J. J., and Brazma, A. (2002), "Design and implementation of microarray gene expression markup language (MAGE-ML)," *Genome Biol*, 3, RESEARCH0046.

Sterrenburg, E., Turk, R., Boer, J. M., van Ommen, G. B., and den Dunnen, J. T. (2002), "A common reference for cDNA microarray hybridizations," *Nucleic Acids Res*, 30, e116.

Storey, J. and Tibshirani, R. (2003), "Statistical significance for genomewide studies," *Proc Natl Acad Sci U S A*, 100, 9440–9445.

Tan, P., Downey, T., Spitznagel, E., Xu, P., Fu, D., Dimitrov, D., Lempicki, R., Raaka, B., and Cam, M. (2003), "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Res*, 31, 5676–5684.

Thareau, V., Déhais, P., Serizet, C., Hilson, P., Rouzé, P., and Aubourg, S. (2003), "Automatic design of gene-specific sequence tags for genome-wide functional studies," *Bioinformatics*, 19, 2191–2198.

The International Human Genome Sequencing Consortium, T. (2001), "Initial sequencing and analysis of the human genome," *Nature*, 409, 860–921.

Tijo, J. and Levan, A. (1956), "The chromosome number of man." *Hereditas*, 42, 1–6.

Trask, B. (2002), "Human cytogenetics: 46 chromosomes, 46 years and counting," *Nat Rev Genet*, 3, 769–778, historical Article.

Tukey, J. W. (1949), "One degree of freedom test for non-additivity." *Biometrics*, 5, 232–242.

Vandenabeele, S., Van Der Kelen, K., Dat, J., Gadjev, I., Boonefaes, T., Morsa, S., Rottiers, P., Slooten, L., Van Montagu, M., Zabeau, M., Inze, D., and Van Breusegem, F. (2003), "A comprehensive analysis of hydrogen peroxide-induced gene expression in tobacco," *Proc Natl Acad Sci U S A*, 100, 16113–16118.

Vandenabeele, S., Vanderauwera, S., Vuylsteke, M., Rombauts, S., Langebartels, C., Seidlitz, H. K., Zabeau, M., Van Montagu, M., Inze, D., and Van Breusegem, F. (2004), "Catalase deficiency drastically affects gene expression induced by high light in Arabidopsis thaliana," *Plant J*, 39, 45–58.

Vanderauwera, S., Zimmermann, P., Rombauts, S., Vandenabeele, S., Langebartels, C., Gruissem, W., Inze, D., and Van Breusegem, F. (2005), "Genome-wide analysis of hydrogen peroxide-regulated gene expression in Arabidopsis reveals a high light-induced transcriptional cluster involved in anthocyanin biosynthesis," *Plant Physiol*, 139, 806–821.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995), "Serial analysis of gene expression," *Science*, 270, 484–487.

Vermeesch, J. R., Melotte, C., Froyen, G., Van Vooren, S., Dutta, B., Maas, N., Vermeulen, S., Menten, B., Speleman, F., De Moor, B., Van Hummelen, P., Marynen, P., Fryns, J.-P., and Devriendt, K. (2005), "Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis," *J Histochem Cytochem*, 53, 413–422.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M. (1995), "AFLP: a new technique for DNA fingerprinting," *Nucleic Acids Res*, 23, 4407–4414.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001), "Assessing gene

significance from cDNA microarray expression data via mixed models," *J Comput Biol*, 8, 625–637.

Wu, Z. and Irizarry, R. (2005), "Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays." *J Comput Biology*, 12, 882–93.

Yang, Y., Buckley, M., Dudoit, S., and Speed, T. (2000), "Comparison of methods for image analysis on cDNA microarray data," .

Yuen, T., Wurmbach, E., Pfeffer, R., Ebersole, B., and Sealfon, S. (2002), "Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays," *Nucleic Acids Res*, 30.

Curriculum vitae

Joke Allemeersch was born in Diest on April 11, 1979. She obtained a Master's degree in Mathematics at the Katholieke Universiteit Leuven in 2001. Her master thesis on integration techniques "Vergelijkende studie van integratietechnieken" was performed under supervision of Prof. Dr. J. Quaegebeur. After graduation, she obtained in 2002 a Master of Science in Statistics with summa cum laude distinction, also at the K.U.Leuven. For this degree, she made a study on "A wavelet-based estimator of the Hurst parameter applied to Ethernet traffic Data", under the guidance of Prof. Dr. W. Van Assche. In October 2002 she joined ESAT-SCD as a Ph.D. student in the field of Bioinformatics under the supervision of Prof. Dr. Y. Moreau and Prof. Dr. B. De Moor.