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

NORMALIZING MICROARRAY DATA:
ESTIMATING ABSOLUTE EXPRESSION LEVELS

Jury:
Prof. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. K. Marchal, co-promotor
Prof. dr. ir. J. Suykens
Prof. dr. ir. J. Vanderleyden
Prof. dr. J. Winderickx
Prof. dr. T. Ayoubi (UM, Nederland)
Prof. dr. ir. Y. Moreau

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Kristof ENGELEN

UDC 681.3*J3 December 2005

D/2005/7515/97

ISBN 90-5682-669-7

Voorwoord

Toen ik, nu meer dan vier jaar geleden, aan mijn doctoraat begon, was ik één van de eerste bioingenieurs in de BIOI-groep van ESAT-SCD. Ik kwam terecht in de vreemde wereld van 'die andere' ingenieurs. Een wereld die ik misschien nog steeds niet helemaal begrijp, maar ik heb de voorbije jaren op ESAT wel ontzettend veel geleerd, en wil dan ook in de eerste plaats mijn promotor Prof. Bart De Moor bedanken. Niet alleen voor de kansen die hij me gegeven heeft, maar ook voor de steun en interesse die hij altijd getoond heeft voor mijn, soms wel eigenzinnig onderzoek. Het schrijfproces van mijn thesis is uit noodzaak van erg korte duur geweest, en daardoor werden de meeste geschreven regels van de administratieve afhandeling met de voeten getreden. Ik wil de leden van mijn jury en begeleidingscommissie oprecht bedanken voor de tijd die ze vrijgemaakt hebben (en de flexibiliteit die ze daarbij getoond hebben) voor het nalezen van mijn tekst, ondanks ongetwijfeld overvolle agenda's.

Niemand ben ik meer voor mijn doctoraat verschuldigd dan Kathleen (ik zal de 'Prof. Marchal' achterwege laten, ik weet dat je daar maar niet aan kan wennen). Het enthousiasme waarmee je mijn onderzoek dagdagelijks hebt begeleid is ongeëvenaard. In al die jaren dat ik je ken, ben je meer een grote zus dan een baas geweest: ik kan me geen moment herinneren dat je niet klaarstond voor mij (of één van je andere studenten), en dat is meer dan bewonderenswaardig. Ik heb je meer dan eens het bloed van onder de nagels gehaald, en dat zal zeker nog gebeuren. Ik kan niet zeggen dat dat me spijt, maar ik beloof plechtig dat ik je nooit meer zal doen wenen.

Ik wil ook het IWT bedanken, dat mij vier jaar lang financieel gesteund heeft, en zonder hetwelk dit onderzoek nooit was mogelijk geweest.

Door de aard van mijn onderzoek heb ik de gelegenheid gehad om met verschillende biologische en biomedische onderzoeksgroepen samen te werken. Ik ben betrokken geweest bij de meest interessante, en uiteenlopende onderzoeksprojecten. Ik wil dan ook iedereen bedanken die

zo dapper was zijn of haar data, en/of het design van hun experimenten aan mij toe te vertrouwen. Een bijzondere vermelding verdienen de mensen die bereid waren mij te volgen in mijn, misschien wat onorthodoxe opvattingen over microarrays, en daarin kosten noch moeite gespaard hebben: Jos en Sigrid (CMPG), Johan en Bart (Afd. Planten en Micro-organismen), en Bart en Koen (ISLab).

Doctoraatsonderzoek staat niet los van de groep waarin het gevoerd wordt, en de BIOI-groep van SCD is er één die alle hoeken van de Arenbergcampus gezien heeft: van de broeihete ESAT-zolder, naar de schroeiend hete 200F van scheikunde, en weer terug naar ESAT, ditmaal naar een ijskoude toren. Al die verhuizingen deden niets af aan de werklust en sfeer binnen de groep. Ik wil iedereen op BIOI, vroeger en nu, bedanken voor de geweldige tijd die ik er beleefd heb, en een speciaal bedankje voor de mensen waarmee ik nauw samengewerkt heb: Frank (had jij me niet op weg geholpen, had mijn doctoraat nooit op tijd klaar geraakt!), Ruth (Toki Toki Boom Boom?), Pieter (immaand oem teege te ziejevere in ef eige toal), Karen, Nathalie, Tijl, Tim en Thomas. Ook mag ik Bart, Ida, en Ilse niet vergeten voor al hun hulp doorheen de administratieve rompslomp.

Ongeveer op hetzelfde moment als ik, begon ene Bert Coessens aan zijn doctoraat in de BIOI-groep. Bert is een wat timide jongen, wars van discussies, maar zondermeer een hele toffe pee. Hij was mijn huisgenoot gedurende drie jaar en dat was niet altijd even gemakkelijk. Tenminste voor hem niet: samenwonen met iemand die in constante ontkenning van de afwas vertoeft, moet met momenten een hele opgave geweest zijn. En oh ja, Bert, het spijt me nog steeds heel erg van die gaten in je keukentafel. Ondervinding is een goede leerschool: een keuken is geen schrijnwerkerij.

Hoewel sommige professoren hun studenten het tegendeel willen wijsmaken (althans, dat heb ik van horen zeggen), is het leven meer dan doctoreren alleen. Doctoraatszorgen relativeren, vergeten, of verdrinken bij pot en pint, daar zijn vrienden (en broers!) voor. Met mijn collega-muzikanten bij *Kokain* en *The Mob Stories* heb ik de voorbije jaren de meest memorabele momenten meegemaakt. Niets is zo goed om een mens zijn frustraties weg te nemen dan ‘Gaaaaas geeeveeeeeuh!’ Muziek heeft altijd een centrale plaats gehad in mijn leven. De zeldzame keren dat ik de laatste jaren eens vóór zeven uur opgestaan ben, was om samen met mijn voor-zolang-ik-me-kan-herinneren beste vriend te lakken, schuren en polijsten aan een blok padouk om er een bespeelbaar instrument van te maken. Dus Steven, wanneer beginnen we aan de volgende?

Karel, ik herinner me nog de eerste mei in 2000 toen jij er voor gezorgd hebt dat mijn eindverhandeling netjes ingebonden klaar was om in te dienen. En nu weer; zonder jou had dit boekje nooit op tijd klaar geweest. Bedankt!

Terwijl ik de vorige alinea's aan het typen was riep mijn moeder me toe vanuit de keuken: "Weet je nog toen je begon te studeren en wij zeiden dat je één keer mocht proberen en als het dan niet lukte…goh, en nu ga je je doctoraat afleggen!" Dat weet ik nog al te goed, ma, pa en broer. Ik weet ook dat jullie altijd voor mij klaar stonden en het nooit nagelaten hebben van mij op alle vlakken te steunen en mijn weg te laten kiezen. Daarvoor kan ik jullie niet genoeg bedanken!

En lieve Loo, ik ben nog het meest dankbaar voor het feit dat ik jou heb leren kennen, en voor je volharding om mij daarvan te overtuigen toen ik dat zelf nog niet wist…

Kristof Engelen

December 2005

Abstract

The microarray platform is a relatively complex technology that permits the simultaneous assessment of mRNA expression levels of thousands of genes in a single hybridization assay. Normalization of spotted microarray measurements, the first step in a microarray analysis trajectory, aims at removing consistent and systematic sources of variations to allow mutual comparison of measurements acquired from different slides and experimental settings. Data normalization largely influences the results of all subsequent analyses and the biological interpretation of these results, and is therefore a crucial phase in the analysis of microarray data. Over the past years, the field of microarray analysis finally seems to have adapted a few generally applied methodologies for data normalization. Although some approaches inherently work with absolute intensities, in general, normalization of spotted microarrays largely revolves around the calculation of the log-ratios of the measured intensities. Moreover, these techniques generally show little interest in the underlying causes of the observed systematic and random variation in microarray data.

The normalization methods we pursue in this thesis differ in spirit from standard log-ratio approaches. The basic premise is to acknowledge the physical and biological reality of the process and address the normalization problem starting from units of absolute intensities. These measured intensities are to be modelled as a function of systematic sources of variation in a physically and experimentally meaningful way, and should allow for the calculation of an absolute value of expression instead of being limited to the relative nature of intensity ratios. During the initial research stage, the use of ANOVA for microarray normalization, at the time the only available method that allowed for calculation of absolute expression values, was evaluated and compared to ratio based approaches. Based on these results, further research was conducted towards the development and deployment of generic (applicable to any experimental setup) ANOVA models for microarray normalization. ANOVA approaches nevertheless suffer from several

shortcomings. To circumvent these issues we developed a novel normalizing method for spotted microarray data, using external control spikes to fit a calibration model. External control spikes serve to estimate the model parameters. The obtained parameters values are then employed to estimate absolute levels of expression for the remaining genes. We illustrate the workings and principles of this method by applying it to a publicly available benchmark data set.

Korte inhoud

Microroosters zijn een relatief complexe technologie, die toelaten de mRNA-expressieniveaus van duizenden genen tegelijkertijd te meten. Normalisatie van de metingen is de eerste stap in de analyse van microroosterdata. De bedoeling ervan is het verwijderen van consistente en systematische bronnen van variatie, zodat metingen van verschillende microroosters en biologische condities onderling vergeleken kunnen worden. Normalisatie van de data heeft een substantiële invloed op de resultaten van alle daaropvolgende analyses en de biologische interpretatie ervan. Gedurende de voorbije jaren zijn verscheidene methodes voor de normalisatie van microroosterdata ontwikkeld die als standaard kunnen beschouwd worden. Hoewel sommige van deze aanpakken inherent werken met absolute intensiteiten, is het verwerken van microroosterdata grotendeels gebaseerd op het berekenen van log-ratio's van de gemeten intensiteiten. Daarnaast vertonen deze normalisatietechnieken weinig interesse in de onderliggende oorzaken van de geobserveerde systematische en willekeurige variaties van de gemeten intensiteiten.

De normalisatiestrategieën die in deze thesis uitgewerkt zijn, zijn anders in opzet. De achterliggende idee is om rekening te houden met de fysische en biologische realiteit van het proces en om het normalisatieprobleem aan te pakken vertrekkende van absolute intensiteiten. De gemeten intensiteiten worden gemodelleerd op een fysisch en experimenteel betekenisvolle manier, om het zodoende mogelijk te maken om absolute waarden van genexpressie te schatten, in plaats van beperkt te zijn door de relatieve aard van intensiteitsratio's. Initieel onderzoek bestond uit de evaluatie van procedures voor microroosternormalisatie steunend op ANOVA-modellen, en een vergelijkende studie met op ratio's gebaseerde technieken. Verder onderzoek was gericht op de ontwikkeling van generische (toepasbaar op elk experimenteel design) ANOVA-modellen voor normalisatie van microroosterdata. Deze aanpak vertoonde echter verschillende tekortkomingen en daarom werd een geheel nieuwe methode ontwikkeld

gebaseerd op een fysisch gemotiveerd calibratiemodel. Externe controles zijn een centraal onderdeel van deze methode aangezien ze toelaten de parameters van het calibratiemodel te schatten, dewelke op hun beurt kunnen gebruikt worden om absolute expressiewaarden voor de overige genen te berekenen.

Notation

Symbols

y	Measured intensity
$Cy3$, $Cy5$	Used as subscripts; indicate whether a parameter applies to the Cy3 or the Cy5 channel.
I	Logarithm transformed intensities
M	Log-ratios
A	Average of the logarithm transformed intensities over Cy3 and Cy5
M_{corr}	Corrected ratio based on an intensity dependent normalization
I_{ijklm}, I_{ijklmn}	Logarithm transformed intensities (ANOVA)
ε_{ijklm}, ε_{ijklmn}	Model error terms (ANOVA)
C_j	Condition effect parameter
D_l	Dye effect parameter
B_m	Batch effect parameter

A_k , $A_{k(m)}$	Array effect parameter
G_i , $G_{i(n(m))}$	Gene effect parameter
GC_{ij} , $GC_{ij(n(m))}$	Gene×condition interaction effect parameter
GA_{ik}	Gene×array interaction effect parameter
$R(GA)_{m(ik)}$, $R(G)_{m(i)}$	Replicate spot effect parameter
$P_{n(m)}$, $PA_{nk(m)}$	Pin-group effect parameter
x_0	Target concentration in the hybridization solution
x_s	Amount of target hybridized to a spotted probe
s	Remaining spot capacity
s_0	Total spot capacity
μ_s	Average spot capacity
ε_s	Spot capacity error
σ_s	Spot capacity error variance
K_A	Hybridization constant
p_1	Saturation function slope
p_2	Saturation function intercept
ε_m	Multiplicative intensity error
σ_m	Multiplicative intensity error variance
ε_a	Additive intensity error
σ_a	Additive intensity error variance

Acronyms

ANOVA	Analysis of Variance
AQBC	Adaptive Quality Based Clustering
aRNA	Antisense RNA
cDNA	Complementary DNA
CGH	Comparative Genomic Hybridization
ChIP	Chromatin Immunoprecipitation
DNA	Deoxyribonucleic Acid
ERCC	External RNA Control Consortium
EST	Expressed Sequence Tags
GNA	Global Normalization Assumption
INCLUSive	Integrated Clustering and Upstream Sequence Retrieval
LOWESS	Locally Weighted Scatter Plot Smoothing
MIAME	Minimum Information About a Microarray Experiment
mRNA	Messenger RNA
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcriptase PCR
SAM	Significance Analysis of Microarrays

Nederlandse samenvatting

Normalisatie van microroostermetingen: schatten van absolute expressiewaarden

Hoofdstuk 1: Inleiding

Hoge-doorvoer data en microroosters

In traditioneel genetisch en moleculair biologisch onderzoek werden genen, eiwitten en andere moleculen een voor een bestudeerd als geïsoleerde entiteiten. Technologische vernieuwingen hebben, voornamelijk gedurende het voorbije decennium, hier grondig verandering in gebracht. De toepassing van *hoge-doorvoer* (*high-throughput*) technologieën (genomica, transcriptomica, metabolomica) laat immers toe om in een zeer korte tijd de DNA-sequentie van hele genomen in kaart te brengen, gelijktijdig de expressie van duizenden genen of proteïnen in een organisme te analyseren, de aard en concentratie van metabolieten te evalueren en de interacties tussen deze verschillende genetische entiteiten te identificeren. De focus van biologisch onderzoek is verschoven van alleenstaande, of een beperkt aantal genen en proteïnen, naar de analyse van hele populaties.

Het voordeel van een dergelijke holistische aanpak is dat men een beter inzicht kan bekomen in de fundamentele, moleculair biologische processen, aangezien een gen gesitueerd wordt in een globale context, als deel van een complex regulatorisch netwerk. Een cel of organisme wordt beschouwd als een systeem dat interageert met zijn omgeving en waarvan het gedrag wordt bepaald door de dynamische interacties tussen genen, proteïnen en metabolieten op het niveau van het regulatorisch netwerk (i.e. *systeembiologie*).

Hoge-doorvoer experimenten hebben onderzoekers niettemin voor verscheidene uitdagingen gesteld. De analyse van data die gegenereerd wordt op zulk een grote schaal is verre van triviaal. *Bioinformatica* is een jong en snel groeiend interdisciplinair onderzoeksdomein, hetwelk kan

gedefinieerd worden als de wetenschap die zich bezighoudt met het computationele management en de analyse van diverse soorten van moleculair biologische data, of deze nu betrekking hebben op genen en gerelateerde moleculen, cellen, organismen of zelfs hele ecologische systemen.

De opkomst van *microroosters* (*microarrays*) was -en is nog steeds- een drijvende kracht achter de verdere ontwikkeling en wereldwijde inburgering van hoge-doorvoor technologieën. Het doel van de meeste microroosterexperimenten is de identificatie van genen die differentieel tot expressie komen in RNA-stalen die geëxtraheerd zijn uit verschillende celtypes of cellen groeiend in verschillende condities. Veel van de principes van moderne microroosters stammen uit de late jaren '80 en de prille jaren '90 toen gekloneerde cDNA probes, gepositioneerd op membraanfilters, werden gehybridiseerd met complexe mengsels van *target* moleculen om verschillen in genexpressie te quantificeren [37,84,122,129,191]. Een grote doorbraak kwam medio jaren '90, toen Pat Brown, Ron Davis en collega's hun onderzoek publiceerden dat de werking beschreef van een tweekleuren, intern comparatieve techniek waarbij cDNA probes in hoge densiteit machinaal op een vaste drager werden bevestigd [49,175,176]. Deze studies hebben geleid tot de ontwikkeling van DNA-microroosters die toelaten de relatieve expressie van duizenden mRNA-transcripten simultaan te bestuderen.

Microroosters zijn een complexe technologie die kan rekenen op de interesse van specialisten uit uiteenlopende onderzoeksdomeinen (niet alleen moleculair biologen en genetici, maar ook chemici, fysici, ingenieurs, wiskundigen, computerwetenschappers, etc.) en het gebruik ervan heeft geleid tot belangrijke resultaten en inzichten in uiteenlopende sectoren, gaande van fundamenteel biologisch onderzoek, tot biomedische en industriële toepassingen. Het onderzoek dat beschreven wordt in deze doctoraatsthesis is volledig gesitueerd in het gebied van de analyse van microroosterdata. Het handelt over de normalisatie van intensiteiten die bekomen worden van gescande beelden van een microroosterexperiment.

Motivatie van het onderzoekswerk

Normalisatie van de metingen is de eerste stap in de analyse van microroosterdata. De bedoeling ervan is het verwijderen van consistente en systematische bronnen van variatie, zodat metingen van verschillende microroosters en biologische condities onderling vergeleken kunnen worden. Normalisatie van de data heeft een substantiële invloed op de resultaten van alle daaropvolgende analyses en de biologische interpretatie ervan. Het is daarom een cruciale fase in de analyse van microroosterdata.

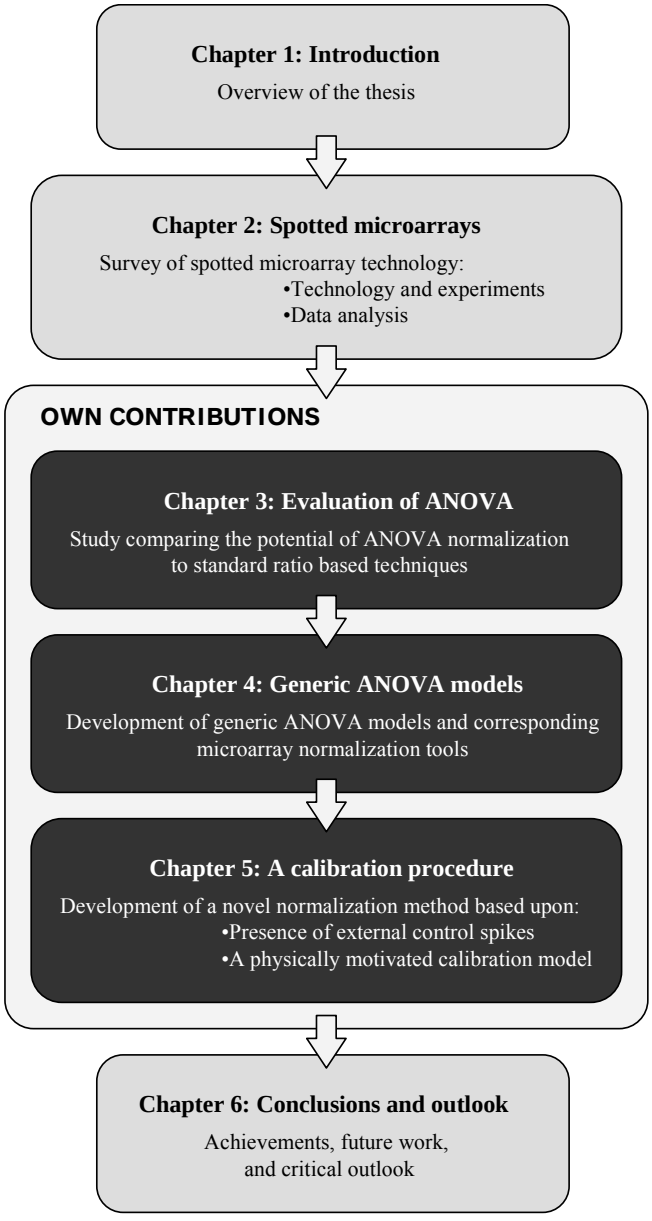
Gedurende de voorbije jaren zijn verscheidene methodes voor de normalisatie van microroosterdata ontwikkeld die als standaard kunnen

beschouwd worden (enkele goede overzichtsartikels zijn vb. Leung and Cavalieri, 2003 [123], Quackenbush, 2002 [156], and Bilban *et al.*, 2002 [22]). Hoewel sommige van deze aanpakken inherent werken met absolute intensiteiten (e.g. ANOVA [113,221]), is het verwerken van microroosterdata grotendeels gebaseerd op het berekenen van *log-ratio's* van de gemeten intensiteiten. Dit is te wijten aan het inherent differentieel karakter van microroosterexperimenten: twee verschillende stalen, gelabeld met verschillende fluorochromen (Cy3 en Cy5), worden gelijktijdig gehybridiseerd op hetzelfde microrooster. Gezien de vergelijkende aard van microroosterexperimenten is het nemen van ratio's van de gemeten intensiteiten een logische benadering voor de analyse van de resultaten. Het gebruik van dergelijke ratio's is echter niet zonder nadelen. Vanuit een theoretisch standpunt zullen ratio's de ruis op de metingen vergroten door de experimentele fout op de intensiteiten te vermenigvuldigen. Daarnaast houden ratio's geen rekening met mogelijk nuttige informatie in verband met het absolute niveau van genexpressie (een bepaalde intensiteitsratio kan bijvoorbeeld wijzen op een significant verschil in expressie in het geval van relatief hoge individuele intensiteiten, terwijl eenzelfde ratio voor lagere intensiteiten geen betekenis heeft omwille van een hogere onbetrouwbaarheid). Het gebruik van ratio's heeft ook verscheidene praktische implicaties. Zo is het moeilijk om voor complexe experimentele designs meerdere biologische condities met elkaar te vergelijken, vooral wanneer deze niet vergeleken werden met dezelfde referenties.

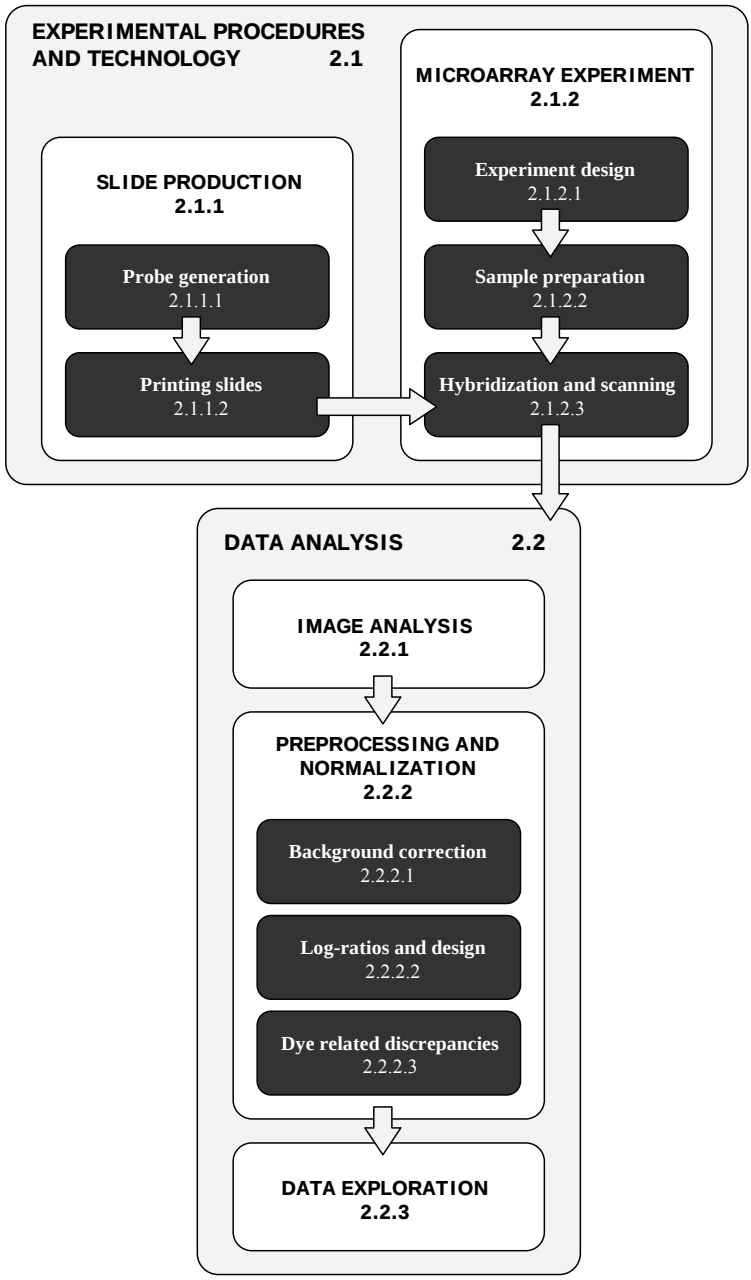
Een ingeburgerde normalisatiestap is de linearisatie van Cy3- versus Cy5-intensiteiten (e.g. LOWESS [226]). Dergelijke methoden nemen aan dat de distributie van genexpressiewaarden weinig globale veranderingen vertoont en gebalanceerd is ten opzichte van de geteste biologische condities (i.e. de *Globale Normalisatie Assumptie*), een assumptie waarvan werd aangetoond dat ze verre van altijd opgaat [206-208]. Microroosterdata worden dus over het algemeen genormaliseerd door de berekende ratio's te transformeren naar een maat van differentiële expressie waaraan men verwacht dat de onderliggende biologische realiteit beantwoordt. Ratio-normalisatietechnieken vertonen weinig interesse in de onderliggende oorzaken van de geobserveerde systematische en willekeurige variaties in intensiteiten.

De normalisatiestrategieën die in deze thesis uitgewerkt zijn, zijn anders in opzet (een overzicht van de thesis zelf wordt gegeven in Figuur N.1). De achterliggende idee is om rekening te houden met de fysische en biologische realiteit van het proces en om het normalisatieprobleem aan te pakken vertrekkende van absolute intensiteiten. De gemeten intensiteiten zullen gemodelleerd worden op een fysisch en experimenteel betekenisvolle manier, om het zodoende mogelijk te maken om absolute waarden van

genexpressie te schatten, in plaats van beperkt te zijn door de relatieve aard van intensiteitsratio's.



Figuur N.1: Organizatie van de thesis. Hoofdstukken die handelen over het eigen onderzoek zijn in zwarte kaders weergegeven..



Figuur N.2: Microroostertechnologie. Bovenste paneel: overzicht van de experimentele procedures betrokken bij een microroosterexperiment, gaande van de productie van roosters tot het eigenlijke uitvoeren van de experimenten. Onderste paneel: data-analysecomponent van een microroosterexperiment.

Hoofdstuk 2: Microroosters

In dit hoofdstuk wordt een overzicht gegeven van de technologische en experimentele principes van microroosters (sectie 2.1), gevolgd door een bespreking van enkele typische datakenmerken en analysetechnieken (sectie 2.2). In het laatste deel van dit hoofdstuk (sectie 2.3) worden enkele toepassingen van microroosters behandeld, die niet gericht zijn op het meer gebruikelijke monitoren van genexpressie.

Technologie

Deze sectie handelt over de technologieën en procedures die betrokken zijn in het uitvoeren van een microrooster experiment (zie Figuur N.2), gaande van de productie van de microroosters die de DNA-***probes*** bevatten (sectie 2.1.1), tot de preparatie van hybridisatieoplossingen (bevatten de ***target*** moleculen) en de eigenlijke hybridisatiereactie en scannen van het rooster (sectie 2.1.2).

De eerste stap in de productie van microroosters is het genereren van probe-oplossingen die fungeren als stocks van het DNA dat op de roosters kan gepositioneerd worden. Tegenwoordig worden ofwel cDNA fragmenten, ofwel synthetische oligonucleotiden (oligomeren) gebruikt als probes voor microroosters. Het eigenlijke printen van de microroosters kan gebeuren via *contact printing* [121,175], de methode die gebruikt werd voor het maken van de eerste microroosters [175] en nog steeds erg populair is, of door *non-contact printing* (inkt jet) [91,178]. De meest kritieke factoren die een invloed hebben op de kwaliteit van de geproduceerde microroosters zijn het gebruikte type van printpin en de karakteristieken van het roosteroppervlak (een glazen plaatje met een coating die toelaat dat het probe-DNA gemakkelijk kan gebonden worden). Daarnaast spelen ook eigenschappen van de geautomatiseerde printer (beweging van de printpinnen en positionering van de microtiterplaten en microroosters), de samenstellingen van de probe-DNA oplossing, en controle over omgevingsfactoren zoals temperatuur en vochtigheidsgraad een belangrijke rol . Het plaatsen van DNA-probes op welomlijnde, discrete posities op een glazen drager mag conceptueel eenvoudig lijken, de precieze en betrouwbare productie van microroosters in de praktijk is niet zonder uitdagingen.

Het uitvoeren van de eigenlijke microroosterexperimenten begint met het bedenken van een gepast experimenteel design, dat zoveel mogelijk biologisch relevante informatie oplevert, en terwijl rekening houdt met de beperkingen van microroostertechnologie, zoals de kostprijs van de experimenten en de beschikbaar van de biologische stalen. De eerste fase in het genereren van hybridisatiestalen is het isoleren en zuiveren van mRNA uit celculturen of weefsels. Wanneer slechts een beperkte hoeveelheid RNA voorhanden is (vb. geisoleerd uit een kleine hoeveelheid tumorweefsel),

wordt gewoonlijk een extra amplificatiestap ingelast. Daarna worden deze stalen gelabeld, i.e. worden fuorochromen geïncorporeerd in de target sequenties. De populairste fluorochromen zijn de carbocyanines Cy5 en Cy3 [231], respectievelijk de 'rode' en de 'groene'. Het hybridisatieproces bestaat uit het incuberen van het gelabelde target-DNA met het probe-DNA dat vastgehecht is op het microrooster: fluorescente target sequenties hybridiseren met complementaire probes. De uitgezonden fluorescentie kan gemeten worden met een confocale laserscanner en is een indicatie van de hoeveelheid geïmmobiliseerd target-DNA.

Verwerking van de data

Het uitvoeren van de experimentele procedures is slechts een eerste fase in een microroosterstudie, de daaropvolgende data-analyse (sectie 2.2) is evenzo belangrijk. Deze sectie bespreekt een typische data-analyse pijplijn zoals geïllustreerd in Figuur N.2, beginnende met beeldanalyse (sectie 2.2.1), gevolgd door normalisatie van de intensiteiten (sectie 2.2.2), en tot slot exploratie van de data op hoger niveau (sectie 2.2.3).

De beeldanalyse van gescande microrooster converteert de bekomen scans naar numerieke waarden, geassocieerd met individuele probe-spots, die dienen als maat voor de hoeveelheid gehybridiseerd target. Dit proces kan onderverdeeld worden in drie stappen: *gridding* (of *addressing*; het toekennen van coördinaten aan elk van de geprinte probes), *segmentatie* (het classificeren van de pixels van het beeld als voorgrond, i.e. behorende tot een spot van probe-DNA, of achtergrond), en *intensiteitsextractie* (het berekenen van voorgrond- en achtergrondintensiteiten voor elke spot op het microrooster voor zowel Cy5 als Cy3).

Normalisatie van de ruwe, geëxtraheerde intensiteiten is een noodzakelijke stap vooraleer verdere analyses worden uitgevoerd die kunnen leiden tot biologische interpretaties (sectie 2.3). In plaats van een exhaustieve lijst te voorzien van alle beschreven methodes, handelt dit deel van het hoofdstuk over typische karakteristieken en gerelateerde problemen van microroosterdata, en enkele van de standaardtechnieken die gebruikt worden om hiermee om te gaan:

- ***Achtergrondcorrectie*** (sectie 2.2.2.1) is de eerste stap van het normaliseren van microroosterdata. De bedoeling is om de 'voorgrond' spotintensiteiten te corrigeren voor achtergrondcontributies, zoals niet-specifieke hybridizatie, residuele Cy5- en Cy3-moleculen, en fluorescentie afkomstig van andere delen van het rooster (*overshining*). Het is algemeen aanvaard dat het effect van achtergrond additief is met respect tot de gemeten spotintensiteiten [34] (achtergrondcorrectie wordt dan ook vaak *achtergrondsubtractie* genoemd). Het is helaas onmogelijk om de echte achtergrond te meten voor elke spot. Als gevolg zijn er

verschillende methodes ontwikkeld om deze achtergrond bij benadering te kwantificeren. In dit deel van het hoofdstuk geven we een korte bespreking van de voor- en nadelen van methodes die gebruik maken van een constante achtergrond, een locale achtergrond, een achtergrondmodel, en het simpelweg werken met de ruwe intensiteiten (i.e. geen achtergrondcorrectie uitvoeren).

- Zoals reeds eerder vermeld is microroostertechnologie fundamenteel ontworpen met het oog op het meten van relatieve genexpressie. Zodoende zijn ***log-ratio's*** (sectie 2.2.2.2), het logaritme van de ratio's Cy5- over Cy3-intensiteiten, de basiseenheden die gebruikt worden om de data te interpreteren. Het wordt aangenomen dat zulke ratio's de grote, spot-gerelateerde variaties in intensiteiten teniet doen. De motivatie voor de logaritmische transformatie is tweevoudig. Microroosterdata vertonen buiten de additieve achtergrond ook multiplicatieve fouten die kunnen opgevangen worden door het nemen van een logaritme. Daarnaast vergemakkelijkt dergelijke transformatie de interpretatie van de berekende ratio's. De relatieve aard van microroosterdata en het gebruik van log-ratio's heeft belangrijke gevolgen voor de experimentele setup van complexere experimenten (i.e. experimenten met meer dan twee biologische condities). De centrale designkeuze is altijd of twee biologische stalen direct (op hetzelfde rooster) of indirect (op verschillende roosters) vergeleken worden. In dit deel van het hoofdstuk bespreken we verder drie standaarddesigns: de *colour-flip*, het *loop design*, en het *reference design*.

- Het gebruik van log-ratio's omzeilt theoretisch gezien alle systematische fouten die afkomstig zijn van spots, printpinnen en roosters. De meeste normalisatiestrategieen voor microroosters zijn daarom gefocust op het verwijderen van ***fluorochroom-gerelateerde verschuivingen*** (sectie 2.2.2.3). Dergelijke systematische variaties veroorzaken een significante distortie in de distributie van log-ratio's, en zijn het gevolg van verschillende factoren, voornamelijk de fysische eigenschappen van de carbocyanines en de efficiëntie van de incorporatie van deze labels, maar ook verschillen in de hoeveelheid aan input RNA, en scanner-specifieke excitatie- en meeteigenschappen. Gewoonlijk worden *alle genen* gebruikt om te compenseren voor een fluorochroom gerelateerde verschuiving. Men neemt aan dat dit niet onredelijk is omdat **1)** slechts een relatief kleine proportie van alle genen significant van expressie zal variëren tussen twee mRNA stalen van distincte biologische condities, en **2)** dat er symmetrie is in de hoeveelheid op- en neergereguleerde genen. In de praktijk is de

geobserveerde verschuiving niet constant binnen een rooster en over verschillende roosters heen, wat aanleiding heeft gegeven tot intensiteitsafhankelijke herschalingsprocedures (e.g. LOWESS [226]), dewelke in dit deel verder besproken worden.

Nadat de data genormaliseerd zijn, kunnen verdere analyses gebeuren met het doel van biologisch betekenisvolle resultaten te bekomen. De biologische en biomedische vraagstukken die bestudeerd worden kunnen vrij uiteenlopend zijn, zodat verscheidene methodes en algoritmes uit het domein van de statistiek, *data mining* en *machine learning* hun weg gevonden hebben naar de verwerking van microroosterdata. Dit deel van het hoofdstuk geeft een bondig overzicht van enkele van de meest wijdverbreide data-exploratiemethodes, zoals de selectie van genen met significant differentiële expressie, clustering van genexpressieprofielen, clustering van de geteste biologische condities, classificatie van de geteste biologische condities en inferentie van regulatorische (genetische) netwerken.

Andere toepassingen

Microroosters worden voornamelijk gebruikt om de expressieprofielen van specifieke celtypes en weefselstalen te bestuderen. De differentiële labels en het daaruit volgende relatieve karakter van de experimenten, maakt microroosters echter uitermate geschikt voor andere types van genomische analyses. In deze sectie worden twee van de meest courante applicaties besproken, namelijk *Comparatieve GenoomHybridisatie* en *Chromatine-ImmunoPrecipitatie* op microroosters (respectievelijk *CGH-arrays* en *ChIP-chip*). CGH is een methode die toelaat sites met een variabel kopienummer te identificeren en in kaart te brengen voor het hele genoom. ChIP-chip is een populaire technologie die toelaat de bindingsplaatsten van DNA-bindingsproteïnen op het DNA te bepalen.

Hoofdstuk 3: Evaluatie van ANOVA-normalisatie

In dit hoofdstuk werd het gebruik van ANOVA voor microroosternormalisatie geëvalueerd. Omdat er geen directe manier bestaat om een normalisatieprocedure te beoordelen (de daadwerkelijke expressieniveaus zijn immers niet gekend), werden significant differentiële genen geselecteerd o.b.v. ANOVA-genormaliseerde data en vergeleken met genen die geïdentificeerd werden als significant differentieel tot expressie komend o.b.v. de gemeten log-ratio's [132]. Om de invloed van de gebruikte selectieprocedure te verminderen, werden de ANOVA-resultaten vergeleken met die van drie verschillende methodes die steunen op het gebruik van log-ratio's.

Een eerste deel van dit hoofdstuk (sectie 3.1) beschrijft de principes van de op ANOVA gebaseerde, normalisatie van microroosters. Het tweede deel (sectie 3.2) doet hetzelfde voor de op log-ratio's gebaseerde methodes, die gebruikt werden voor de identificatie van genen met differentiële expressie. In een laatste deel (sectie 3.3) worden de resultaten weergegeven en besproken.

ANOVA modellen voor normalizatie

ANOVA (ANalysis Of VAriance) wordt steeds meer gebruikt voor de normalisatie van microroosterdata [104,113,221]. Een ANOVA-normalisatie modelleert de gemeten expressieniveaus van elk gen als lineaire combinaties van predictorvariabelen, die, in de context van deze studie, de belangrijkste bronnen van variatie in een microroosterexperiment vertegenwoordigen (e.g. microrooster, fluorochroom, conditie, printpin, etc.). De parameterisaties van de *GC*-variabele (*genxconditie* interactie) kunnen beschouwd worden als genormaliseerde data: ze beschrijven voor elk gen de conditie-geaffecteerde verandering in expressie. Door het fitten van een ANOVA-model bekomt men bovendien een residuele foutendistributie, een schatting van de experimentele foutendistributie. Deze residu's kunnen gebruikt worden om significante genen te identificeren door betrouwbaarheidsintervallen op te stellen op het verschil in *GC*-factorniveaus. Meestal vertoont deze residuverdeling echter grote afwijkingen van normaliteit. In dat geval is het gebruik van Gaussiaanse statistiek ongepast; 'bootstrapping' [50,67,68] (voor het eerst op microroosters toegepast door Kerr *et al*., 2000 [113]), een virtuele herbemonsteringsmethode (*resampling*), kan dan gebruikt worden als alternatief voor statistische inferentie.

Verschillende ANOVA-modellen werden geëvalueerd. Deze modellen verschilden van elkaar in het aantal additionele interactievariabelen (voor het beschrijven van spot-gerelateerde variabiliteit). Het model met de beste performantie bestond uit een eigen adaptatie van eerder beschreven modellen [111-113], en werd dan ook gebruikt in de vergelijkende studie:

$$I_{ijklm} = \mu + G_i + C_j + A_k + D_l + R(G)_{m(i)} + (AG)_{ki} + (GC)_{ij} + \varepsilon_{ijklm} \quad \text{(N.1)}$$

In dit model is μ het gemiddelde signaal over alle intensiteiten heen, stelt G_i het effect van het i^{de} *gen* voor, stelt C_j het effect van de j^{de} *conditie* voor, stelt A_k het effect van het k^{de} *rooster* voor, stelt D_l het effect van de l^{de} *fluorochroom* voor, stelt $(GA)_{ik}$ de interactie voor tussen het i^{de} gen en het k^{de} rooster, $(GC)_{ij}$ de interactie tussen het i^{de} gen and de j^{de} conditie. De foutentermen ε_{ijkl} worden verondersteld identiek verdeeld en onafhankelijk te zijn. Het $R(G)_{m(ik)}$ effect representeert de m^{de} replica van een gen dat meerdere malen gespot werd op elk rooster. Deze typische, geneste structuur werd gekozen om variabiliteit, die kan toegewezen worden aan de

probe-oplossingen voor genen die meerdere malen gespot worden. De probes van een enkel gen op verschillende roosters stammen immers van dezelfde PCR-reactie of dezelfde oligo-set.

Identificeren van differentiële expressie o.b.v. log-ratio's

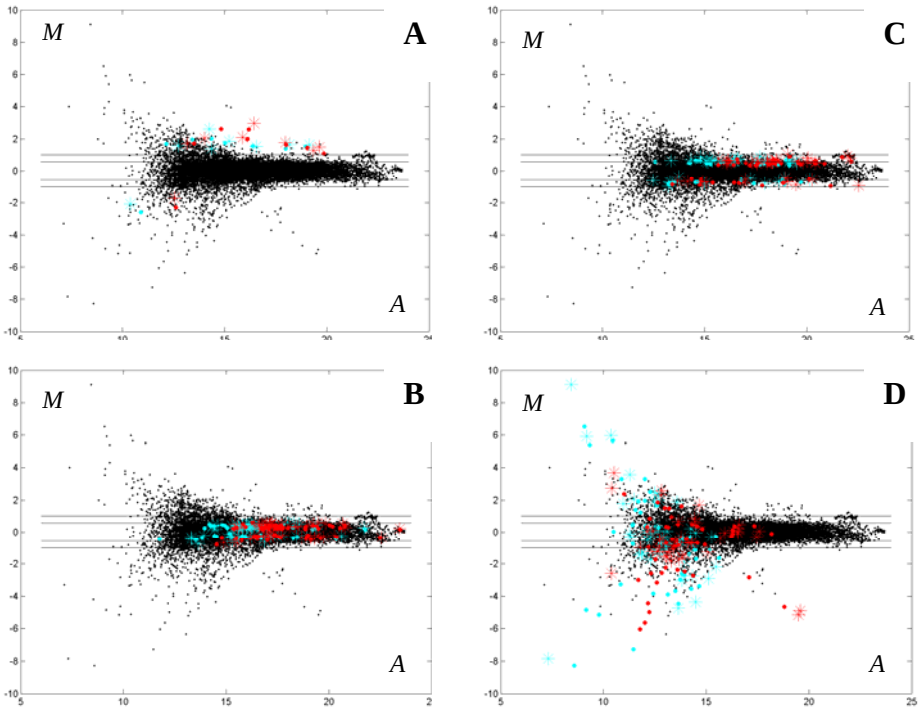
Op basis van een uitgebreide literatuurstudie konden bestaande methodes ingedeeld worden in afzonderlijke klassen, naargelang de gebruikte teststatistiek, distributie van de nulhypothese en hun onderliggende assumpties. Drie methodes, die elk kunnen beschouwd worden als vertegenwoordiger van een verschillende klasse, werden geselecteerd: een 'fold' test [157], een gepaarde *t-test* [12], en SAM (Significance Analysis of Micro-arrays [204]). Deze drie methodes werden exhaustief bestudeerd en vergeleken met de ANOVA-normalisatie. In deze paragraaf wordt een korte beschrijving gegeven van elk van deze drie, distincte methodes.

De ***'fold' test*** is een eenvoudige selectieprocedure die gebruik maakt van een arbitrair gekozen drempelwaarde; ze is gebaseerd op het principe dat een grotere verhouding ('fold change') tussen test en referentie met grotere zekerheid kan beschouwd worden als een sterkere respons t.o.v. omgevingssignalen dan een kleinere verhouding. Voor elk gen wordt een log-ratio berekend, en indien metingen gerepliceerd zijn wordt een gemiddelde ratio berekend. Genen waarvan de ratio's een bepaalde drempelwaarde overschreiden (het meest gebruikelijke is tweevoud) worden beschouwd als diferentieel tot expressie komend [157].

Een ***t-test*** is geschikter dan een eenvoudige 'fold' test om tot statistisch relevante besluiten te komen i.v.m. de al dan niet differentiële expressie van een gen. Als standaardstatistiek voor het vergelijken van twee populaties (i.e. gemeten intensiteiten in test vs. gemeten intensiteiten in referentie), houdt ze, in tegenstelling tot de 'fold' test, niet alleen rekening met het verschil tussen de gemiddelde logratio's, maar ook met de consistentie van de metingen, gebruikt om deze gemiddelde logratio's te bekomen. Een gepaarde *t-test* zorgt voor nog meer sensitiviteit ('power') doordat intrinsiek rekening gehouden word met variatie over spots en arrays. Het theoretische voordeel van een (gepaarde) *t-test* t.o.v. de 'fold' test, is dus dat kleinere verschillen tussen test en referentie als significant kunnen beschouwd worden wanneer de expressieniveaus voor het betrokken gen met grote nauwkeurigheid (hoge consistentie) werden gemeten, terwijl grotere verschillen als niet-significant kunnen worden geduid wanneer de metingen met lage consistentie werden bekomen. In deze evaluatie werd de gepaarde *t-test* van Baldi en Long, 2001, gebruikt [12].

SAM (*Significance Analysis of Micro-arrays*) berekent voor elk gen een zogeheten 'Relative difference d(i)', die kan beschouwd worden als een gemodificeerde *t-test*statistiek. Een groter verschil met de (gepaarde) *t-test* echter, is dat SAM geen assumpties maakt m.b.t. de distributie van de

nulhypothese. De SAM-procedure is gebaseerd op een niet-parametrische rangstatistiek: i.p.v. p-waarden te berekenen, worden differentiële genen geïdentificeerd via ordening en permutatieanalyse. Een extra voordeel van deze methode is dat een schatting kan gemaakt worden van het aantal vals-positieven. Voor meer technische informatie wordt verwezen naar het oorspronkelijke artikel van Tusher *et al*., 2001 [204].



Figuur N.3: Gedetailleerde voorstelling van verschillende groepen van geselecteerde genen. Gemiddelde log-intensiteiten *A* zijn uitgezet tegen LOWESS-genormaliseerde log-ratio's *M* voor beide microroosters in elke plot. Zwart: alle 3785 genen; rood en cyaan: geselecteerde genen op de 1ste resp. 2de array. Horizontale lijnen markeren de 1,5- en 2-voudige over- en onderexpressiegrenzen. De aangeduide genen werden geselecteerd door **A)** alle methodes, **B)** de gepaarde *t-test*, **C)** gepaarde *t-test* en SAM en **D)** 'fold' test en ANOVA-bootstrap.

Resultaten en conclusies

Door elke methode op eenzelfde dataset toe te passen, en de karakteristieken van de verschillende groepen van genen te vergelijken, konden besluiten gevormd worden m.b.t. de performantie van, en inherente verschillen tussen, deze vier selectieprocedures. In Figuur N.3 worden de belangrijkste bevindingen geïllustreerd.

Een van de meest opmerkelijke vaststellingen was de lage graad van overeenkomst tussen de verschillende methodes: slechts acht genen werden door elke methode gedetecteerd (Figuur 1.1, plot A). Genen die alleen door de gepaarde *t-test* werden geselecteerd (Figuur 1.1, plot B), waren erg consistent gemeten, maar ogenschijnlijk te weinig differentieel tot expressie komend, om biologisch relevant te zijn. De *t-test* heeft mogelijk een veel te lage sensitiviteit, gezien het kleine aantal replica's. De genen die zowel door de *t-test* als door SAM werden geïdentificeerd, zijn weergegeven in Figuur 1.1, plot C. Deze metingen waren consistent en de gemiddelde logratio's voldoende verschillend van nul. Tot slot zijn de genen, die zowel door ANOVA-bootstrap, als de 'fold' test werden geselecteerd, weergegeven in Figuur 1.1, plot D. Door hun hoge gemiddelde expressiewaarde worden deze genen door de selectieprocedures als significant beschouwd, maar de consistentie van deze metingen was opmerkelijk laag. Bovendien was er een sterke heteroscedasticiteit in de data (grotere variantie voor lagere intensiteiten), waardoor de bekomen ratio's voor lagere intensiteiten meer onbetrouwbaar werden, een verschijnsel dat nefast is voor de 'fold' test. Om dezelfde reden werd de variatie bij lagere intesiteiten systematisch onderschat door de bootstrap-gebaseerde confidentie-intervallen, en overschat bij hogere intensiteiten, met als resultaat ongetwijfeld sterke vertegenwoordiging van zowel valspositieven als valsnegatieven in de geselecteerde genen.

Zoals dikwijls het geval met statistische analyses, lijkt de betrouwbaarheid van de gebruikte methode hier sterk afhankelijk van de dataset: SAM presteerde duidelijk beter dan de andere methodes omdat de dataset beter voldeed aan de onderliggende assumpties. Hoewel de ANOVA-gebaseerde selectieprocedure duidelijk mindere prestaties leverde, werd toch besloten op deze methode verder te bouwen voor het genereren van een normalisatie- en identificatiemethodologie. Met het oog op meer complexe experimentele designs (t.o.v. colour-flip), die in het kader van de genetische netwerkinferentie dienen geanalyseerd te worden, biedt dit hele concept theoretisch gezien immers enkele belangrijke voordelen:

- Inherent aan ANOVA is een normalisatie die, in tegenstelling tot de meer gebruikelijke slide-per-slide procedures, verschillende bronnen van variatie over het gehele experiment in rekening brengt door informatie te extraheren uit alle metingen. Een goed normalisatie is niet onbelangrijk, aangezien de kwaliteit van

netwerkinferentieprocedures grotendeels zal afhangen van de invoerdata.

- De residu's die bekomen worden na het fitten van het ANOVA-model kunnen gebruikt worden voor verdere statistische inferentie, zoals het identificeren van genen met differentiële expressie of het opsporen van inconsistente metingen.

Een groot nadeel van de in de literatuur beschreven ANOVA-modellen is echter dat, voor elk experimentdesign, een andere analytische oplossing moet berekend, én geïmplementeerd worden. Een eigenschap die verder werd onderzocht in hoofdstuk 4.

Hoofdstuk 4: Generische ANOVA-modellen

Om aan beschreven tekortkomingen te beantwoorden, werd een analyseprocedure voor microroosterdata gecreëerd, gebaseerd op een generisch ANOVA-model. Dit hoofdstuk beschrijft achtereenvolgens de problemen met beschreven ANOVA-modellen en hun toepassing op verschillende experimentele designs (sectie 4.1), de ontwikkeling van generische (toepasbaar op eender welk design) ANOVA-modellen voor microroosternormalisatie (sectie 4.2) en de implementatie van een dergelijk model in een gebruiksvriendelijke web-interface (section 4.3). Enkele belangrijke observaties die voortkwamen uit dit onderzoek worden besproken in het laatste deel (sectie 4.4).

ANOVA-modellen en experimentdesign

In dit deel van het hoofdstuk wordt beschreven hoe de parameters van een ANOVA-model geschat kunnen worden, hoe deze schatters beïnvloed worden door het design van het experiment. Deze principes worden geïllustreerd aan de hand van drie simpele, maar conceptueel verschillende designs: een *colour-flip* design, een *reference* design, en een *loop* design. Het belangrijkste designprobleem van ANOVA-modellen is inherent aan de microroostertechnologie: het aantal condities dat tegelijkertijd kan gemeten worden op eenzelfde microrooster is beperkt tot twee. Rooster en conditie zullen daarom nagenoeg nooit orthogonaal zijn, met uitzondering van vb. een simpel *colour-flip* design.

Een *reference* design is wat dat betreft veel complexer. Conditie-effecten zijn in dat geval 'volledig verward' met fluorochroomeffecten, aangezien elke conditie maar gelabeld is met een type fluorochroom. Men kan dus niet zowel conditie-effecten als fluorochroom-effecten in rekening brengen in het model wanneer men een *reference* design wil analyseren. Een alternatief hiervoor zijn loop designs, die conditie- en fluorochroom-gerelateerde

effecten gedeeltelijk ontwarren en bovendien meer vrijheidsgraden overlaten voor schatting van de experimentele fout en dus een betere basis bieden voor verdere statistische inferenties. Afhankelijk van het gebruikte design kunnen effecten ook 'gedeeltelijk verward' zijn. In dat geval is het wel mogelijk om schatters te bekomen voor elk effect, met het nadeel dat deze gecorreleerd zullen zijn.

De geschiktheid van eender welk ANOVA-model voor de normalisatie van microroosterdata wordt bepaald door de typische eigenschappen van het gebruikte experimentdesign, hoe deze gerelateerd zijn aan de variabelen in het model, en het aantal vrijheidsgraden dat overblijft om de experimentele foutenverdeling te benaderen. Bovendien zijn de drie designs die in dit deel van het hoofdstuk besproken werden verre van de enige die gebruikt worden voor microroosterexperimenten. Meer nog, vaak dienen zij enkel als bouwstenen voor complexere designs, zodat de evaluatie van ANOVA-modellen voor elk ander design een vervelende taak wordt.

Generische ANOVA-modellen

Dit deel van het hoofdstuk beschrijft het ontwerp van generisch ANOVA-modellen voor microroosternormalisatie. Deze modellen bieden verscheidene voordelen t.o.v. de modellen van Kerr *et al*. [111-113]:

- Het belangrijkste voordeel (en de primaire focus tijdens de constructie ervan) is het generisch karakter met respect tot het experimentdesign, i.e. het kan elk type van design normaliseren in een enkele analyse. Om te compenseren voor conditieafhankelijke variatie werd geopteerd voor een *arrayxdye* interactievariabele, aangezien het gebruik van een conditiefactor de analytische oplossingen van het model afhankelijk zou maken van het experimenteel design.
- Incorporatie van een *batch* variabele: een batch kan gedefinieerd worden als een collectie van slides die dezelfde set van genen (representatief voor een deel van het genoom) bevatten. Deze factor is van toepassing wanneer de gehele set van onderzochte genen te groot is om op een enkel rooster gespot te worden.
- Incorporatie een pin-variabele: om 'overfitting' tegen te gaan wordt spottingvariabiliteit per pingroep en niet per individuele spot gemodelleerd.

Twee modellen warden ontwikkeld die aan deze kenmerken voldoen. Ze verschillen in de manier waarop de pingroep variabele gestructureerd is met respect tot de batch en array variabelen. In een eerste model is de pingroep variabele verondersteld genest te zijn in de batch variabele, m.a.w. is een pingroep effect constant voor alle microroosters van dezelfde batch. De

verantwoording hiervoor kan gevonden worden in het feit dat microroosters in serie geprint worden en met dezelfde printpinnen, zodat onregelmatigheden die aan deze pinnen te wijten zijn gelijkaardig zullen zijn voor roosters van dezelde serie. Het tweede model veronderstelt een verschillend effect voor elke pin op elk rooster. Dit is een meer algemene modellering die slechts enkele vrijheidsgraden meer vereist dan het eerste model. Dit tweede model (N.2) is hieronder weergegeven, samen met de analytische oplossingen (N.3) van de weerhouden parameters:

$$I_{ijklmn} = \mu + B_m + D_l + A_{k(m)} + AD_{kl(m)} + PA_{nk(m)} + G_{i(n(m))} + GC_{ij(n(m))} + \varepsilon_{ijklmn} \tag{N.2}$$

$$\hat{\mu} = I_{.....}$$

$$\hat{B}_m = I_{....m.} - I_{.....}$$

$$\hat{A}_{k(m)} = I_{..k.m.} - I_{....m.}$$

$$\hat{D}_l = I_{...l..} - I_{.....}$$

$$\hat{AD}_{kl(m)} = I_{..klm.} - I_{..k.m.} - I_{...l..} - I_{.....}$$

$$\hat{PA}_{nk(m)} = I_{..k.mn} - I_{..k.m.}$$

$$\hat{G}_{i(n(m))} = I_{i.....} - I_{.....} - \underset{i}{\text{avg}}\left[\hat{B}_m\right]$$

$$\hat{GC}_{ij(n(m))} = I_{ij....} - I_{i.....} - \underset{ij}{\text{avg}}\left[\hat{D}_l + \hat{A}_{k(m)} + \hat{AD}_{kl(m)} + \hat{PA}_{nk(m)}\right] \tag{N.3}$$

MARAN: een webapplicatie voor de normalisatie van microroosterdata

Normalisatiemodel (N.2) werd gebruikt als uitgangspunt voor MARAN , een geïntegreerde analyseprocedure voor microroosterdata die online beschikbaar werd gesteld [72] (in samenwerking met ir. B. Coessens; http://www.esat.kuleuven.be/maran). Een overzicht van de functionaliteit van MARAN is gegeven in Figuur N.4.

Normalisatie van microroosterdata met MARAN is gebruiksvriendelijk en redelijk vanzelfsprekend. Enkel predictorvariabelen, die relevant zijn voor bestudeerde experimentdesign, worden automatisch in rekening gebracht. Alle andere factoren kunnen door de gebruiker in de analyses meegenomen of weggelaten worden.

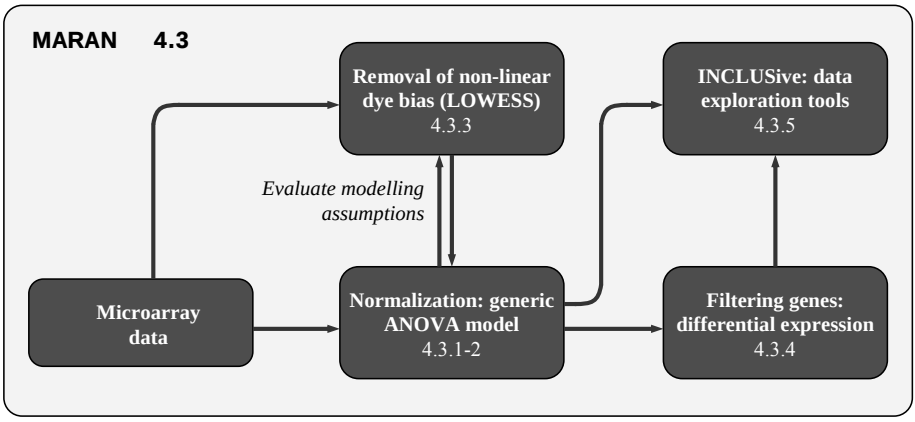


Figure N.4: Schematic representation of the MARAN web application. After the data have been uploaded, they can be normalized by means of a generic ANOVA model, optionally with a preceding LOWESS step to remove nonlinear dye biases. The model and/or LOWESS procedure can be rerun at any time based on an evaluation of model fitting results (evaluation of modelling assumptions). A module for filtering the data and a module that integrates MARAN into INCLUSive (for e.g. clustering, motif detection), are also available.

Daarnaast bevat deze implementatie enkele bijkomende functionaliteiten, zoals figuren om de fit van het model te beoordelen, en een optie om eventuele niet-lineare fluorochroom verschuivingen te verwijderen m.b.v. een LOWESS-fit [226], en een module om genen met een significante verandering in expressie te selecteren. Zoals reeds eerder vermeld kunnen, na de ANOVA normalisatie, de bekomen residu's gebruikt worden voor een statistische analyse van de model parameters. MARAN bevat ook een module om genen met een significante verandering in expressie te detecteren, steunend op de assumptie dat de foutentermen normaal verdeeld zijn, of door gebruik te maken van *bootstrap*-technieken [50,67,68] om de foutenverdeling te benaderen.

MARAN werd ondergebracht in een vernieuwde versie van INCLUSive [41,198] (http://www.esat.kuleuven.be/inclusive): een suite van, grotendeels op ESAT-SCD ontwikkelde, algoritmes en methodes voor genexpressieanalyse en de ontdekking van regulatorische motieven. Alle applicaties van INCLUSive zijn beschikbaar via verschillende webpagina's en als webservices.

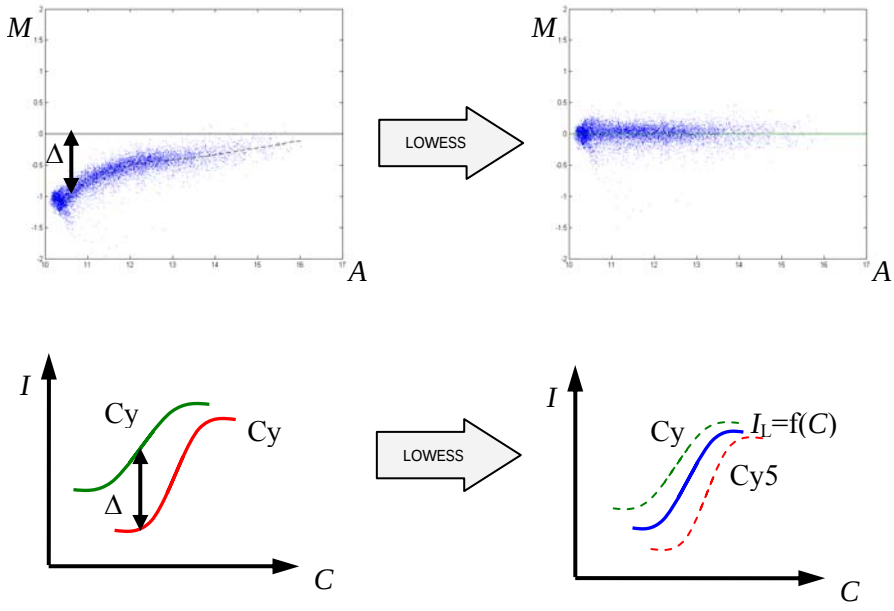
Conclusies

In het laatste deel van dit hoofdstuk bespreken we nog enkele kritieke punten aangaande het gevoerde onderzoek, zoals het belang van replica's (in feite een centrale kwestie ongeacht de gekozen normalisatiemethode) en de implicaties ervan voor ANOVA-gebaseerde microroosteranalyse.

Daarnaast leidde de toepassing van dit onderzoek tot enkele vreemde vasstellingen en noemenswaardige complicaties met betrekking tot de LOWESS-procedure. De bedoeling van het werk zoals beschreven in dit hoofdstuk was een globaal normalisatiemodel te ontwikkelen waardoor een residudistributie bekomen kan worden die voldoet aan onderliggende assumpties, i.e. een model dat een goede beschrijving biedt van de data. Helaas is het voor lineaire modellen (zoals ANOVA) onmogelijk om alle curvilineariteiten uit de data te verwijderen, zelf na een slide-per-slide intensiteitafhankelijke normalisatie (e.g. LOWESS). De reden hiervoor is niet zozeer dat in de MARAN-procedure conceptuele fouten gemaakt worden, maar wel dat huidige lineaire en niet-lineaire normalisatiemethodes niet in staat zijn de niet-lineariteit, inherent aan microroosterdata, op een adequate wijze te compenseren.

Een verklaring voor de niet-lineaire trends -en de specifieke manier waarop deze zich manifesteren- kan gevonden worden in de aanname dat de relatie tussen target-concentratie en intensiteit niet over het hele bereik lineair is, maar verzadigingskarakteristieken kan vertonen voor hogere en lagere intensiteiten. Ze biedt eveneens een verklaring voor het feit dat een niet-lineaire, slide-per-slide normalisatie zoals LOWESS, vooraleer het fitten van een lineair normalisatiemodel (e.g. ANOVA), niet in staat is volledig te compenseren voor de geobserveerde niet-lineariteiten. Zoals geïllustreerd in Figuur N.5 kunnen deze methodes enkel de niet-lineariteiten tussen de Cy3- en Cy5-intensiteitsmetingen verwijderen, maar nooit tussen de gemeten intensiteiten en de fluorochroom/cDNA-concentratie. Wanneer hierna een lineair normalisatiemodel gefit wordt, dat verschillende bronnen van systematische variabiliteit over het gehele experiment in rekening brengt (e.g. een ANOVA-model zoals dat van MARAN), zal dit leiden tot residu's waarin nog steeds uitgesproken niet-lineaire trends worden waargenomen.

De constructie van een globaal niet-lineair normalisatiemodel, uitgaande van deze bevindingen, wordt in detail beschreven in hoofdstuk 5.



Figuur N.5: Een verklaring voor hardnekkige niet-lineariteiten. Twee verschillende saturatiecurves (cy3 en Cy5) beschrijven de relatie tussen fluorofoorconcentratie en gemeten intensiteit. Niet-lineaire, microrooster-gebaseerde normalisatieprocedures (e.g. LOWESS) herschalen de Cy3 en Cy5 intensiteit-concentratie curves tot een nieuwe functie, die in feite gecentreerd is tussen de Cy3 en Cy5 curves. Ze verwijderen dus de niet-lineaire relatie tussen beide fluorofoorintensiteiten, maar niet tussen de fluorofoorintensiteiten en de overeenkomstige fluorofoorconcentraties.

Hoofdstuk 5: Een calibratiemethode voor microroosters

In dit hoofdstuk wordt een nieuwe methode besproken voor de normalisatie van microroosterdata [13]. Deze aan pak steunt op het gebruik van externe controles (*spikes*; een bespreking kan gevonden worden in sectie 5.1) om een calibratiemodel te fitten op de data. Het calibratiemodel dat de kern is van deze normalisatieprocedure (sectie 5.2) bestaat uit twee componenten, die enerzijds de hybridisatie van gelabelde targetmoleculen op hun complementaire probes, en anderzijds de meting van fluorescentiesignalen van deze gehybridiseerde targets beschrijven. De parameters van het model en de geïncorporeerde foutenverdelingen worden geschat op basis van metingen van externe controles, en kunnen gebruikt worden om absolute expressieniveaus te bekomen voor elk gen in elk van de biologische condities die in het experiment bestudeerd werden.

De resultaten die bekomen werden door het toepassen van deze methode op een publiek beschikbare dataset worden eveneens besproken (sectie 5.3). We tonen aan dat de procedure in staat is de typische niet-lineariteiten van microroosterdata te verwijderen, zonder enige assumpties te maken met betrekking tot de distributie van verschillen in genexpressie tussen biologische condities (i.e. zonder te steunen op de GNA). In een volgend deel wordt de methode vergeleken met de combinatie LOWESS en ANOVA. Aangezien het model targetconcentratie linkt aan gemeten intensiteit, tonen we bovendien aan hoe abolute waardes voor expressie kunnen bekomen worden. Tot slot bespreken we nog de invloed van de veel gebruikte lokale achtergrondcorrectie in relatie tot de ontwikkelde methode.

Mathematische modellen en algoritmes

De hybridisatiereactie die vervat zit in het calibratiemodel relateert de hoeveelheid van gehybridiseerd target (x_s) met de concentratie van het overeenkomstig transcript (x_0) in de hybridisatieoplossing. De hybridisatieconstante wordt constant geacht voor alle metingen afkomstig van hetzelfde rooster.

$$x_0 + s \overset{K_A}{\Leftrightarrow} x_s \tag{N.4}$$

Er wordt verondersteld dat deze reactie zijn evenwicht bereikt heeft wanneer de eigenlijke metingen plaatsvinden, en dat ze kan gemodelleerd worden met een eerste-orde benadering (in de praktijk komt dit neer op de veronderstelling dat x_0 constant is). De hoeveelheid geprint DNA van een spot die beschikbaar is voor hybridisatie daarentegen neemt wel af met een stijgende hoeveelheid aan gehybridiseerd target ($s = s_0 - x_s$, met s_0 de 'spot capaciteit' of maximale hoeveelheid probe), zodat bij thermodynamisch evenwicht kan geschreven worden:

$$\frac{x_s}{x_0\left(s_0 - x_s\right)} = K_A \tag{N.5}$$

De spotcapaciteit s_0 volgt een zekere verdeling rond een gemiddelde spotcapaciteit μ_s: $s_0 = \mu_s + \varepsilon_s$ of $s_0 = \mu_s e^{\varepsilon_s}$ waar $\varepsilon_s \sim \mathcal{N}\left(0, \sigma_s\right)$ de spotfout is. Welke distributie het meest geschikt is, zal grotendeels afhangen van het type microrooster en de printprocedure die gebruikt werd. De spotparameters μ_s en σ_s kunnen gelijk beschouwd worden voor alle metingen afkomstig van een microrooster, of verschillend op basis van pingroep.

Een tweede component van het model is de saturatiefunctie, dewelke de relatie beschrijft tussen de gemeten intensiteit y en de hoeveelheid aan gelabeld target x_s dat gehybridiseerd is op een enkele spot van het rooster:

$$y = p_1 x_s e^{\varepsilon_m} + p_2 + \varepsilon_a \tag{N.6}$$

Deze saturatiefunctie is een simpele lineaire vergelijking die een additieve en multiplicatieve fout op de intensiteiten in rekening brengt, respectievelijk $\varepsilon_a \sim \mathcal{N}(0, \sigma_a)$ en $\varepsilon_m \sim \mathcal{N}(0, \sigma_m)$ (dit type van functie werd reeds gebruikt in andere normalisatiestrategieën [62,98,165]). De parameters p_1 en p_2 zijn specifiek voor elke combinatie van microrooster en fluorochroom.

De modelparameters worden geschat voor elk microrooster afzonderlijk, gebaseerd op de gemeten intensiteiten y van de externe controles en hun gekende concentratie x_0 in de hybridisatieoplossing. Schatters voor σ_m en σ_a kunnen relatief gemakkelijk bekomen worden. Schatters voor alle andere parameters kunnen bekomen worden door een kleinste-kwadraten oplossing, met name door de variatie (*sum of squares*) van de spotfouten ($SSE_s = \sum_i \varepsilon_{s,i}^2$) te minimaliseren met betrekking tot $p_{1,Cy3}$, $p_{2,Cy3}$, $p_{1,Cy5}$, $p_{2,Cy5}$ en K_A. De individuele spotfouten die nodig zijn om deze te berekenen voor een gegeven set van modelparameters zijn evenwel ongekend. Ze worden geschat door volgende kostfunctie te minimaliseren voor elk paar van metingen die afkomstig zijn van dezelfde spot:

$$Q_{estim} = Q_{estim}^{Cy3} + Q_{estim}^{Cy5} \tag{N.7}$$

Met:

$$Q_{estim}^{D} = \underset{\varepsilon_m, \varepsilon_a}{\arg\min} \left(\left(\frac{\varepsilon_m}{\sigma_m \sqrt{2}} \right)^2 + \left(\frac{\varepsilon_a}{\sigma_a \sqrt{2}} \right)^2 \right)_D \tag{N.8}$$

$$D = Cy3, Cy5$$

Gegeven vergelijkingen (N.5) en (N.6)

De bekomen parameterwaarden kunnen gebruikt worden om een $x_0(i, j)$ (i.e. het expressieniveau van gen *i* in de biologische conditie *j*) te schatten gebaseerd op alle metingen die bekomen werden voor deze combinatie van gen en conditie. Hoewel elk microrooster zijn eigen set van parameters heeft, kan deze normalisatie niettemin beschouwd worden als zijnde 'globaal'. Immers, voor elke combinatie van een gen en een geteste conditie wordt een enkele, absolute expressiewaarde berekend, ongeacht het aantal

microroosters, of het aantal gerepliceerde probes op een rooster, waarop deze combinatie werd gemeten. Het formaat van de resultaten van dergelijke normalisatie is dus vergelijkbaar met de *genexcondition* interactiefactor van de ANOVA-modellen in hoofdstuk 3 en hoofdstuk 4. De $x_0(i,j)$ worden geschat door volgende objectfunctie te minimaliseren:

$$Q_{norm} = \sum_{C} \sum_{S_j} Q_{norm}^{S_j(k)} \tag{N.9}$$

Met:

$$Q_{norm}^{S_j(k)} = \left(\underset{\varepsilon_m, \varepsilon_a}{\arg\min} \left(\left(\frac{\varepsilon_m}{\sigma_m \sqrt{2}} \right)^2 + \left(\frac{\varepsilon_a}{\sigma_a \sqrt{2}} \right)^2 \right) + \left(\frac{\varepsilon_s}{\sigma_s \sqrt{2}} \right)^2 \right)_{S_j(k)} \tag{N.10}$$

Gegeven vergelijkingen (N.5) en (N.6)

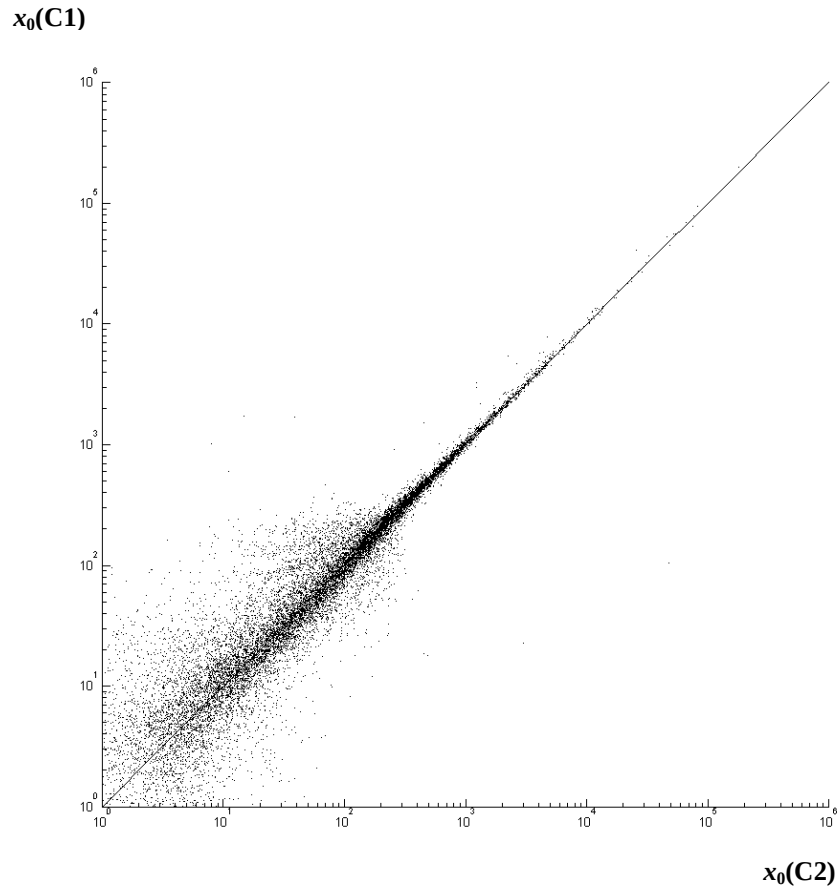
Toepassing en resultaten

De beschreven normalisatiemethode werd geëvalueerd met een publiek beschikbare dataset [96], bestaande uit 14 hybridisaties. Dit specifieke experiment had verschillende eigenschappen die het uitermate geschikt maakten voor de validatie van onze methode, met name:

1. Ze bevatten de noodzakelijke probes om externe controles, die vereist zijn om de parameters van het calibratiemodel te schatten, in de experimenten te incorporeren.
2. Het experimentdesign bevatte slechts een biologische conditie. Elk microrooster bevatte dus een *self-self* hybridisatie.
3. Alle microroosters werden voorzien van een extra set externe controles.

Doordat voor het hele experiment slechts expressiewaarden gemeten werden voor een en dezelfde biologische condities, kon het normalisatiepotentieel van de methode geëvalueerd worden door gebruik te maken van *mock* designs. Een voorbeeld hiervan is weergegeven in Figuur N.6, waar de geschatte expressiewaardes van ca. 19.000 genen geplot zijn voor twee hypothetische condities, afkomstig van een colour-flip design. Aangezien beide condities in werkelijkheid een en dezelfde zijn, duidt de centrering van de punten rond de bissectrice erop dat de methode op een adequate wijze kan omgaan met de typische niet-lineariteiten van microroosterdata. Door gelijkaardige designs te normaliseren met een ANOVA-model,

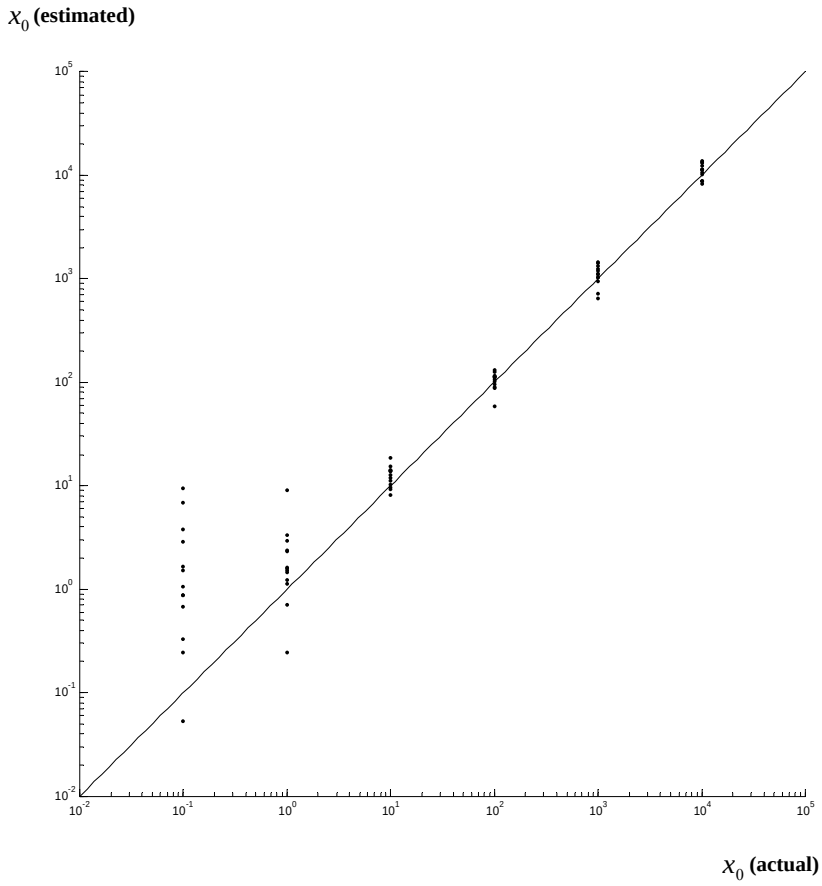
voorafgegaan door een LOWESS-fit, werd de methode vergeleken met meer standaard normalisatiestrategieën.



Figuur N.6: Normalizatie van niet-lineaire artefacten. Geschatte expressieniveaus voor C1 zijn uitgezet tegen die van C2 na normalizatie van een hypothetisch *colour*-flip experiment. C1 en C2 zijn in feite dezelfde biologische conditie. De centrering van datapunten rond de bisectrice is een teken dat de typische microrooster niet-lineariteiten genormaliseerd werden.

De geschatte target concentraties zouden moeten vergeleken worden met de daadwerkelijke concentraties in de hybridisatieoplossing om hun accuraatheid te verifiëren. Dit doen voor de hele populatie van transcripten is onmogelijk, aangezien deze concentraties voor de meeste genen ongekend zijn. De gebruikte dataset bevatte echter een extra set van niet-commerciële controles waarvan deze concentraties wel gekend zijn. Figuur N.7 toont aan dat, met uitzondering van de allerlaagste concentraties, de geschatte waardes goed overeenkomen met de echte concentraties in de hybridisatieoplossing.

Door verschillende factoren, zoals consistente spotfouten of genspecifieke hybridisatie-efficiënties, kunnen de geschatte waarden wel onderhevig zijn aan gengebonden herschalingen. Ze kunnen niettemin geïnterpreteerd worden als absolute niveaus van expressie wanneer verschillende concentraties van een gen vergeleken worden.



Figuur N.7: Evaluatie van geschatte, absolute expressieniveaus. Geschatte mRNA-concentraties (*copy number per cell*) voor alle 13 controles zijn uitgezet tegen de echte, gekende concentraties. De zwarte lijn is de bissectrice. Met uitzondering van de laagste concentraties komen de geschatte waardes goed overeen met de daadwerkelijke mRNA-concentraties in de hybridizatieoplossing

In een laatste deel van de resultaten illustreren we hoe onze methode kan toegepast worden op zowel ruwe intensiteiten als achtergrond-gecorrigeerde intensiteiten (zelfs als deze negatieve waarden vertonen). Welke van de twee aan te raden is hangt grotendeels van het experiment zelf af: over het algemeen observeerden we dat achtergrond-gecorrigeerde metingen een groter lineaire bereik hebben, maar dat dit ten koste gaat van grotere meetfouten voor de lagere concentraties.

Discussie

Hoewel het gebruikte calibratiemodel een vereenvoudiging van de fysische realiteit inhoudt, waar storingsfactoren behandeld worden in een globale, niet-genspecifieke manier, tonen de resultaten aan dat ze in staat is microroosterdata op een adequate manier te normaliseren. Een belangrijk verschil met de meeste bestaande methodes is dat onze methode niet steunt op aannames die betrekking hebben op de distributie van genexpressieniveaus van verschillende biologische condities. Als gevolg is de beschreven procedure uitermate geschikt om experimenten te normaliseren waarvoor de GNA niet geldig is. De procedure biedt een nieuwe aanpak voor de normalisatie van microroosterdata, die het beste combineert van ANOVA-modellen, aangezien er ook absolute expressiewaarden geschat worden, en methodes die een data linearisatie uitvoeren (e.g. LOWESS).

Hoofdstuk 6: Conclusies en vooruitzichten

Het onderzoek voorgesteld in deze doctoraatsthesis handelde volledig over de normalisatie van data afkomstig van microroosterexperimenten. De strategieën die gevolgd werden, verschillen conceptueel van de standaardtechnieken. De ingeburgerde, op ratio's gebaseerde methodes zijn sterk gebonden aan assumpties aangaande de distributie van genexpressiewaardes. De meeste van deze normalisatiemethodes vertonen weinig interesse in de onderliggende oorzaken van de systematische en willekeurige variaties van de gemeten intensiteiten. Het uitgangspunt van dit onderzoek was om zoveel mogelijk de fysische en biologische realiteit van het proces te erkennen en het normalisatieprobleem aan te pakken vertrekkende vanaf absolute intensiteiten. In plaats van beperkt te zijn tot de relatieve aard van intensiteitsratio's, hebben we getracht een absolute maat van expressie te bekomen door de gemeten intensiteiten te modelleren in functie van de systematische bronnen van variatie op een experimenteel betekenisvolle manier.

Onderzoek

Initieel onderzoek (beschreven in hoofdstuk 3) bestond uit de evaluatie van procedures voor microroosternormalisatie steunend op ANOVA-modellen en een vergelijkende studie met op ratio's gebaseerde technieken. Verder onderzoek was gericht op de ontwikkeling van generische (toepasbaar op elk experimenteel design) ANOVA-modellen voor normalisatie van microroosterdata (hoofdstuk 4). In hoofdstuk 5 tenslotte, wordt beschreven hoe externe controles inzicht verschaften in vele van de problemen die opgemerkt werden in het voorgaand onderzoek, en werd een geheel nieuwe methode ontwikkeld gebaseerd op een fysisch gemotiveerd calibratiemodel.

Voor toekomstig onderzoek is het in eerste instantie van belang dat deze ontwikkelde methode toegankelijk wordt voor een groot publiek. Concreet zal een implementatie van de methode vrij beschikbaar worden gemaakt in de vorm van een 'BioConductor Package' [76] (http://www.bioconductor.org) . Deze implementatie zal gepaard gaan met een verdere uitdieping van het fysisch model waarop onze normalisatiemethode steunt, zoals een uitbreiding met parameters en foutenverdelingen die meer lokale storingsfactoren in rekening brengen, om zo de variantie op de geschatte mRNA-concentraties verder te verkleinen. Daarnaast dient gewerkt te worden aan een statistische beschrijving van de complexe foutenverdeling op de geschatte mRNA-concentraties om verdere statistische inferenties te vergemakkelijken. Door de universele basisprincipes waarop de normalisatieprocedure steunt, kan deze makkelijk aangepast worden zodat ze ook voor andere moleculair biologische high-throughput technieken kan gebruikt worden. Zo kan de methode compatibel gemaakt worden met de, op microroosters gebaseerde ChIP-chip technologie, maar ook andere technieken die voor een deel steunen op dezelfde principes als microroosters (e.g. differentiële labeling en relatieve expressie in 2D-DIGE) zouden baat kunnen hebben van gelijkaardige benadering.

Vooruitblik

Tijdens het doctoraatsonderzoek werd getracht een normalisatiemethode voor microroosterdata te ontwikkelen die verder gaat dan het analyseren van intensiteitsratio's. Er werd uitgegaan van de veronderstelling dat elke gemeten intensiteit een representatie is van de aanrijking van een specifiek mRNA-transcript, onderhevig aan een serie van experimentele factoren die al dan niet mathematisch kunnen gemodelleerd en in rekening gebracht worden. We zijn van mening dat de analyse van microroosterdata voordeel zou hebben van een meer methodologische aanpak, in tegenstelling tot aanvaarde technieken die over het algemeen weinig aandacht schenken aan de experimentele karakteristieken van een microroosterexperiment. In dit deel van het hoofdstuk wordt deze visie verder toegelicht alsook de

mogelijke implicaties ervan, zoals de absolute vereiste van externe controles. Daarnaast wordt verder ingegaan op de vraag wat differentiële expressie eigenlijk behelst en wat de invloed van de fysiologische toestand van de cellen, representatief voor de geteste biologische condities, hierop is (i.e. verschillen in hoeveelheid totaal RNA per cel en verschillen in hoeveelheid mRNA per totaal RNA).

Publications

1. Marchal K, Engelen K, De Brabanter J, Aert S, De Moor B, Ayoubi T *et al.*: Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays. *Journal of Biological Systems* 2002, 10: 409-430.
2. Engelen K, Coessens B, Marchal K, De Moor B: MARAN: normalizing micro-array data. *Bioinformatics* 2003, 19: 893-894.
3. Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, Engelen K *et al.*: INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Research* 2003, 31: 3468-3470.
4. Marchal K, De Smet F, Engelen K, De Moor B: *Computational biology and toxicogenomics.* In Predictive toxicology. Edited by Helma C. M. Dekker; 2004.
5. De Smet F, Moreau Y, Engelen K, Timmerman D, Vergote I, De Moor B: Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer* 2004, 91: 1160-1165.
6. Roosen J, Engelen K, Marchal K, Mathys J, Griffioen G, Cameroni E *et al.*: PKA and Sch9 control a molecular switch important for the proper adaptation to nutrient availability. *Molecular Microbiology* 2005, 55: 862-880.
7. De Bie T, Monsieurs P, Engelen K, De Moor B, Cristianini N, Marchal K: Discovering transcriptional modules from motif, chip-chip and microarray data. *Pacific Symposium on Biocomputing* 2005, 483-494.
8. De Smet F, Pochet N, Engelen K, Van Gorp T, Van Hummelen P, Marchal K *et al.*: Predicting the clinical behavior of ovarian cancer from

gene expression profiles. *International Journal of Gynecological Cancer* 2006, Accepted for publication.

9. De Keersmaecker SC, Marchal K, Verhoeven TL, Engelen K, Vanderleyden J, Detweiler CS: Microarray analysis and motif detection reveal new targets of the Salmonella enterica serovar Typhimurium HilA regulatory protein, including hilA itself. *Journal of Bacteriology* 2005, 187: 4381-4391.

10. Verlinden L, Eelen G, Beullens I, Van Camp M, Van Hummelen P, Engelen K *et al.*: Characterization of the condensin component Cnap1 and the protein kinase Melk as novel E2F-target genes down-regulated by 1,25-dihydroxyvitamin D3. *Journal of Biological Chemistry* 2005, 280: 37319-37330.

11. Denolet E, De Gendt K, Allemeersch J, Engelen K, Marchal K, Van Hummelen P *et al.*: The effect of a Sertoli cell-selective knockout of the androgen receptor on testicular gene expression in prepubertal mice. *Molecular Endocrinology* 2005, Accepted to appear in February 2006.

12. Engelen K, Naudts B, De Moor B, Marchal K: A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics* 2006, Accepted for publication.

Contents

Chapter 1

Introduction

1.1 A high-throughput revolution

> *"Thus, the strength of genomic studies lies in the global comparisons between biological systems rather than detailed examination of single genes or proteins. Genomic information is often misused when applied exclusively to individual genes. If one is interested only in one particular gene, there are many more conclusive experiments that should be consulted before using the results from genomics datasets. Therefore, genomic data should not be used in lieu of traditional biochemistry, but as an initial guideline to identify areas for deeper investigation and to see how those results fit in with the rest of the genome."*
>
> Greenbaum *et al*., 2001 [83]

Molecular biology has traditionally been directed towards understanding the role of a single, or a limited number of genes or proteins in a molecular biological process. Over the past decades the advent of novel *high-throughput* techniques has dramatically changed the scope of biological research and has given rise to large scale experimental methods for genome sequencing, expression analysis, and the identification of protein-protein interactions or protein-DNA interactions. Genetic and molecular biological research has shifted its focus from targeting single genes to analyzing whole-cell populations of genes and metabolites simultaneously.

These holistic approaches offer the advantage of a better understanding of fundamental molecular biological processes, as one can study the function or expression of a gene in a global cellular context. A genetic entity never acts on its own, but is always embedded in a larger network and should be treated accordingly (i.e. *systems biology*). On the other hand, high-throughput

approaches pose several novel challenges to molecular biology, because the analysis of such large scale data turned out to be far from trivial. *Bioinformatics* is a young and rapidly growing interdisciplinary research area, which may be defined as the scientific field that deals with the computational management and analysis of all kinds of molecular biological information, whether it may be about genes and their products, whole organisms or even ecological systems. There is an inseparable relationship between the experimental and the computational aspects. On the one hand, data resulting from high-throughput experimentation require intensive computational interpretation and evaluation. On the other hand, computational methods use empirical data to build a knowledge base for predictions. Furthermore, they sometimes produce questionable predictions that should be reviewed and confirmed through experiments.

A piece of history

The start of the high-throughput revolution may be dated as far back as the late 1970s, with the emergence of a branch of biology now called *genomics*. At the Laboratory of Molecular Biology in Cambridge, Sanger and his colleagues developed a revolutionary method [170] to sequence strains of DNA and managed to unravel the genomes of bacteriophage ϕX174 [171], the human mitochondrion [6], and bacteriophage λ [172]. Among other discoveries, the complete ϕX174 sequence revealed the existence of overlapping genes, and the mitochondrial sequence showed that it used alternative codons. Sanger introduced the notion that the sequence of the entire genome of a genetically defined entity formed a good start to understanding its biology. This pioneering work inspired much larger projects, culminating in recent years in the sequencing of the human genome [119,212]. The huge quantity of high quality sequence information in the public databases, a measure of Sanger's legacy, presents researchers worldwide with the challenge of progressing from sequence to function for genomes and organisms of high complexity. Computer based methods of analysis –also pioneered in Sanger's group [188]- go a long way to extracting biologically relevant information from the sequences, but computers and sequences can only go so far. In order to measure levels of gene expression, experimental methods were needed that could be applied on a scale commensurate with the large size of the complex genomes.

Global, but crude, surveys of gene expression had in fact been undertaken before, in the mid 1970s. Polysomal RNA isolated from lines of cultured mammalian cells was transcribed in vitro into radiolabeled single-stranded cDNA, which was then hybridized with an excess of its unlabelled mRNA template. Hybridization kinetics indicated that the mRNA comprised three kinetic classes, differing in sequence complexity and abundance in the mRNA population [24,77,78]. Work carried out during the next 25 years

amply confirmed these general conclusions. Over the course of a quarter century, a number of mRNAs belonging to each abundance class were catalogued, mapped, and quantified by a combination of procedures, especially developed to measure the abundance of target mRNA. The most popular of these procedures –northern blots [5] and ribonuclease protection [18]- remain in common use today. Both methods suffer from the same limitations: they utilize labelled, specific DNA probes to detect one (or at best a few) specific mRNA in the preparation; quantification of mRNA is achieved indirectly by comparing the signal obtained from the target mRNA with that of a housekeeping gene, or of a known concentration of an artificial target RNA. Similar deficiencies apply to other techniques, such as quantitative RT-PCR. Despite their limitations, methods to analyze the abundance of one or a few species of mRNAs have provided keen insight into the biology of a wide range of 'single gene' function puzzles. It was always clear however, that integrated comprehensive maps of cellular transcription could not be built bottom up from studies of individual mRNAs, and that the understanding and classifying of complex problems would require techniques to monitor global changes in gene expression.

Microarrays

The advent of DNA microarrays provided researchers with the means to monitor such global changes in gene expression. In fact, microarrays became the central driving force behind the further development and worldwide acceptance of high-throughput techniques. Many of the principles of modern microarrays were established in the late 1980s and early 1990s, when cloned cDNAs, arrayed on membrane filters, were hybridized to complex targets and used to quantify differences in expression of mRNAs over a wide dynamic range [37,84,122,129,191]. A major breakthrough came in the mid 1990s, when Pat Brown, Ron Davis, and their colleagues published papers describing the use of a two colour, internally comparative technique to probe cDNAs arrayed robotically at high density on solid substrates [49,175,176]. These studies led to the development of DNA microarrays to screen the relative abundance of thousands of mRNAs simultaneously.

The objective of most microarray projects is to identify genes expressed at different abundances in complex samples of RNA extracted from different cells or from the cells growing under different conditions. Differential gene expression analysis has uncovered networks of genes within common pathways of regulation [136,236], and has revealed differences between cancers that cannot be distinguished by conventional ways [3,186]. None of these important results could have been achieved as simply or speedily by any other means of analysis. The technology has also produced many significant results in quite different areas of application. Analysis of transcripts has been used to discover exons and genes for the annotation of

the draft sequence of the human genome [181], and analysis of genomic DNA detects amplifications and deletions found in tumours [75,97].

The microarray platform is a relatively complex technology and has drawn together a vigorous community of interest from several disciplines: engineers, materials scientists, mathematicians, and chemists, in addition to molecular biologists, geneticists, and computer scientists. The research presented in this PhD thesis is based entirely in the field of microarray data analysis. More precisely, it deals with the normalization of the intensity measurements that are obtained from scanned images of spotted microarrays.

1.2 Motivation

Normalization of spotted microarray measurements, the first step in a microarray analysis trajectory, aims at removing consistent and systematic sources of variations to allow mutual comparison of measurements acquired from different slides and experimental settings. Data normalization largely influences the results of all subsequent analyses and the biological interpretation of these results, and is therefore a crucial phase in the analysis of microarray data. It could be argued that the extraction of intensity values from scanned images in itself is a process subject to various experimental and computational factors, and that its proper execution should not be ignored within the framework of data preprocessing. However, such procedures are often highly dependent on the type of equipment and are generally implemented in software that is provided by the manufacturer of the instruments. More importantly, the availability of scanned images to researchers is rather low. Laboratories that outsource their microarray experiments do not always have access to the scanned images of the analysis they commissioned, and microarray data that are submitted to public databases are rarely accompanied with the relevant image files. Indeed, image files are not required in the results format of the widely accepted standard for reporting microarray data (Minimum Information About a Microarray Experiment or ***MIAME*** [29]), mainly due to the large size of such files and the implications for long term storage. The basic measurement unit that is used throughout this thesis are therefore the intensities that are extracted from scanned microarray images, and little attention is given to the actual extraction process.

Normalization of spotted microarray data

Over the past years, the field of microarray analysis finally seems to have adapted a few generally applied methodologies for data normalization (for overviews, see for instance Leung and Cavalieri, 2003 [123], Quackenbush, 2002 [157] and Bilban *et al.*, 2002 [22]). Although some approaches

inherently work with absolute intensities (e.g. ANOVA [113,221]), in general, preprocessing of spotted microarrays largely revolves around the calculation of the log-ratios of the measured intensities. The reason can be found in the inherent differential nature of spotted microarrays: two different samples, labelled with different fluorescent dyes (Cy3 and Cy5), are hybridized to the same microarray, and their intensities are compared. Given these experimental features, taking ratios is a genuinely logical approach to analyzing the data. The use of intensity ratios however, is not without drawbacks. From a theoretical point of view, ratios increase the measurement noise by multiplying the intensity errors. Moreover, ratios disregard possibly useful information regarding the absolute level of gene expression (e.g. a certain ratio might indicate a significant change in expression for high intensity values, while the same ratio might be meaningless for lower intensities due to experimental error characteristics). Ratios also have severe practical implications, as for complex experimental designs, their use complicates comparing multiple biological conditions, especially when they are not measured with the same reference condition.

A common ratio normalization step consists of the linearization of the Cy3 versus Cy5 intensity ratios (e.g. LOWESS [226]), sometimes followed by, or inherently combined with, techniques for variance stabilization [62,98]. These methods assume that the distribution of gene expression shows little overall change and is balanced between the biological samples tested (referred to as the 'Global Normalization Assumption'). If this assumption is violated, for instance when comparing drastically different biological conditions or when working with dedicated arrays, using such a normalization may yield erratic results that propagate throughout every step of the subsequent analysis and the biological interpretation of the results.

Beyond differential expression: a different approach

Normalization of intensity ratios is heavily bound to assumptions concerning the distribution of gene expression; they are guided by how expression levels are presumed to change across different biological conditions. Put simply, microarray data are often normalized by transforming the calculated ratios into a measure of differential expression to which the concealed biological reality is expected to conform. Ratio normalization techniques generally show little interest in the underlying causes of the observed systematic and random variation in microarray data. The normalization methods we pursue in this thesis differ in spirit from the ones mentioned in the previous section. The basic premise is to acknowledge the physical and biological reality of the process and address the normalization problem starting from units of absolute intensities. These measured intensities are to be modelled as a function of systematic sources of variation in a physically and experimentally meaningful way, and should allow for the calculation of an absolute value of expression instead of being limited to the relative nature of

intensity ratios. When done properly, such an approach could circumvent most of the problems that seem to be inherent to the calculation of intensity ratios. Moreover, estimates of absolute expression can greatly simplify the analysis of large, complex designs comparing multiple biological conditions and could aid inter-platform and inter-laboratory comparisons of expression analysis [13,101,120].

1.3 Thesis outline and achievements

An overview of the organization of the thesis can be found in Figure 1.1. The relationship between the different chapters is of a chronological and causal nature: each chapter is a logical continuation of the research described in its predecessor. Apart from chapter 2, which serves as an introduction to microarray technology, each chapter is associated with our own contributions to a research topic. A brief chapter-by-chapter overview of the conducted research is given below. A list of publications that resulted from this work can be found at the beginning of this text (see p. xli).

Chapter 2: Spotted microarrays

Since the prime focus of this PhD is with the normalization of spotted microarray data, this chapter will serve as a general introduction to microarray experiments and data analysis. The first part gives an overview of the basic technology and experimental principles, followed by a survey of the most important features of microarray data and standard data analysis techniques. A final part of this chapter will list some extended applications of spotted microarrays, other than the widespread monitoring of gene expression levels.

Chapter 3: Evaluation of ANOVA normalization

This chapter represents a first phase of this PhD research, where the use of ANOVA for microarray normalization was evaluated and compared to ratio based approaches (reported in Marchal *et al.*, 2002 [132]). The performance of any normalization procedure, especially one that estimates absolute expression levels, is hard to assess as the actual levels of mRNA abundance in any particular condition are normally unknown. In order to nevertheless appraise the presence of any markedly flawed features in ANOVA normalized data, genes thought to be differentially expressed were selected for both log-ratios and data preprocessed by ANOVA. To minimize the influence of the used selection method, several different ratio based procedures for selecting differentially expressed genes were tested and compared to a selection procedure based on ANOVA normalization.

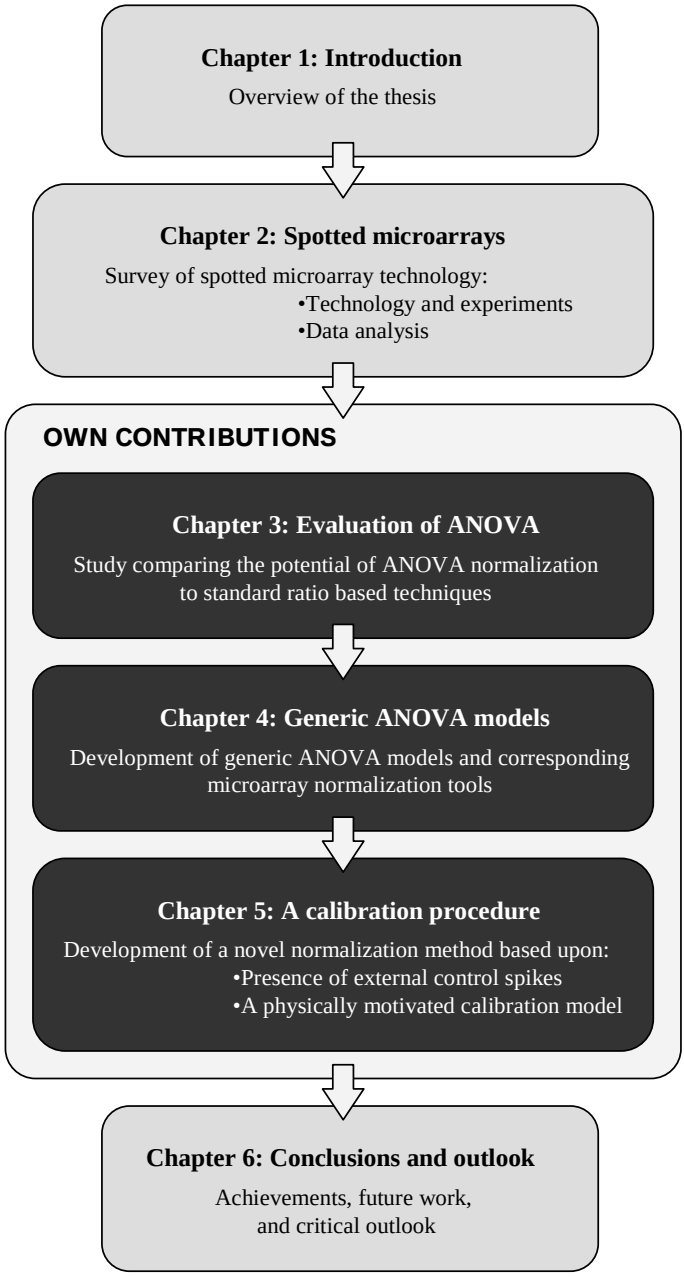


Figure 1.1: Organization of the thesis. Chapters that deal with our own research contributions are shown in black frames.

A first part of this chapter describes the principles of ANOVA based normalization of microarrays, and details some of the particular models [112,113] that can be used for normalizing colour flip designs, as well as the statistics that can be used to select differentially expressed genes based on ANOVA normalized data. A second part will provide a background to the log-ratio based methods for selecting differentially expressed genes, namely the fold test [157], the paired *t*-test [126], and a method called Significance Analysis of Microarrays or SAM [204]. The third section describes the results of performing the analysis, a discussion of which can be found in the final section.

Chapter 4: Generic ANOVA models

ANOVA models for microarray normalization can not readily be applied to any type of experimental setup of a microarray experiment. This chapter describes the issues that are encountered when attempting to fit published ANOVA models to different experimental designs, and the development of generic (applicable to any experimental setup) ANOVA models for microarray normalization. The following section is dedicated to the implementation of such a generic model in a user friendly web application, dubbed MARAN (http://www.esat.kuleuven.be/maran; Engelen *et al.*, 2003 [72]).

The final part of this chapter discusses some interesting features that were revealed during the course of this research. These results seem to indicate that a LOWESS normalization may not be able to completely alleviate intensity dependent nonlinear tendencies in the data (despite of harsh assumptions with regards to the distribution of gene expression from one biological condition to the next), and fuelled the research described in chapter 5.

Chapter 5: A calibration procedure for spotted microarrays

In this chapter we develop a normalizing method for spotted microarray data, using external control spikes to fit a calibration model (Engelen *et al.*, 2006 [73]). This model incorporates parameters and error distributions representing both the hybridization of labelled target to complementary probes, and the subsequent measurement of fluorescence intensities. External control spikes serve to estimate the model parameters. The obtained parameters values are then employed to estimate absolute levels of expression for the remaining genes. For each combination of a gene and a tested biological condition, a single absolute target expression level is estimated, taken the specificities of the design.

We discuss results that were obtained from applying our method to a publicly available data set, and show that the procedure is capable of adequately removing the typical non-linearities of microarray data, without

making any assumptions on the distribution of differences in gene expression from one biological sample to the next. Next, we compare our method to results obtained from normalizing the data with a standard LOWESS procedure prior to fitting an ANOVA model. Since our model links target concentration to measured intensity, we further demonstrate how absolute expression values of transcripts in the hybridization solution can be estimated. Finally, we illustrate the effect of local background correction and the models capacity to deal with negative (background corrected) intensity values.

Chapter 6: Conclusions and outlook

The results and observations that culminated from this work are summarized in this chapter, together with a short description of some concrete problems that will be studied in the future and an outlook on microarray normalization and spotted microarrays in general.

1.4 Cooperations

During the entire term of the PhD, (in)formal cooperations were made with several molecular biological and biomedical research groups, such as the Centre for Microbial and Plant Genetics (CMPG), the Molecular Physiology of Plants and Micro-organisms Section, the Laboratory for Molecular Cell Biology, the Microarray Facility (MAF; Flanders Interuniversity Institute for Biotechnology), the Experimental Medicine and Endocrinology Section (LEGENDO), the Gynaecology Section, the Intelligent Systems Lab (ISLab; University of Antwerp), and the Laboratory for Malting and Brewing Sciences. These research groups provided us with the data sets of their microarray experiments, a necessary means for the evaluation of our algorithms and implementations. In exchange, the obtained data were extensively analyzed. Some of these analyses, usually together with extra experimental validation in a *wet lab* environment, have led to various publications [41,43,44,46-48,131,168,213], but are not discussed in this dissertation.

Chapter 2

Spotted microarrays

High-throughput experiments allow measuring the expression levels of mRNA (genomics), proteins (proteomics) and metabolite compounds (metabolomics) for thousands of entities simultaneously, and can provide a wealth of data that can be used to develop a global insight into the cellular behaviour. The most powerful experimental designs consist of surveying a biological system in a wide array of responses, phenotypes or conditions. The combination of these experimental data and the right computational tools can lead to powerful new findings with applications in drug discovery, disease management, metabolic engineering, etc. One of the main contributors to the surge of high-throughput applications in biological and biomedical research and industries is the development of DNA microarray technologies.

DNA microarrays are a technology that permit the simultaneous assessment of mRNA expression levels of thousands of genes in a single hybridization assay. An array consists of a reproducible pattern of different DNAs (primarily PCR products or oligonucleotides) attached to a solid support. Each spot on an array represents a distinct coding sequence of the genome of interest. There are two main microarray platforms that can be distinguished from each other in the way that DNA is attached to the support, and the specifics of how the hybridization reaction is performed: spotted microarrays, and GeneChip or Affymetrix arrays.

- ***Spotted microarrays*** (sometimes still referred to as *cDNA microarrays* for historical reasons) are small glass slides on which pre-synthesized single stranded DNA or double-stranded DNA is spotted. These DNA fragments can differ in length depending on the platform used (cDNA microarrays versus spotted oligomer arrays). Usually the probes contain several hundred of base pairs and are derived from Expressed Sequence Tags (ESTs) or from known coding sequences from the organism under study. Usually

each spot represents one single gene or Open Reading Frame (ORF). A high-density spotted array can contain up to 25000 different spots.

- ***GeneChip*** oligonucleotide arrays (Affymetrix Inc., Santa Clara) are high-density arrays of oligonucleotides synthesized *in situ* using light-directed chemistry. Each gene is represented by 15-20 different oligonucleotides (25-mers), which serve as unique sequence specific detectors. In addition, mismatch control oligonucleotides (identical to the perfect match probes except for a single base-pair mismatch) are added to allow the estimation of cross-hybridization. An Affymetrix array can represent over 40000 genes.

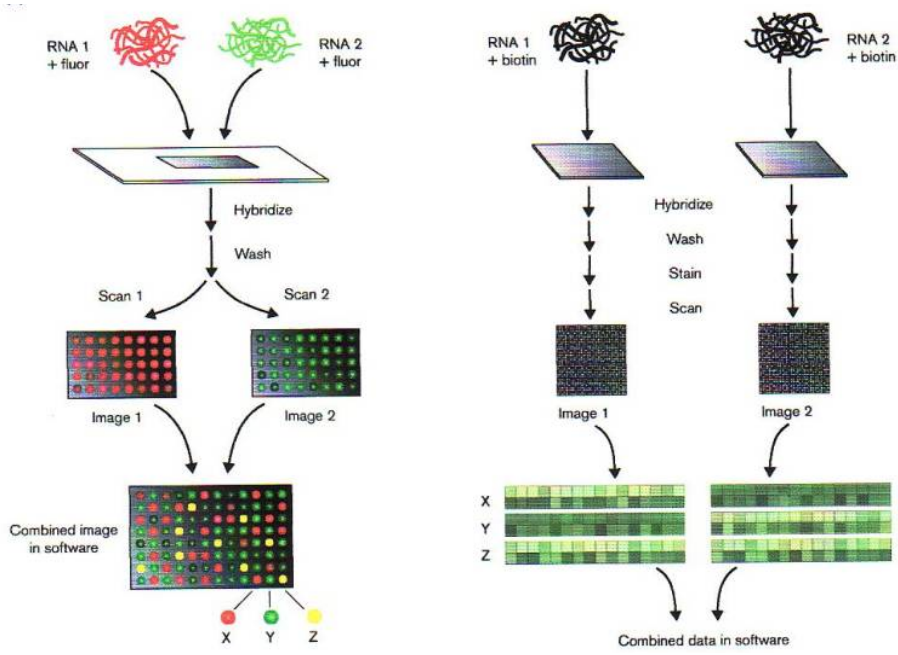


Figure 2.1: Spotted microarrays versus Affymetrix GeneChips. The main conceptual difference between spotted microarrays (left) and Affymetrix GeneChips, is that spotted microarrays allow two-colour hybridization, which permits simultaneous, relative analysis of two samples on the same array. Affymetrix arrays on the other hand, can only measure a single biological condition can array. Taken from Harrington *et al.*, 2000 [86].

As illustrated in Figure 2.1, the main conceptual difference between spotted microarrays and Affymetrix GeneChips, lies in whether or not multiple samples are hybridized simultaneous to a single microarray. Since the prime focus of this PhD is with the normalization of spotted microarray data, this chapter will give an overview of the basic technology and experimental principles (section 2.1), followed by a survey of the general data analysis techniques (section 2.2). A final part of this chapter (section 2.3) will list some extended applications of spotted microarrays, other than the widespread monitoring of gene expression levels.

2.1 Technology and experimental procedures

This section describes the technology and procedures that are involved in a spotted microarray experiment (Figure 2.2), from production of the microarray slides (section 2.1.1), to the preparation of hybridization samples, the hybridization reaction, and fluorescence scanning of the hybridized samples to their complementary DNA on the microarray (section 2.1.2). For sake of clarity, we abide to the convention [152] of referring to the material spotted on the microarray as *probes*, and the material to be hybridized on the microarray as *targets* (contrary to the accepted terminology for the single gene equivalent *Northern blots* or *quantitive PCR* techniques).

2.1.1 Slide Production

2.1.1.1 Probe generation

The first step in the production of spotted microarrays is the generation of arraying material, which serves as the probe feedstock for printing. These days, probes for microarrays are constructed using either cDNA fragments or synthetic oligonucleotides (oligomers).

During the 1990s the rate of gene discovery has been greatly accelerated through the use of large-scale sequencing of cDNA libraries to generate expressed sequence tags (EST). Craig Venter was one of the initial promoters of such endeavours [212] and projects to identify and catalogue ESTs in a wide range of species are still ongoing in both commercial and academic laboratories. EST sequences from large-scale sequencing projects are deposited in dbEST [26], a division of GenBank [17], where an automated process called UniGene compares ESTs and assembles overlapping sequences into clusters. *Clone sets*, compromising a single representative of each cluster, are a resource for microarray probes and can be obtained from authorized distributors (http://image.llnl.gov/image/html/idistributors.shtml). Most researchers who

array cDNA fragments work with such off-the-shelf clone sets, possibly supplemented with individual ESTs that are appropriate to their particular needs. DNA for arraying is typically prepared from clone sets by high-throughput polymerase chain reaction (PCR), rather then by purification of recombinant constructs (e.g. plasmids). Because clone sets usually employ a restricted range of cloning vectors, it is often possible to use universal primers for the amplification of cDNA inserts.

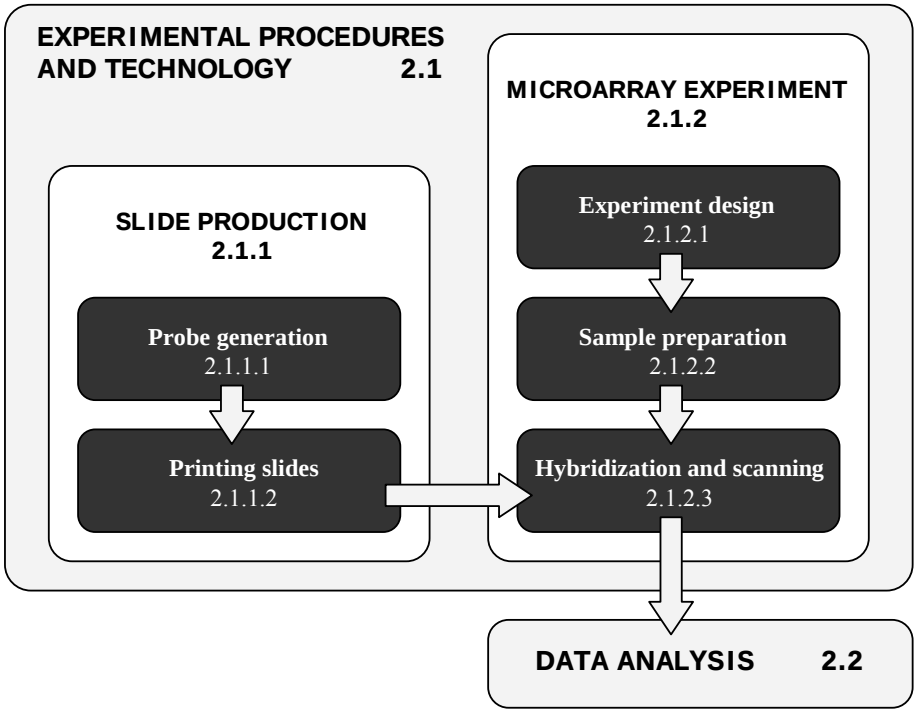


Figure 2.2: Microarray technology and experimental procedures. Overview of the technology and experimental procedures that are involved in a spotted microarray survey, ranging from the production of the slide, to the actual performance of the microarray experiment.

The disadvantages of using cDNA clone sets for arraying are that the physical handling of large numbers of clones, including replication and amplification, is labour-intensive, and the complexity of the processes involved creates opportunities for errors. Spotted arrays generated with sets of long synthetic oligonucleotides (60–70 mers) represent an attractive alternative to the arraying of cDNA-derived PCR products. The use of commercially available oligonucleotide sets obviates much of the work in the development of array-ready material and takes advantage of the growth in genome sequence information. Using EST databases and open reading frame (ORF) predicting programs, oligomer sequences are designed to abide by several constraints, such as a narrow range of melting temperature and the prevention of cross-hybridization [27,99,100]. Initial experiments indicated a high level of concordance between results obtained with PCR amplified cDNA fragments and oligomers [99], findings that were confirmed in recent publications [88].

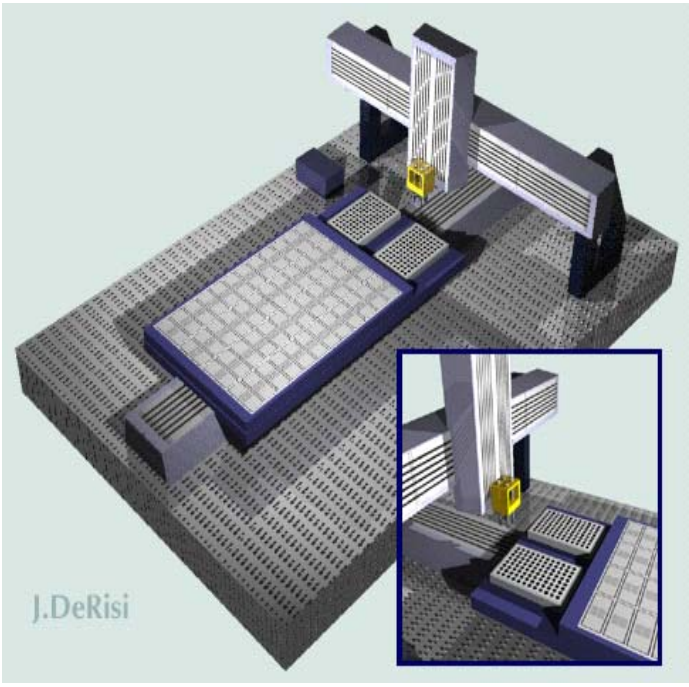


Figure 2.3: **A conceptual 3D representation of a contact arrayer.** The pinhead (yellow) that holds an arrangement of printing pins is shown rigged to a computer controlled robotic arm. Right in front of the pinhead are two 96-well microtiter plates that contain the probe spotting solutions. Further away from the pinhead are four series of glass slides layed out, ready to be printed. Taken from the DeRisi Lab website (http://derisilab.ucsf.edu/).

2.1.1.2 Printing slides

The first glass slide microarrays were produced at Stanford University [175] by an XYZ axis gantry robot that used banks of printing pins to ferry small volumes of DNA solutions from 96-well plates to the prepared surfaces of a series of glass slides (Figure 2.3). This procedure of *contact printing* [121,175] is still one of the workhorse techniques for the in-house production of microarrays, although *non-contact* (ink jet) [91,178] printing methods are increasing their market share. Commercial arrayers, both contact and noncontact, are available from several companies and building ones own arrayer remains an attractive and affordable option for the technically ample investigator (http://cmgm.stanford.edu/pbrown/mguide).

A critical factor that influences the quality of the microarrays produced with contact printing, are the types of printing pins used in the spotting process. Key features include shape, reproducibility, durability, and surface roughness of the printing pins. Proper cleaning of the pins at the end of each spotting cycle, so that the same pins may be used with different clones without significant cross-contamination, will also have a considerable effect on spot quality. Two of the most used pin types are [27]:

- *Solid pins*: made from solid steel, these pins are robust, highly uniform, easy to manufacture, and easily cleaned between cycles. A major disadvantage is that simple solid pins are capable of only one round of printing per visit to the source plate. The constant movement of the printing head between source plates and slides generates heat, so that evaporation from the source plate can be a problem.
- *Quill pins*: these pins fill up by capillary action, and were developed to allow continuous printing of a series of slides after the pins have been fully loaded at the source plate (depending on the pin this can up to 200 slides and over). Quill pins are capable of generating way smaller spots than solid pins (90-250 μm range), making them the ideal choice for generating high-density arrays. The major drawback is the sensitivity of these pins to damage and blockage. Careful cleaning between clones is essential to prevent particle and bubble retention and to prevent cross-contamination or carryover between the wells of the different source plates.

Other factors that greatly affect spot morphology, for both contact and non-contact printing, are the characteristics of the slide surface. Glass slides have been a favoured solid support for immobilization of probes because of their intrinsic material properties (low fluorescence and high transparency, good thermal properties, excellent rigidity and nonporous nature of glass), easy availability, and not in the least a surface that can readily be modified for stable DNA binding [27]. Several types of coatings can be applied that

attach firmly to the slide and tightly bind DNA spotted onto the surface. Some of the more popular substrates, along with their major features, are listed below:

- *Poly-L-lysine*: this coating was used on the first cDNA microarrays [176] and is still widely favoured due to its ease of manufacture, accessibility, and overall good performance. The binding of DNA is complex, but essentially involves charge interactions that can be converted to covalent bonding by baking or UV-irradiation.
- *Amino silane*: a popular coating alternative to poly-L-lysine. DNA is bound to the surface through electrostatic interactions [85].
- *Aldehyde*: can covalently bind to chemically modified DNA. This coating showed superior results in at least one head to head comparison of several commonly used substrates [234].
- *Activated polymers*: covalently binds DNA and holds it away from the slide surface, thereby making it more available for hybridization. The hydrophilic polymers reduce non-specific binding to the slide surface, resulting in lower background signals [159] (http://www.moterola.com/lifesciences).

Because of the physics of nanoliter spot delivery, spot quality and array sensitivity depend largely on an interplay among not only printing surface and pin type and pin performance, but also printer characteristics (pin head movement and slide and plate mounting), the composition of the DNA probe solution, and control over environmental factors such as temperature and humidity. While placing DNA probes at discrete positions on a glass support may be conceptually straightforward, the precise and reliable manufacturing of microarrays in practice is still not without challenges.

2.1.2 Performing a spotted microarray experiment

2.1.2.1 Experiment design

Performing spotted microarray experiments is a costly undertaking, even when considering a decade of growing appliance and declining price tags. Good experimental design should therefore simplify analysis and empower the interpretation of data, while balancing these aims against the constraints of microarray cost and availability, and the amount of RNA available for testing and replication. In spotted microarray experiments, the choice of design however, is not only influenced by financial considerations and the priorities of the biological questions underlying the experiment, but is heavily driven by the intrinsic relative nature of spotted microarray

expression measurements. The central design choice is whether two samples will be compared directly (on one slide) or indirectly.

A more detailed discussion and overview of different types of experimental designs is given in section 2.2.2.2.

2.1.2.2 Sample preparation

The first step in producing samples for hybridization is the isolation and purification of mRNA from tissues or cell cultures. Success in expression analysis hinges on the quality of the isolated RNA [174]. RNA contaminated with salts, polysaccharides, DNA, proteins, or lipids will label inefficiently and can mediate non-specific binding of labelled DNA to matrix surfaces, generating high backgrounds during hybridization. If the RNA is partially degraded, labelling may be biased towards sequences that lie at the 3' termini or toward sequences that are relatively resistant to attack by RNAses. This may distort the relative proportions of various targets detected by hybridization to DNA microarrays.

On average, mRNA constitutes between 1% and 5% of the total cellular RNA, and mRNA species are heterogenous in size, abundance, and sequence [124]. For eukaryotic mRNAs, a tail of polyadenylate residues (poly(A)$^+$ tail) at the 3' termini makes them distinct from the rest of the RNA population, and this unique characteristic allows their purification and separation from the other RNA species by means of chromatography on poly(dT) cellulose [8,140]. The mRNA poly(A)$^+$ tail can be conveniently exploited for other means. Reverse transcriptase, an oncoretroviral enzyme, can be used to convert sample to target cDNA by means of a short primer (usually provided by an oligo(dT) fragment) that initiates cDNA synthesis. When dealing with prokaryotic (e.g. bacterial) organisms, mRNAs are not uniformly polyadenylated, which makes matters more complicated. Methods have been proposed to selectively polyadenylate prokaryotic mRNA in the presence of total RNA [220], but working with total RNA and random primers for the reverse transcriptase reaction is also a viable option.

When only limited amounts of RNA are available (e.g. isolated from a small sample of tumour tissue), an extra amplification step is usually performed. PCR [169] is a highly efficient method for exponentially amplifying a population of single-stranded cDNA. However, the nonlinear amplification results in a target population in which sequence representation is skewed compared with the original mRNA pool. As an alternative, the amplified antisense RNA (aRNA) procedure [64,65,151,211] is a linear procedure that produces a target population more representative of the initial mRNA pool.

A final step in the preparation of target samples is the labelling process, the incorporation of fluorochromes into the target sequences. The most popular fluorochromes are without a doubt the carbocyanine dyes Cy3 and Cy5 [231]

(other examples include the Alexa dyes [145]), due to their specific fluorescence spectra referred to as the 'green' and 'red' dyes respectively (Figure 2.4). Incorporation of the dyes can be done either *directly* or *indirectly*, but both procedures are dependent on the incorporation of modified oligonucleotides during the reverse transcriptase or amplification reaction. Direct labelling [237,238] makes use of Cy3 or Cy5 labelled nucleotides, while indirect labelling [232] uses aminoallyl-nucleotides to which modified fluorescent dyes (Cy3, Cy5, or others) can be attached in a subsequent step. In both cases, the goal is to achieve a density where an average of 1 base in 8 carries a fluorescent label. Labelling at higher densities is counterproductive as quenching reduces the fluorescent yield [162].

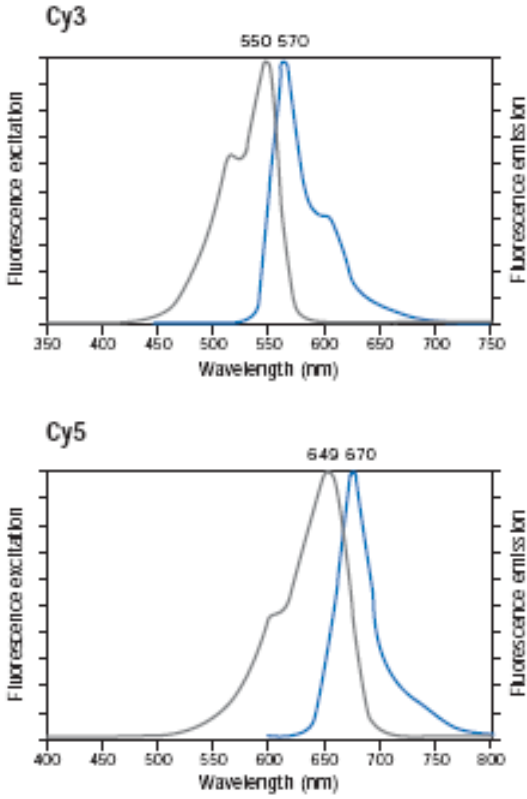


Figure 2.4: Carbocyanine dyes. Fluorescence excitation (grey curve) and fluorescence emission (blue curve) spectra for Cy3 and Cy5, the two cyanine dyes most common in spotted microarray experiments. Taken from the Amersham Biosciences website (http://www.amershambiosciences.com).

2.1.2.3 Hybridization and scanning

Hybridization is the process of incubating the labelled target DNA with the probe DNA tethered to the microarray substrate. Fluorescent target DNA hybridizes to complementary probe DNA on the slide and the emitted can be measured as an indication of the amount of immobilized target DNA. Hybridization to the probe DNA should therefore ideally be linear, sensitive (detection of low abundance transcripts) and specific (no cross-hybridization). Both the amount of probe on the slide and the concentration of target in the hybridization solution are critical factors in this process. When the amount of probe DNA is limiting, the dynamic range of the system is limiting and estimates of differential expression get compressed [92,233]. If the concentration of the target in the hybridization mixture is too low, annealing will be slow and the attenuated signal may not be detected by the fluorescence scanner. After hybridization, the array goes through a series of washes to remove all unbound labelled target DNAs. These washing steps are the most critical steps in obtaining consistently low backgrounds.

After the hybridization and washing steps, the array is scanned to obtain a measure of the amount of target bound to each probe spot. The arrays are stimulated with a laser, and the emitted fluorescence is captured by a CCD camera, non-confocal, or confocal laser scanner. Typically, the scanner produces two 16-bit images (usually TIFF files), one for each fluorescent dye, containing intensities for a large number of pixels covering the scanning area of the array. Operationally, the dynamic range of the microarray system is probably defined more by the scanners than by the concentration of target DNA. Although a number of scanner settings can usually be adjusted, such as the PMT voltage (photo multiplier tube) which is often tuned to the brightest pixels, the 16-bit nature of scanning equipment restricts their dynamical range to 2^{16} [27,128].

2.2 Data analysis

Performing the actual experiments is only a first phase in any microarray survey. Subsequent data analysis [139,156] is equally important and comprehensive. This chapter will discuss a typical data analysis flow as illustrated in Figure 2.5, divided into image analysis (section 2.2.1), preprocessing and normalization (section 2.2.2) and high-level data exploration (section 2.2.3).

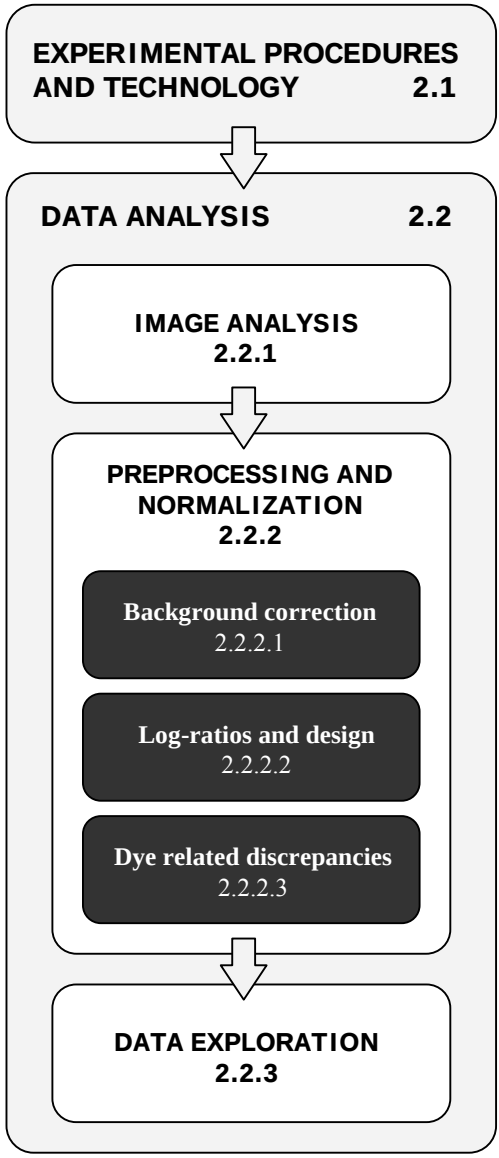


Figure 2.5: A typical data analysis flow for spotted microarrays. Starting from image analysis, followed by data preprocessing and normalization and ending with exploration and mining of the data to obtain biologically relevant results.

2.2.1 Image analysis

The analysis of scanned microarray images converts the image into spot associated numerical values that serve as a measure of target abundance. Several commercial or non-commercial packages are available that are tailored specifically to this task. The image analysis process can be divided into three major tasks: gridding, segmentation and intensity extraction.

Gridding (or ***addressing***) is the process of assigning coordinates to each of the spotted probes. The basic layout of a microarray is known as it is determined by the spot deposition by the arrayer. Gridding is meant to alleviate deviations from the exemplary spot positions, i.e. the translation of individual spots or the displacement of entire grids of spots caused by slight variations in print tip positions. Other parameters that may need to be considered are misregistration of the Cy3 and Cy5 channels, overall position and rotation of the array in the scanned image, and deviation from symmetry due to printer or scanner artefacts. Gridding procedures are very varied and have not been well documented [27,130].

Segmentation procedures classify the pixels of the image as either foreground (the *spot mask*), i.e. belonging to a printed spot of probe DNA, or background. According to the geometry of these spot masks, each segmentation method can be categorized into one of four groups (three of which are shown in Figure 2.6). *Fixed circle* segmentation fits a circle with constant diameter to all the spots in the image. In *adaptive circle* segmentation, the diameter of the circle that defines the spot mask is estimated independently for each spot. *Adaptive shape* segmentation algorithms [1,214] are not bound to a circular delineation of the spot masks and have the advantage of being able to cope with irregular spot shapes. *Histogram* segmentation differs from the other methods in that they do not explicitly classify pixels into foreground or background. Instead, these methods estimate foreground intensities from the distribution of pixels within a designated region (*target mask*).

Intensity extraction is the final step in the image analysis and involves calculating foreground and background intensities for each spot on the array in both channels (Cy3 and Cy5). Each pixel value in a scanned image is assumed to represent the level of hybridization at a specific location on the slide, and the total amount of hybridization at a particular probe spot should be proportional to the total fluorescence at the spot. The natural measure of spot intensity is therefore the sum of pixel intensities within the spot mask.

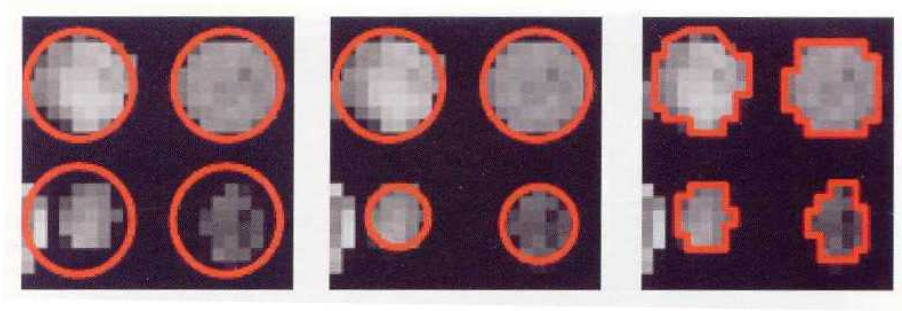


Figure 2.6: Segmentation procedures. Red circles are a representation of the segmentation effect of fixed circle (left), adaptive circle (middle), and adaptive shape (right) segmentation; grey pixels represent the actual spot. All three of these produce a spot mask where pixels inside the spot mask are considered as foreground, i.e. a measure of gene expression levels. Taken from Bowtell and Sambrook, 2002 [27].

2.2.2 Preprocessing and normalization

Normalization of the raw, extracted intensities is a necessary step before proceeding to any high-level analysis (section 2.2.3). Normalization aims to remove consistent and systematic sources of variation to ensure comparability of the measurements, both within and across slides. It largely influences the results of all subsequent analyses (such as e.g. identification of differentially expressed genes, clustering, etc.), and is therefore a crucial phase in the analysis of microarray data.

Performing microarray experiments is a complex, multi-step procedure (section 1.2), with equally vast opportunities for introducing variation that will ultimately contribute to the measured intensities [177]. Apart from human errors that can arise at various stages of the experiment (e.g. pipetting errors), critical factors include: the quality of the mRNA preparations, characteristics of the reverse transcriptase and the labelling reaction (number and density of dye incorporation), surface properties of the slide and composition of the spotting solution, deficiencies in the spotting equipment, stringency of the hybridization reaction and efficiency of the washing procedure, and equipment settings during slide scanning. As such, consistent sources of variation that manifest themselves in the data can be attributed to individual (or sets of) spots, genes, biological conditions under survey, dyes (Cy3 and Cy5), and arrays.

Since the emergence of microarrays in the mid 1990s, a plethora of (sometimes redundant) methods for normalizing spotted microarray data have been proposed. Instead of providing an exhaustive listing of different

techniques, this section outlines typical characteristics and related problems of spotted microarray data, and the widely accepted remedial measures to deal with them: background correction (section 2.2.2.1), log-ratios and experiment design (section 2.2.2.2), and dye related discrepancies (section 2.2.2.3).

2.2.2.1 Background correction

A first step in microarray data normalization is to correct the 'foreground' spot intensities for background, as the measured intensity of each spot includes a contribution of non-specific hybridization, residual Cy3 and Cy5 dyes and fluorescence emitted form other parts of the array (*overshining*) and/or the slide substrate itself. It is generally accepted that the background contribution is additive with respect to the spot intensity [34] (background correction is therefore often referred to as *background subtraction*). Unfortunately, it is impossible to measure the true background for each and every spot. The true background being a measurement of the fluorescence of a spot after the hybridization reaction, but with no complementary transcripts bound to it. As a result, several methods have been developed to quantify the intensity of background signals, all of which can merely provide an approximation of the true background.

The use of a ***constant background*** is by all means the simplest. It employs the mean or median of the whole image background (as determined in the segmentation process) as a measure of background intensity. Constant background correction is seldom used in real applications, due to its difficulty of dealing with inhomogeneous backgrounds.

Local background intensities are estimated by focusing on small regions surrounding the spot masks. Usually, the background estimate is the mean or median of pixel values within those specific regions. Figure 2.7 illustrates different local background adjustment methods. Local background correction remains one of the most popular techniques to this day, regardless of the multiple objections in the relevant literature [32,66,82,115,135,201]. Apart from the substantial variance in local background intensities, the main critique is focused on the occurrence of higher than signal backgrounds, resulting in negative corrected intensities, which are of course insensible. Negative corrected intensities have been shown to not arise because the spot is being incorrectly located during image segmentation, but rather because more fluorescent compounds are actually binding to the area surrounding the spot than to the spotted probe itself [32]. This is thought to be caused by differences in the chemistry of non-specific binding of target and/or residual fluorochromes to the DNA-free substrate and the (non-homologous) spotted DNA. Negative corrected intensities pose problems during further data analysis (e.g. the calculation of log-ratios, see section 2.2.2.2) and, as a consequence, need to be omitted or replaced by arbitrary values [45,202].

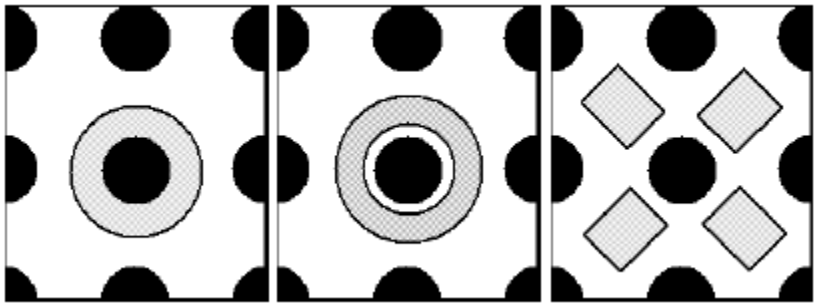


Figure 2.7: Local background correction. Schematical representation of different local background corrections. Black regions represent the spot mask. Bounded grey regions represent the regions used for background calculation as used by software packages such as ScanAlyze (left), Imagene or QuantArray (middle), and GenePix or Spot (right).

To avoid the high variability of local background estimates, while still providing a less rigid approximation than the all-slide, constant background, algorithms have been developed that essentially fit a ***background model*** to the image data. The *morphological opening*, a non-linear smoothing filter, can be considered as such a technique [184], and a multitude of others have been proposed since [32,66,82,115,117,165,224,229,230].

The comparison of different background correction methods indicates that estimates based on local neighbourhoods, and occasionally estimates based on background models, are quite noisy and tend to greatly inflate the standard deviation of the log-ratios [225]. At the other extreme, one can consider the possibility of ***no background*** adjustment at all. This option is not without drawbacks either: not performing a background correction could possibly hamper the ability to identify differentially expressed genes [225].

It has been suggested that it may be more meaningful to estimate background on the basis of a set of negative control spots [56,159,165,225]. Up to this day, a widely accepted way to correct for background has yet to emerge. This is reflected in the many applied and methodological microarray papers that are published each year, where the background correction procedures that are used, or discarded altogether, seem to be chosen depending on whichever strategy renders the most suitable results.

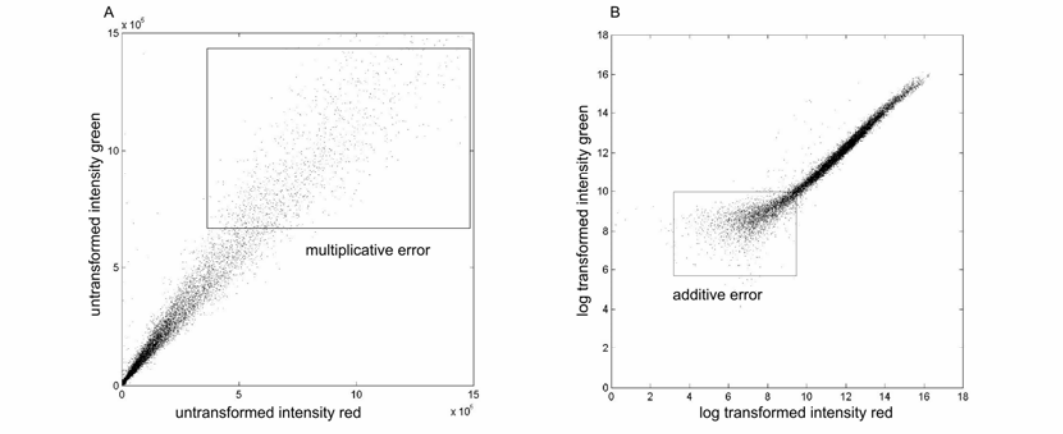


Figure 2.8: Multiplicative and additive intensity error in microarray data. The random variation, as measured by the standard deviation of the intensities, typically increases approximately linearly with the average signal strength, deteriorating the reliability of most statistical test. Removal of these multiplicative errors can be done by log-transforming the data (right hand plot), showing an increased variance at low intensity levels corresponding to the additive background error.

2.2.2.2 Log-ratios and experimental design

Spotted microarray technology is fundamentally designed towards the measurement of ***relative*** gene expression. Hybridization is performed simultaneously with two differentially labelled samples (Cy3 and Cy5) and the resulting data consists of per spot intensities of both channels. Although methods exist that work with the logarithm of the absolute intensities [104,113,221], the logarithm of the ratios of Cy5 over Cy3 intensities (***log-ratios***) for each spot are the basic 'unit' of data interpretation. These intensity ratios are thought to alleviate the large, spot related variations that occur in microarray data. The motivation to perform a log-transformation on the other hand, is twofold:

- Apart from the additive background error, microarray intensities also show a pronounced multiplicative error [34] (illustrated in Figure 2.8). The random variation, as measured by the standard deviation of the intensities, typically increases approximately linearly with the average signal strength, deteriorating the reliability of most statistical test. Removal of these multiplicative errors can be done by log-transforming the data. The increased variance at low intensity levels (Figure 2.8, panel B) is intuitively plausible, as low expression levels are generally assumed to be less reliable [71].

- Interpretation of intensity ratios is facilitated by performing a logarithmic transformation. Levels of over and under expression are brought to the same scale, i.e., values of under expression no longer range between 0 and 1 (a log2-base transformation is usually applied for convenience).

The relative nature of spotted microarray measurements has severe repercussions on the setup of the appropriate experiments (***experiment design***). As pointed out before, the choice of experimental design is not only influenced by financial considerations and the priorities of the biological questions underlying the experiment, but is heavily driven by the differential labelling inherent to spotted microarrays. The central design choice is whether two samples will be compared directly (on one slide) or indirectly (across slides). Some excellent reviews on experimental design were published by Churchill, 2001 [38] and Yang and Speed, 2002 [227].

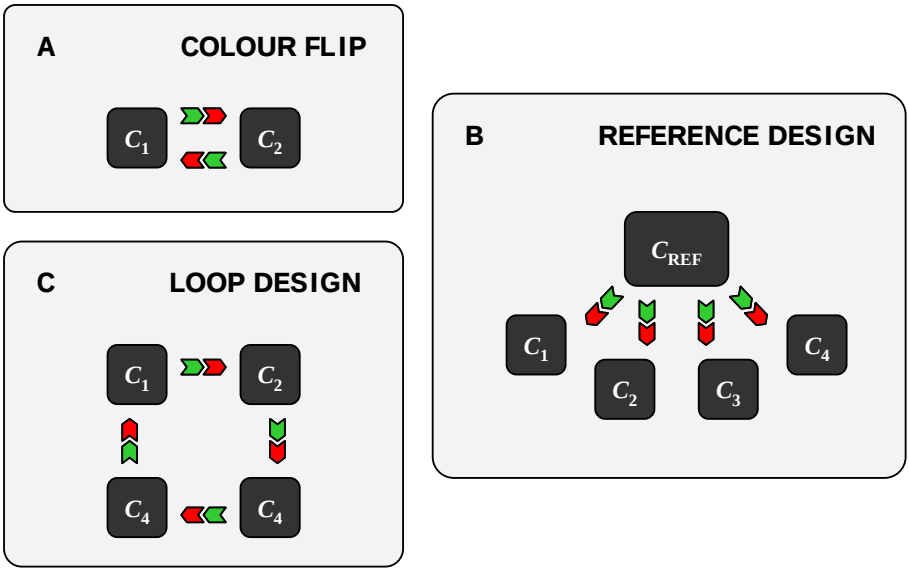


Figure 2.9: **Experimental design.** Schematical representation of some basic experimental designs: A) Colour flip design, B) Reference design (4 conditions and a reference), and C) Loop design (4 conditions). Black boxes represent the different biological conditions. Arrows represent the arrays on which indicated conditions are hybridized, either labeled in Cy5 (red part of arrow) or Cy3 (green part of array).

The most basic designs that are commonly applied are depicted in Figure 2.9 and are further discussed below. The simplest microarray experiments compare expression in two distinct conditions. A test condition (e.g. a cell line triggered with a drug compound) is compared to a reference condition (e.g. a cell line triggered with a placebo). Usually the test is labelled with Cy5 (red dye) while the reference is labelled with Cy3 (green dye). Performing replicate experiments is mandatory to infer relevant information on a statistically sound basis. However, instead of just repeating the experiments exactly as described above, a more reliable approach is to perform ***colour flip*** experiments (also called *dye swap* experiments). The same test and reference conditions are measured once more as a repeat on a second array but the dyes are swapped, i.e. on this second array, the test condition is labelled with Cy3 (green dye) while the corresponding reference condition is labelled with Cy5 (red dye). This allows better compensating for dye specific biases, to the extent that these biases are repeatable across slides. Generally, colour flipped pairs are recommended whenever possible [27,38,227].

When the multiple distinct biological conditions are compared (e.g. different mutant strains, different drug treatments, etc.), or when the conditions under study reflect the biological behaviour during the course of a dynamic process (e.g. a time course experiment), more complex designs are required. Customarily used, and still preferred by many molecular biologists, is the ***reference design***: different test conditions are each paired with the same reference condition on separate arrays. The reference condition can be artificial and does not need to be biologically significant. Its main purpose is to have a common baseline to facilitate mutual comparison between the samples. There are two main disadvantages to this approach. Firstly, half of the measurements (and consequently half of the experiment costs) are replicates of the condition in which one is not primarily interested (i.e. the reference condition). Secondly, genes that demonstrate a low expression level in the reference condition (or no expression at all), will produce unreliable ratios or even missing values. In order to retain most signals, the choice of reference is therefore not trivial. An independent sample is often chosen as reference, such as a mixture of mRNA, isolated from a wide range of biological conditions (*mRNA pools*). The use of genomic DNA for bacterial microarrays can also be considered as an independent reference.

An alternative to the reference design is the ***loop design***. A loop design can be viewed as an extended colour flip experiment. Every condition is measured twice, each time on a different array and labelled with a different dye. For an equal number of arrays, a loop design offers more balanced replicate measurements of each condition than a reference design, while the dye specific biases are conceptually compensated for. The main disadvantage of a loop design manifests itself when comparing two

conditions on opposite ends of the loop. Such a comparison requires the evaluation of ratios upon ratios, significantly increasing the error variance for each step of the loop that separates the two conditions.

These basic designs are by no means the only ones used in microarray experiments. They often serve as templates or building blocks for larger and more complex designs (e.g. a reference design extended with a colour flip for every array is not uncommon).

2.2.2.3 Dye related discrepancies

The use of log-ratios theoretically removes all systematic errors originated from spots, printing pins or array effects. As a result, normalization strategies for spotted microarrays are mostly centred on the removal of dye-related discrepancies from the log-ratios. These dye biases can cause a significant distortion of log-ratio distributions and stem from a variety of factors, namely physical properties of the dyes and efficiency of dye incorporation, but also differences in the amount of input RNA, and the scanner-specific excitation and collection process. In practice, Cy5 intensities often tend to be lower than the Cy3 intensities, and the observed imbalance is usually not constant within and across arrays.

For any procedure intended on removing these consistent dye biases, a decision must be made as to which set of genes to use for the normalization (for a review, see Kroll and Wolfl, 2002 [118]). Different approaches have been described, such as the use of *spikes* [208] (also called *control spots*, *external controls*), *housekeeping genes* (genes expected not to alter their expression level under the conditions tested), a *microarray sample pool* [226], or iterative procedures to find a set of invariant genes [173,203], but customarily, *all genes* are used to perform a correction for dye biases. This is because it is assumed reasonable that **1)** only a relatively small portion of genes will vary significantly in expression between two mRNA samples of distinct biological conditions, and **2)** there is symmetry in the level of up-regulated versus down-regulated genes. We'll further refer to these principles as the '***Global Normalization Assumption***' or **GNA**. The use of the GNA for normalization is less appropriate when the expression patterns between two biological conditions are expected to differ considerably. For instance, in bacterial arrays, where genomic DNA is often used as a reference sample, or in *dedicated arrays*, which contain only a relatively small number of genes all thought to be involved in the studied process, this assumption is no longer valid and global rescaling to a log-ratio of zero is arbitrary.

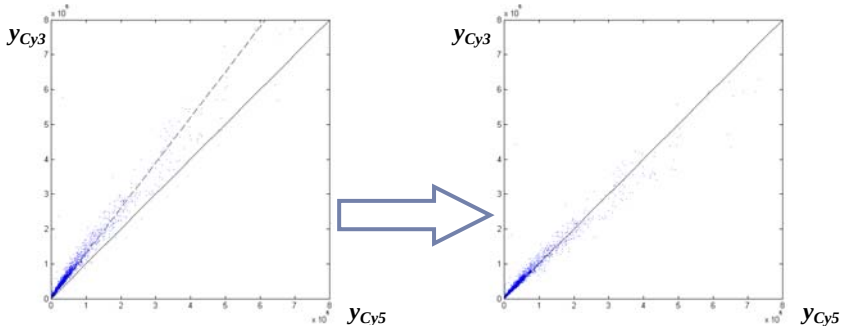


Figure 2.10: Linear normalization of data from a sngle microarray slide. Linear normalization assumes that the intensity measurements in both channels (y_{Cy5} and y_{Cy3}) are related by a constant factor for the entire slide. A common choice for this transformation factor is the mean or median of the log intensity ratios for a given gene set (represented by a dotted line in the left plot).

As dictated by the GNA, removal of dye specific intensity biases from any microarray should produces log-ratios that are evenly distributed around zero. A ***linear normalization*** (Figure 2.10) assumes that the intensity measurements in both channels (y_{Cy5} and y_{Cy3}) are related by a constant factor k for the entire slide:

$$y_{Cy5} = k \cdot y_{Cy3} \tag{2.1}$$

A common choice for the transformation factor k is the mean or median of the log intensity ratios for a given gene set. Alternatively the constant normalization factor can be determined by linear regression of the Cy5 signal versus the Cy3 signal. More complex approaches, that use an iterative method to estimate the constant normalization factor [34,59], have also been described.

Unfortunately, assuming a linear relationship between the measurements in both channels is an oversimplification. Variation between the Cy5 and Cy3 channels is seldom constant, but changes as a function of the intensity of the signal and is most pronounced at extreme intensities (either high or low). *MA-plots* that plot the log-ratios $M=\log(y_{Cy5}/y_{Cy3})$ against the average intensity $A=\log(\sqrt{y_{Cy5}.y_{Cy3}})$ are often used to visualize this phenomenon. Performing an ***intensity dependent normalization*** can be done by generating a best-fit curve through the middle of an MA-plot, and setting this as the new

zero line for the vertical axis (Figure 2.11). A corrected log-ratio M_{corr} is calculated by shifting the log-ratio M by a quantity that depends on the corresponding A value:

$$M - c(A) = M_{corr} \tag{2.2}$$

Several intensity-dependent normalizations have been proposed [74,107], but most popular, and widely applied procedure was first described by Yang *et al.* (2002) [226]. The estimate of *c(A)* is made using a LOWESS (LOcally WEighted Scatter plot Smoother; more detailed information can be found in appendix A) function [39] to perform a local scatter plot smoothing to the MA-plot. The scatter plot smoother, a type of regression analysis, performs robust locally linear fits by calculating a moving average along the A axis. Robust in this context means that the curve is not affected by a small to moderate percentage of differentially expressed genes that appear as outliers in the MA-plot. A user defined parameter f is the fraction of the data used for smoothing at each point; the larger the f value, the smoother the fit. Typically, a value of f between 30% and 40% is recommended.

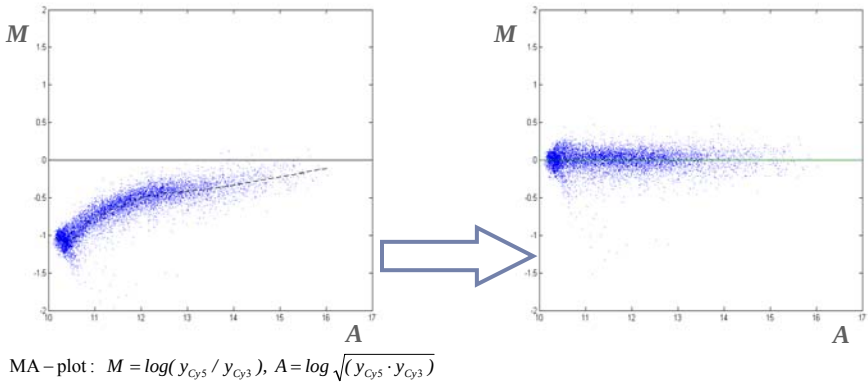


Figure 2.11: Intensity dependent normalization of data from a single microarray slide. Intensity dependent normalization techniques assume that the dye bias is non-linear in nature. *MA-plots* that plot the log-ratios M=log(y_{Cy5}/y_{Cy3}) against the average intensity A=log($\sqrt{y_{Cy5}.y_{Cy3}}$) are often used to visualize this phenomenon. Performing an intensity dependent normalization can be done by generating a best-fit curve through the middle of an MA-plot, and setting this as the new zero line for the vertical axis. A corrected log-ratio M_{corr} is calculated by shifting the log-ratio M by a quantity that depends on the corresponding A value. The curve that is shown in the left hand plot, and that serves as the basis for the non-linear rescaling, was estimated by performing a LOWESS fit [226].

2.2.3 Data exploration

After the data have been normalized, they can be explored in order to extract biologically meaningful results. The biological or biomedical questions that need to be addressed can be quite diverse, and so numerous techniques and algorithms from statistics, data mining and machine learning have found their way into the microarray data analysis field. This section lists an overview, which is by no means exhaustive, of different data exploration methods.

A microarray experiment measures the expression levels from thousands of genes in parallel. Genes that show little or no change in expression levels are typically of no biological relevance. As such, a selection of the genes show a variable expression across the condition tested is often a crucial step in the analysis of any microarray experiment. Over the years many methods have been proposed for the ***identification of significantly differential genes***, some of which are discussed (and evaluated) in detail in chapter 3. Related types of dimensionality reduction in the gene space are those that identify genes for which the expression profile is most correlated with the distinction between different conditions or sets of conditions (e.g. between different mutant strains or different classes drug compounds). These methods can either be *supervised* (distinction is known; e.g. the methods described by Park *et al*., 2001 [147] or Golub *et al*., 1999 [81]) or *unsupervised*. ***Clustering of genes*** is a prominent form of unsupervised dimensionality reduction among researchers that work with microarray data. Genes involved in a similar biological pathway or with a related function often exhibit similar expression behaviours across different biological conditions (*co-expression*). The objective of cluster analysis of gene expression profiles is to identify clusters (subgroups) of such co-expressed genes. The first generation of cluster algorithms that were used in microarray data analysis included standard techniques such as *K*-means [200], self-organizing maps [116,192], principal component analysis [23] and hierarchical clustering [71] (illustrated in Figure 2.12 and still one of the more popular among researchers). Although biologically meaningful results can be obtained with these algorithms, they often lack the fine-tuning that is necessary for biological problems [194]. A panoply of cluster algorithms have been designed since, better tailored to the specifics of microarray data, and the needs of molecular biology research [15,79,87,94,95,127,179,228] (for an overview, see Moreau *et al.*, 2002 [138]).

Other applications main point of interest lays in the space of the tested conditions (e.g. in pharmaceutical or clinical settings: drug discovery, toxicogenomics, disease management, etc.). Examples are methods for ***class discovery***, where the tested biological conditions are subdivided into classes based on the characteristic expression fingerprints of cells exposed to these

conditions (unsupervised), and ***class prediction***, where the class membership of new conditions (e.g. drug compounds) is predicted based on a classifier model which was trained from a data training set (supervised). For a more detailed overview, see Marchal *et al.*, 2004 [131].

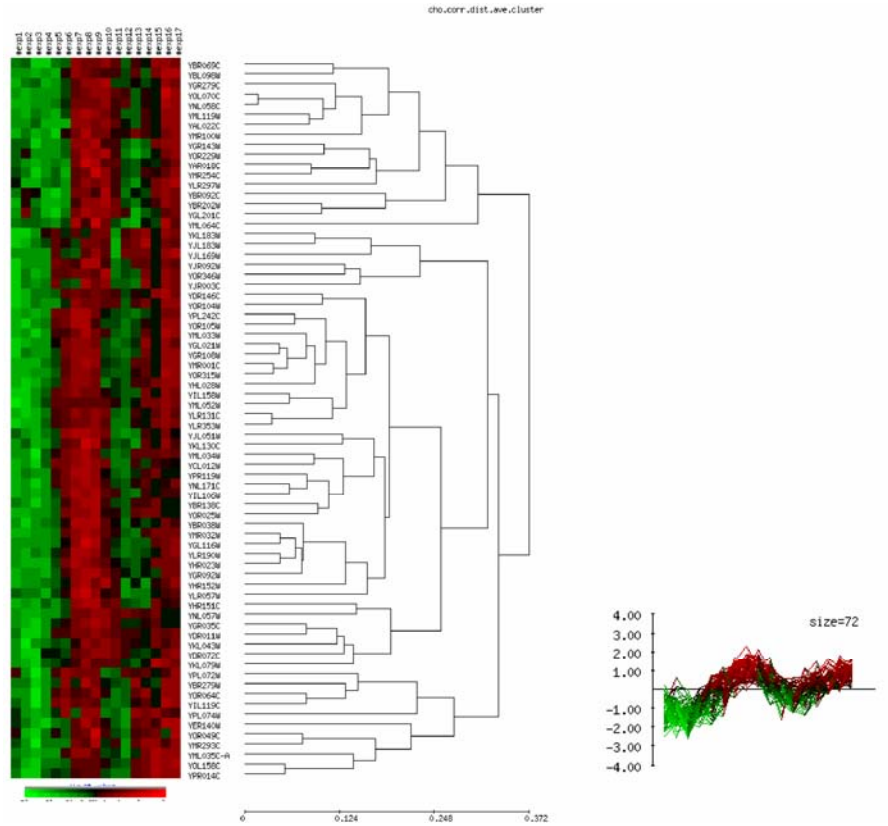


Figure 2.12: Clustering of microarray data. Hierarchical clustering of the dataset of Cho *et al.*, 1998 ([35]) representing the mitotic yeast cell cycle. After a selection of 3000 genes was made ([45]), hierarchical clustering was performed using the Pearson correlelation coefficient and an average linkage distance as implemented in EPCLUST ([30]). Only a subsection of the total tree is shown containing 72 genes. The columns represent individual experiments, the rows show the gene names. Green colour indicates down regulation while a red colour represents upregulation as compared to the reference. In the complete experimental set up a single reference was used (reference design).

One of the major challenges of the field of interdisciplinary biology is the ***inference of regulatory*** (or *genetic*) ***networks*** [9,25,83,146]. From a systems biology approach, a cell is considered a system that continuously interacts with its environment. The cell receives dynamically changing environmental cues and transduces these signals into the observed behaviour (i.e. change of phenotype or change of physiological response). This signal transduction is mediated by the regulatory network. A complete regulatory network can be seen as consisting of proteins interacting with each other, with DNA or with metabolites to constitute a complete signalling pathway [36,37,69,146,219]. Regulatory network inference became a big research topic as microarray technology made its way into mainstream biological research, but the underdetermined nature of the data made the construction of biologically relevant regulatory networks a daunting task. With the advent of diverse types of high-throughput data however, the research in network inference has received a new impulse, often integrating different data types (genomics, transcriptomics, proteomics and metabolomics) to obtain biologically plausible results. A comprehensive review of different methods and approaches can be found in Van den Bulcke *et al.* (2006) [210].

2.3 Extended applications

Spotted microarrays are predominantly used to study the expression profiles of specific cell types and tissue samples. Contrary to Affymetrix GeneChip technology, the differential labelling and resulting relative nature of the measurements of spotted microarrays renders them suitable tools for other types of genomic analysis. In this section we describe two such strategies: determining genomic copy number (attempts at which have also been made using GeneChip arrays [21]), and mapping DNA-protein interactions on the genome.

Comparative Genomic Hybridization or **CGH** identifies and maps sites of variation in DNA copy number throughout the genome [60,105,106] in a single measurement. In CGH as originally developed, total genomic DNAs from two (or more) cell populations are labelled with different fluorescent dyes and hybridized to metaphase chromosome spreads from a normal individual. The binding ratio of sequences from the different cell populations at the locations on the chromosomes to which they are complementary is proportional to their relative initial concentrations in the hybridization mixture. In the years since its inception, CGH has provided a wealth of information regarding regions of amplification and deletion in cancer cell lines and tumour samples [93,137,215,216]. Microarray implementations of CGH [153,154,185] have the potential to overcome many of the limitations of traditional cytogenetic CGH: using microarrays of mapped genomic clones permits the resolution of the measurements to be

determined by the spacing of the clones on the array and their size. The exact performance requirements for accurate measurement of copy number nevertheless make array CGH a very demanding microarray technology.

Chromatine immunoprecipitation or **ChIP** is a popular technology to identify the actual binding sites or *loci* for DNA binding proteins. It relies on the use of a specific antibody to immunoprecipitate the protein of interest together with its associated DNA sequences. Covalent cross-linking using formaldehyde allows specific interactions to be detected and has been used for mapping DNA-protein interactions in yeast [190], *Drosophila* [143], and mammalian cells [28]. This approach has been combined with microarray hybridization (***ChIP-chips***), initially in yeast [102,125,163,164], to enable mapping of DNA-protein interactions on a genomic scale. The specifically immunoprecipitated DNA is labelled using one fluorescent dye, and the second dye is used to label a reference sample which can consist either simply of genomic DNA, amplified and labelled in parallel, or DNA derived from a parallel immunoprecipitation reaction that serves as a negative control. The fluorescence log-ratio at an array element represents the enrichment of that locus in the immunoprecipitation sample and hence is indicative of the extent of binding of the protein to that locus. Due to the amount of random sequences that are extracted during the immunoprecipitation reaction however, a large number of replicate measurements are necessary to infer DNA binding sites with any statistical significance.

Chapter 3

Evaluation of ANOVA normalization

The normalization -and further analysis- of data from spotted microarray experiments is based upon measurements of relative expression, usually represented as log-ratios (see chapter 2). The technology however, is by no means restricted to the use of intensity ratios. The estimation of absolute expression levels from the measured intensities is not conceptually impossible. Calculation of log-ratios is thought to alleviate part of systematic variation, so any technique attempting to obtain absolute values of expression will have to be mindful of the various sources of systematic variation in a microarray experiment. The first method to work with absolute intensities, and still the most common to this day, is the use of ANOVA models to normalize microarray data.

This chapter represents a first phase of this PhD research, where the use of ANOVA for microarray normalization was evaluated and compared to a ratio based approach [132]. The performance of any normalization procedure, especially one that estimates absolute expression levels, is hard to assess as the actual levels of mRNA abundance in any particular condition are normally unknown. In order to nevertheless appraise the presence of any markedly flawed features in ANOVA normalized data, genes thought to be differentially expressed were selected for both log-ratios and data normalized by ANOVA. To minimize the influence of the used selection method, several different ratio based procedures for selecting differentially expressed genes were tested and compared to a selection procedure based on ANOVA normalization. A two sample colour flip design (see chapter 2, section 2.2.2.2) was chosen as test data set. Two sample colour flips are relatively simple designs for which several test procedures for identifying differentially expressed genes have been established. Moreover, most biologists start off with such straightforward experiments to roughly identify the genes involved in the biological system studied. Based on the conclusions drawn, more complex experiments are designed afterwards.

A first part of this chapter describes the principles of ANOVA based normalization of microarrays (section 3.1), and details some of the particular models that can be used for normalizing colour flip designs, as well as the statistics that can be used to select differentially expressed genes based on ANOVA normalized data. A second part (section 3.2) will provide a background to the log-ratio based methods for selecting differential genes, namely the fold test [157], the paired *t*-test [126], and a method called Significance Analysis of Microarrays or SAM [204]. The third section (section 3.3) describes the results of performing the analysis, a discussion of which can be found in the final section.

3.1 ANOVA models for normalization

3.1.1 Principles

ANOVA (ANalysis Of Variance; more detailed information is given in appendix B, for a complete dissertation on the topic we refer to Neter *et al.*, 1996 [141]) models are used for studying the relation between a response variable and one or more explanatory or predictor variables Specifically, single-factor studies are utilized to compare different factor level effects, to ascertain the best factor level, and the like. In multifactor studies, analysis of variance models are employed to determine whether the different factors interact, which factors are the key ones, which factor combinations are best and so on. Just as with regression models, the results of an ANOVA model fit can be represented in an *ANOVA table*, showing a factor-wise partitioning of the total sum of squares, the degrees of freedom, the resulting mean squares and possibly the result of statistical significance tests.

The proper use of ANOVA generally requires two major assumptions to be satisfied. At first the data should adequately be described by the linear ANOVA model. Secondly, observations should be normally distributed with constant within group variances equal for all groups. Satisfying both requirements results in the residuals of the fit (i.e. the difference between the observed and the fitted values) being independently and normally distributed random variables with zero mean and constant variance. Residuals are therefore highly useful for examining the aptness of ANOVA models. For instance, visual inspection of the residual plots can reveal much about the models behaviour. If the data can not be fitted by a linear model (i.e. not satisfying the first assumption), residual plots show pronounced non-linear trends, not satisfying the second assumption results in heteroscedasticity (a non-constant error variance), indicated by an observed wedge-shaped trend in the residual plot. Residual analysis can also detect other departures from

the ANOVA model, such as the prescence of outliers, non-indepence of error terms, non-normality of error terms, and the omission of important explanatory variables. It should be noted though, that it is not necessary, nor is it usually possible, for an ANOVA model to fit the data perfectly. ANOVA models are reasonably robust against certain types of departures from the model, such as the error terms being not exactly normally distributed. The major purpose of the examination of the appropriateness of the model is therefore to detect serious departures from the conditions assumed by the model [141].

With regards to microarray data, ANOVA has been used to address several problems that are not necessarily limited solely to the normalization of the data (e.g. assessing the contributions of age, sex and genotype into transcriptional variations [104,221], or between natural populations of the same species [142]). Considering the topic of this research, we will only focus on models that are designed to remove systematic sources of variation from microarray data. Such models were originally proposed by Kerr *et al*., 2000 [113]. When used for normalizing microarray data, ANOVA models describe the measured expression level of each gene as a linear combination of the explanatory variables that reflect the major sources of systematic variation in a microarray experiment. The residuals of the fit can be considered as estimates of random, experimental noise. Several explanatory variables representing condition (also referred to as *varieties* [108,109,111-113]), dye and array related variation and combinations of these variables are taken into account in the models. Only those interaction factors that have a physical meaning in the process to be modelled are retained. Reliable use of an ANOVA model for the normalization of microarray data therefore requires a good insight into this process. The *gene×condition* (*GC*) interaction effect however, is a key variable in all ANOVA models for microarray normalization, because it reflects how the expression of a gene depends on the biological conditions of the experiment (i.e. the condition-specific expression for that gene). This is the effect in which biologists are interested, and is thus referred to as the *factor of interest*. In the context of a two sample colour flip experiment, the difference between the estimated *GC* effects of a single gene reflects the differential expression (it is in fact a rescaled version of the average log-ratio described in chapter 2, section 2.2.2.2) and is called the *contrast of interest*.

If the aforementioned assumptions can be satisfied (measurements can be explained by a linear model, independent and normally distributed residuals), one of the major advantages of using ANOVA for normalization consists of its ability to assess the different sources of variation across the entire experiment (i.e. the entire set of arrays) instead of treating each slide separately. In contrast to a slide by slide approach, all measurements are combined during statistical inference. Moreover, due to the specific way of

modelling measured intensities as a superposition of different experimental factors, absolute values of gene expression (i.e. the estimated values for the *GC* interaction factor levels) are obtained and one is not bound to the use of log-ratios for further analysis of the data.

3.1.2 Models for normalizing colour flips

Three different models were used in this study to normalize a colour flip experiment. Two of these models were originally described by Kerr *et al.* [111,113], the third one is an adaptation on our part of the original models. The models differ from each other in the number of additional combined effects included.

The first more simplified model does only compensate for array, dye and condition effects. Let y_{ijkl} denote the intensity measurement from the i^{th} gene, j^{th} condition, k^{th} array and l^{th} dye, and let I_{ijkl} be the log-transformed intensity of this measurement, i.e. $I_{ijkl} = \log(y_{ijkl})$. The different sources of variation in a microarray experiment can be modelled as

$$I_{ijkl} = \mu + G_i + C_j + A_k + D_l + GC_{ij} + \varepsilon_{ijkl} \tag{3.1}$$

where μ is the overall average signal, the parameter G_i represents the effect of the i^{th} *gene*, the parameter C_j represents the effect of the j^{th} *condition*, the parameter A_k represents the effect of the k^{th} *array*, the parameter D_l represents the effect of the l^{th} *dye*, and the parameter $(GC)_{ij}$ represents the interaction between the i^{th} gene and the j^{th} condition, referred to as *gene×codition*. The error terms ε_{ijkl} are assumed to be independent and identically distributed with mean 0. The array effects A_k account for differences between arrays averaged over all genes, dyes, and conditions. These may arise, for example, because arrays are hybridized under slightly different conditions that result in a change in hybridization efficiency across an array. Similarly, the dye effects D_l account for differences between the average signal from each dye. One dye is often inherently "brighter" than the other, and this must be taken into account in the analysis. Remark that this dye effect corresponds to a simple linear rescaling, and does not compensate for intensity dependent variation (see section 2.2.2.3 of chapter 2). The terms C_j account for overall differences in the conditions. Such differences could arise because of differential concentration of mRNA in the labelled sample. The terms G_i account for average effects of individual genes spotted on the arrays in the experiment, i.e. they represent the 'basal' expression of a gene given all of the biological conditions that are surveyed. As explained before, the effects of interest in model (3.1) are the interactions between conditions and genes, the $(GC)_{ij}$ effects. These terms capture departures from the overall averages that are attributable to the specific

combination of a condition *j* and a gene *i*. Non-zero differences in *gene×condition* interactions across conditions for a given gene indicate differential expression.

In the second (3.2) and third (3.3) model spot effects are added, but each model does this in a different way. In the second model, each spot is modelled individually as if each spot were spotted with a different pin. In the original models of Kerr *et al*, 2000 [113] this was accomplished by incorporating a $(GA)_{ik}$ interaction factor. However, a single $(GA)_{ik}$ effect can only account for the intensity contribution of a single spot, when every spot corresponds to a specific interaction between a gene *i* and an array *k*, i.e. when a probe for a particular gene is only spotted once on every array (as was the case for data set used by Kerr *et al.,* 2000 [113]). In the data set used in this study however, cDNA probes were spotted in duplicate on every microarray. To compensate, an adaptation was made to the proposed model:

$$I_{ijklm} = \mu + G_i + C_j + A_k + D_l + R(GA)_{m(ik)} + GC_{ij} + \varepsilon_{ijklm} \qquad (3.2)$$

where $R(GA)_{m(ik)}$ represents the effect of the m^{th} replicate spot for a single *gene×array* interaction GA_{ik}. In this form, the *replicate* variable *R* is said to be 'nested' within the interaction variable *GA*. The GA_{ik} effect could have been left in the model, but was omitted to restrict the overly use of degrees of freedom that may otherwise serve to estimate the error variance in the experiment. By incorporating spot effects this way, the model stays true to the original concept of every spot on the array having a distinct parameter that quantifies its contribution to the measured intensity.

The third model, an adaptation on our part of the original models is slightly more complex. It is based on empirical observations and assumes a relationship between left and right spots on the same array.

$$I_{ijklm} = \mu + G_i + C_j + A_k + D_l + R(G)_{m(i)} + GA_{ik} + GC_{ij} + \varepsilon_{ijklm} \qquad (3.3)$$

In this model, the $R(G)_{m(ik)}$ effect represents all replicate spots on an array, the difference between these replicate spots is described by the replicate variable that is now nested within the gene variable. This particular nesting structure is meant to account for the variability of the probe source of multi-spotted genes, as the probes of a particular gene on different arrays usually originate from the same PCR amplification reaction or oligo set.

For all models, it is also possible to include other effects, such as *gene×dye* interactions. However, the physical or biological relevance of such additional effects usually does not outweigh the loss of degrees of freedom, needed to estimate the error variance in the experiment, that comes from including them.

3.1.3 Use of ANOVA model residual distribution

Parameters of an ANOVA model can be estimated through a constrained least-squares fit (for details on estimation procedures and obtained solutions, see section 4.1.1 of chapter 4). This results in a set of residuals e_{ijklm} and estimated values $\hat{I}_{ijklm}$, which are a linear combination of the estimated parameter values, so that:

$$I_{ijklm} = \hat{I}_{ijklm} + e_{ijklm} \tag{3.4}$$

As stated in section 3.1.1 the proper use of ANOVA generally requires two major assumptions to be satisfied: the data should adequately be described by the linear ANOVA model, and observations should be normally distributed with constant within group variances equal for all groups. When both assumptions are satisfied and the residual distribution will show only slight deviations from normality (so that the actual model errors can be assumed to be normally distributed) and significantly differentially expressed genes can be identified based on ***normal*** assumptions. In the particular case of a colour flip design, this would be done by constructing confidence intervals on the difference in *GC* effect.

If the distribution of the residuals shows serious deviations from normality, confidence interval construction can still be done, but bootstrap analysis should be used as an alternative. In ***bootstrap analysis*** [50,67,68] no explicit assumption on the distribution of the errors is made (somewhat similar to the permutation analysis of SAM described in section 3.2.3), but confidence intervals are estimated based on novel *in silico* generated datasets. The only assumption is that the errors are identically and independently distributed (*iid*), i.e. assuming a constant error variance. Simulated datasets I^*_{ijklm} are created as:

$$I^*_{ijklm} = \hat{I}_{ijklm} + e^*_{ijklm} \tag{3.5}$$

where the e^*_{ijklm} are drawn independently from $\hat{E}\sqrt{n/(n-p)}$, $\hat{E}$ being the empirical distribution of residuals from the original fit of the model, n the total number of intensity measurements in the experiments, and p the total number of degrees of freedom in the model. Rescaling $\hat{E}$ produces a distribution of which the variance is closer to that of the real error term [113,223]. By adding residuals, randomly sampled with replacement from this rescaled residual distribution, to the estimated expression values in this way, thousands of novel bootstrapped datasets can be generated. In the particular case of a colour flip design, a measure for the differential expression of each gene can be calculated for such a novel dataset as the difference in *GC* effect between the two conditions can be calculated. Based

on these thousands of estimates of the difference in *GC* effect, a bootstrap confidence interval can be obtained [110,111,113].

3.2 Identifying differentially expressed genes from log-ratios

When consistent sources of variation have been removed by normalization, the replicate log-ratio measurements for a particular gene can be combined to find out whether a gene is differentially expressed. A plethora of methods is available to identify differentially expressed genes in a statistically more founded way than the simple heuristic of a *fold test* [12,31,34,58,126,144,157,199,204,222,233]. Distinct classes of models can be discerned (see Table 3.1), differing from each other in the test statistic used, in the way the null hypothesis is modelled, and in their underlying assumptions. In this study, three different ratio based methods for selecting differential genes were compared to the ANOVA-based bootstrapping: the fold test, the *t*-test described by Long *et al.*, 2001 [126] and the SAM (Significance Analysis of Microarrays) method of Tusher *et al.*, 2001 [204]. Though quite advanced, these methods are still most intuitive and straightforward to understand for non-expert users, and cover a broad range of different approaches (see Table 3.1). They are outlined in this section, but for the in depth technical details of each of these methods we refer to the individual references.

3.2.1 Fold test

The fold test is a simple selection procedure that makes use of an arbitrary chosen threshold. For each gene *i* an average log-ratio $\overline{M}_i$ (arithmetic mean of the replicate log-ratios) can be calculated as:

$$\overline{M}_i = \bar{I}_{i,KO} - \bar{I}_{i,WT} \tag{3.6}$$

with $\bar{I}_{i,KO}$ being the average of the logarithm transformed intensities for gene *i* in the knock-out condition, and $\bar{I}_{i,WT}$ being the average of the logarithm transformed intensities for gene *i* in the wild-type condition. Average log-ratios that exceed a certain threshold (usually chosen to correspond to a twofold expression ratio) are retained. The fold test is based on the intuition that a larger observed fold change can be more confidently interpreted as a stronger response to the environmental signal than smaller observed changes. Note that a fold test discards all information obtained from replicates [12].

Table 3.1: Overview of methods to determine differentially expressed genes across two conditions.

Method	Assumptions	Test statistic	Error restrictions	Distribution H_0: $\mu_1 = \mu_2$	Additional modifications
Regular *t*-test for equality of means	Observations are independent Observations are normally distributed Unequal sample variance Unequal sample size	$t_i = \frac{\bar{I}_{i1} - \bar{I}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$	Errors normally distributed	Parameterized: student *t* distribution	Empirical Bayesian estimate of variance
Paired *t*-test for equality of means	Each pair of measurements is independent Differences are normally distributed Unequal sample variance Equal sample size	$t_i = \frac{\bar{I}_{i1} - \bar{I}_{i2}}{\sqrt{\frac{s_{i1}^2 + s_{i2}^2 + 2\,cov(y_{i1}, y_{i2})}{n}}}$	Errors normally distributed	Parameterized: student *t* distribution	
Weighted least squares	Unequal sample variance Unequal sample size	$t_i = \frac{\bar{I}_{i1} - \bar{I}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} \cdot \frac{n_1 - 1}{n_1} + \frac{s_{i2}^2}{n_2} \cdot \frac{n_2 - 1}{n_2}}}$	Unequal error variances acceptable	Parameterized: standard normal distribution	Weighted least squares
Mixture model approach	Unequal sample variance Unequal sample size	$t_i = \frac{\bar{I}_{i1} - \bar{I}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$	Errors equal variance (iid) and symmetrically distributed	Test statistic used in a likelihood ratio test	H_0 estimated by normal mixture models
SAM	Equal sample variance: use of 'pooled' variance Unequal sample size	$t_i = \frac{\bar{I}_{i1} - \bar{I}_{i2}}{s_0 + s_{ip}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s_{ip}^2 = \frac{(n_1 - 1)s_{i1}^2 + (n_2 - 1)s_{i2}^2}{n_1 + n_2 - 2}$	Errors equal variance (iid)	No explicit H_0 distribution but use of order statistics	Addition of s_0 to ensure distribution of t_i is independent of gene expression level

Note on table 3.1: Each of the methods uses variations of a mean and variance normalized test statistic t_i. For each gene i the test statistics t_i is calculated. $\bar{I}_{i1}$ ($\bar{I}_{i2}$): average logarithm transformed expression level of the n_1 (n_2) replicates of gene i in the first (second) condition, s_{i1} (s_{i2}): within variance of this group of replicates (e.g. for the data set presented in this chapter: condition 1 refers to knock-out, condition 2 refers to wild-type). Based on the calculated t_i value, a preset significance level and the degrees of freedom a corresponding p-value is calculated. The p-value expresses the probability of finding a certain value of the test statistics t_i by coincidence, assuming that both genes were not differentially expressed (H_0 hypothesis). The methods also differ from each other in the way the corresponding significance level is calculated. A first class of methods (regular and paired t-test for equality of means [12,126], and weigthed least squares [199]) makes use of simple t-test statistics. As a H_0 distribution a parameterized (Student t-distribution) is used for small sample sizes. Due the small sample size and the corresponding low degrees of freedom, t-tests have a low power, i.e. they tend to miss a number of real positives. Non-parametric alternatives to the t-test and the paired t-test respectively are the Wilcoxon Rank Sum test and the Wilcoxon Signed rank test. For a sufficiently large sample size, the test statistic t_i used by weighted least squares and that of the regular t-test may be considered equal. For small sample sizes the weighted least squares procedure of Thomas *et al.* (2001) [199] makes use of the maximum likelihood estimator of the variance. The advantage is that it does not assume a constant variance of the error term. However, to calculate a significance level, H_0 is assumed to be normally distributed, which might be too strong an assumption considering the small sample size. A t-distribution might have been a more appropriate choice. A second class of models (mixture model approach [144] and SAM [204]) estimates the distribution of H_0 directly by permutation analysis (a comparable method is used by Kerr *et al.*, 2000 [113]). The mixed model described by Pan *et al*. 2002 [144] make use of complex estimation procedures to determine the distribution of H_0 while SAM [204] uses order statistics. Moreover, the SAM method assumes that the standard deviations σ_{i1} and σ_{i2} are equally distributed and therefore uses the pooled standard deviation s_{ip} as an estimator of $\sigma_{i1} = \sigma_{i2} = \sigma_i$.

3.2.2 *t*-test

A *t*-test is a hypothesis test that assumes that the observations are drawn at random from a normal population and that employs a Student *t*-distributed test statistic for confidence interval estimation. The *t*-distribution describes the distribution of a normal variable, standardized with the sample variance s^2 as opposed to the population variance σ^2. It is used for hypothesis testing of normally distributed variables when the population variance σ^2 is unknown, in which case the sample variance s^2 is used as an estimator of σ^2. It is more appropriate to make statistical inference about the differential expression of a gene than a simple fold test, since it does not only take into account how much a gene is differentially expressed, but also the consistency of the individual measurements used to assess the average differential expression level. The non-paired *t*-test evaluates if the average expression level of a gene in the test condition is significantly different from its average expression level in the reference condition. The H_0 hypothesis states that the expression level of the test and reference are equal. The formula to compute the test statistic is depicted in Table 3.1. To calculate the within sample variance of a regular non-paired t-test, the four observations of the test are used to estimate the mean expression level of the gene in the test condition. In the same way the four measurements of the reference are considered as a single group. The standard deviations ($s_{i,KO}$, $s_{i,WT}$) are computed based on the deviation of the different measurements of a group from their respective group means ($\bar{I}_{i,KO}$, $\bar{I}_{i,WT}$). Of course when the within variance is calculated in such a way, it intrinsically contains the consistent variations due to array and spot effects (the absolute expression values instead of the ratios are used to calculate an estimate of the average differential expression level). This problem can be overcome by using a paired *t*-test.

The paired *t*-test is a special case of the two-sample *t*-tests of hypotheses that occurs when the observations on the two populations of interests are collected in pairs (in a cDNA microarray experiment, measurements of the Cy5 and Cy3 channel for a particular gene, assessed on the same array and the same spot, are paired). The difference with an unpaired two-sample *t*-test is that both variables are presumed to be dependent. This translates into the incorporation of the covariance between both variables in the test statistic. As a result, a positive correlation within the pairs can cause the unpaired two-sample *t*-test to considerably understate the significance of the data if it is incorrectly applied to paired samples. In Table 3.1 is outlined how a paired *t*-test is calculated for spotted microarray data. For computation of the variance, a pair of observations can considered as a new variable. The within group variation, as calculated by a paired *t*-test evaluates the deviation of this new variable from the mean of that variable,

taking into account the covariance between log-intensities obtained from the same spot. As such a paired *t*-test, in contrast to a regular non-paired *t*-test intrinsically compensates for the variation over spots and arrays. The lower within group variation increases the power of a paired *t*-test as compared to a regular *t*-test. In practice, the advantage of a (paired) *t*-test is that smaller fold changes are considered significant for genes whose expression levels are measured with great accuracy (high consistency), and large fold changes are considered non-significant if expression levels were not measured accurately (low consistency).

Usually a *t*-test is combined with a correction for multiple testing. When considering a family of tests, the level of significance and power are not the same as those for an individual test. For instance, a significance α of 0.01 for individual gene expression indicates a probability of 1% of finding a ratio similar to the measured ratio under the null hypothesis (no differential expression present). This means that for every 1000 genes tested (a family of 1000 tests), 10 would be expected to pass the test, though not differentially expressed. To limit this number of false positives in a multiple test, a correction is needed. The implementation of Baldi and Long, 2001 (Cyber-T) uses a Bonferroni correction [12]. The choice of the Bonferroni correction factor however (see chapter 4 of Neter *et al.* [141]), is quite arbitrary and due to the immense amount of simultaneous tests (i.e. thousands of genes) the single step adjusted *p*-values, as implemented in the Cyber-T software, decrease the power of the statistical test (ability to detect real positives). To handle these pitfalls, other corrections for multiple testing have been proposed [58]. Long *et al.*, 2001 [126] provide other extensions to their implementation of the *t*-test such as the Bayesian *t*-test, a methodology developed to cope with the low number of replicates.

3.2.3 SAM

SAM (Significance Analysis of Microarrays) is another method for the analysis of paired or unpaired black and white experiments [204]. SAM calculates for each gene a modified *t(i)* statistic, called relative difference and referred to as *d(i)* in the original article. The difference between a *t*-test statistic *t(i)* and the *d(i)* values calculated by SAM, is the constant term s_0, used to compensate for the dependency of the distribution of *d(i)* on the measured expression level. Genes are ranked according to their *d(i)* value and the higher the absolute *d(i)* value, the more likely that the gene will be differentially expressed. Instead of calculating a *p*-value using a Student *t*-distribution, genes called differentially expressed are identified by performing a permutation analysis. New random datasets are generated by permuting the original data. In these permutated datasets, none of the genes is differentially expressed. The *d(i)* values in these randomized datasets are

calculated, ranked, and subsequently used to infer the expected differences, i.e. the *d(i)* value that can be expected if a gene is not differentially expressed. By using a scatter plot (Figure 3.1), ranked *d(i)* values of the experimental dataset are compared to ranked expected *d(i)* values.

The delta value δ, a user-specified parameter determines the number of significant, differentially expressed genes; it expresses how much the measured *d(i)* value should exceed the expected one in order to consider a gene diferentially expressed (δ is measured as a displacement of the *d(i)* value from the $d(i) = d_{expected}(i)$ line). The number of false positives can be estimated as the number of genes present in the permuted dataset, for which the *d(i)* value exceeds the lowest *d(i)* value that was considered significant based on a given setting of the delta value δ. Permutation analysis overcomes the need of a high number of replicates and is used as an alternative to correction for multiple testing. The setting of the delta slider allows choosing a trade-off between the number of false positives (type I error) and the number of false negatives (type II error). The lower the number of false positives, the more stringent the test and the fewer genes will be withheld as significant.

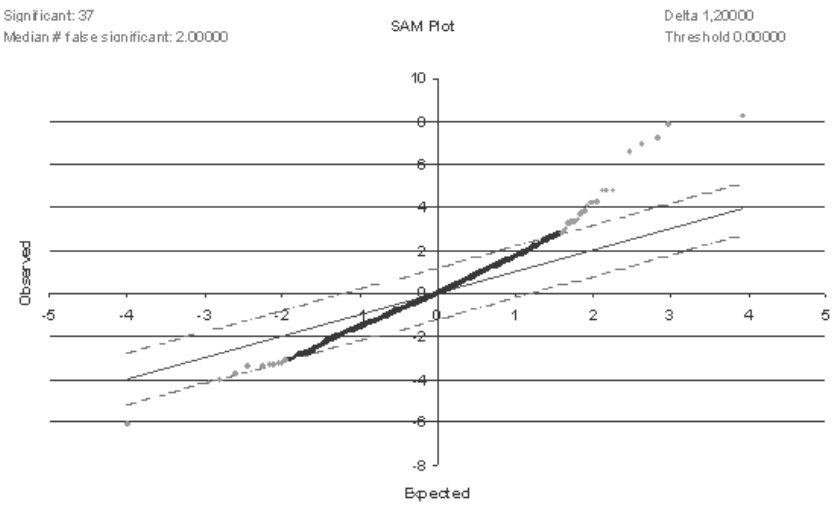


Figure 3.1: SAM analysis output. Expected differences in expression are plotted against observed differences in expression. The value δ determines the threshold (dotted lines) which is used to select genes of which the observed change is in expression is sufficiently different from the expected one, i.e. the genes that can be assumed to exhibit differential expression (grey dots). The method also gives an indication of the number of false positives (False Discovery Rate or FDR). The result depicted is that of the LOWESS normalized dataset (see section 1.3.3).

3.3 Results

3.3.1 Data set

The dataset used in this review compares a spontaneous knock-out (KO) and wild-type (WT) mouse (data kindly provided by Prof. T. Ayoubi; the experiment was conducted at the VIB MicroArray Facility by Dr. P. Van Hummelen). In the spontaneous knock-out mouse, the Hmgi-c gene has been disrupted by a 100kb deletion. Hmgi-c RNA was consequently not transcribed and therefore did not result in a protein. HMGI proteins play a critical role at promoter regions in the correct assembly and stabilization of higher order protein-DNA complexes required for efficient transcriptional activation of genes [103].

From both mice mRNA was extracted, labelled and hybridized on a mouse cDNA microarray containing 4202 cDNA fragments of 0.5 to 2kb. The cDNA fragments were PCR amplified, purified and spotted in duplicate on Type-VII silane coated slides (catalogue number RPK0174, Amersham BioSciences, UK) using a Molecular Dynamics Generation III printer with 12 capillary quill pins (Amersham BioSciences). Duplicate spots were arrayed distant from each other on the left and right hand side of the slide. For the probes, 5 μg of total RNA was amplified using a modified protocol of *in vitro* transcription as described earlier and labelled during a reverse transcription reaction of the amplified RNA [155] with either Cy3-dCTP (green dye) or Cy5-dCTP (red dye). The probes were mixed and hybridized overnight using an automatic slide processor (Amersham BioSciences). Hybridizations were repeated in the following way: in a first analysis, the test sample (KO) was labelled with the Cy5 (red) dye while the corresponding reference (WT) was labelled with the Cy3 (green) dye, and in a second analysis the colours were reversed (i.e. colour flip experiment). Since every gene was spotted in duplicate, this design resulted in four measurements per gene for each condition tested.

3.3.2 Data preparation

Prior to performing the distinct statistical tests, data were preprocessed as outlined in this paragraph. Raw intensities were corrected with a local background subtraction and the resulting values were log-transformed. Genes, for whom at least one intensity measurement contained a value smaller or equal to zero, were treated separately. Dividing by, or taking the logarithm of, negative or zero values during analysis will result in undefined (missing) values. If these genes are not treated separately, the information

about such genes is lost. In a two-sample experiment, below zero values in one particular condition might correspond to genes differentially switched off, i.e. being of great interest to the posed biological question. In this particular example all genes containing below zero intensities behaved inconsistent, indicating that the value of zero was dye dependent rather than condition dependent. Genes consistently switched on in one condition and off in the other were not detected.

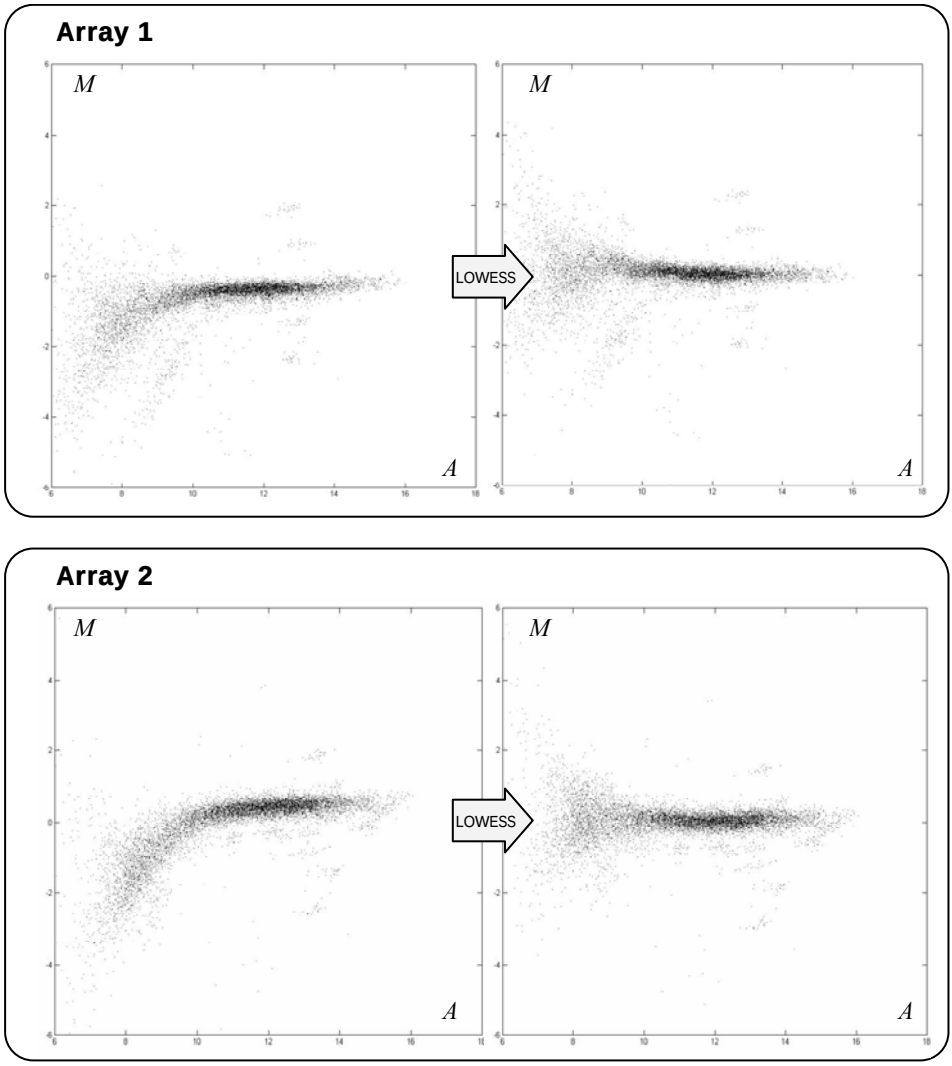
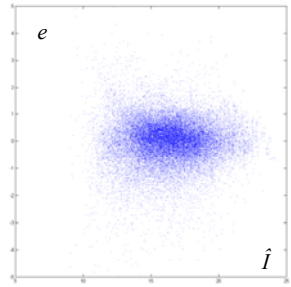


Figure 3.2: LOWESS normalization of the data. Effect of performing a LOWESS normalization on the data illustrated by MA-plots for both arrays. The upper two plots correspond to one array, the lower two correspond to the other. Original background corrected intensities are shown in the left plots, LOWESS corrected intensities are shown in the right plots.

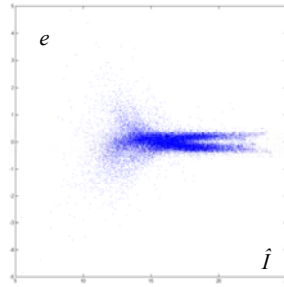
A **MODEL (3.1)**

Source	SS	df	MS
G	133176.3	3784	35.2
C	4536.6	1	4536.6
A	22432.4	1	22432.4
D	4822.5	1	4822.5
GC	1313.6	3784	0.3
Error	19210.4	22708	0.8
Total	185492.2	30279	6.1



B **MODEL (3.2)**

Source	SS	df	MS
G	133176.3	3784	35.2
C	4536.6	1	4536.6
A	22432.4	1	22432.4
D	4822.5	1	4822.5
R(GA)	15005.5	7570	2.0
GC	1313.6	3784	0.3
Error	4204.9	15138	0.3
Total	185492.2	30279	6.1



C **MODEL (3.3)**

Source	SS	df	MS
G	133176.3	3784	35.2
C	4536.6	1	4536.6
A	22432.4	1	22432.4
D	4822.5	1	4822.5
R(G)	2164.7	3785	0.6
GA	9052.6	3784	2.4
GC	1313.6	3784	0.3
Error	7993	15138	0.5
Total	185492.2	30279	6.1

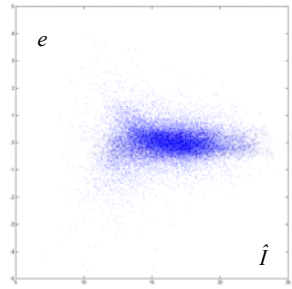


Figure 3.3: Results of the three different ANOVA models tested. Data were log-transformed, genes containing at least 1 zero value were removed, but no normalization by LOWESS was performed. ANOVA models used: μ: overall mean of the expression levels, *A*: array effect, *D*: dye effect, *G*: gene effect, *C*: condition effect, *GC*: effect of interest, *R*: replicate effect, *AG*: combined effect representing a spot effect, *i*: number of genes, *j*: number of conditions, *k*: number of arrays, *l*: number of dies, *m*: number of replicates. ANOVA tables represent for each effect in the corresponding ANOVA model its contribution to the total variance (*SS* = sum of squares error). The residual SS, represented by *Error* is the variation in the dataset that could not be explained by any of the effects. The total variation in the dataset represented by *Total*. *Df* denotes the degrees of freedom, *MS* the mean square error. Corresponding residual plots represent for each ANOVA model the plot of the residuals (*e*) versus the estimated intensity values ($\hat{I}$). If the assumptions underlying an ANOVA model are satisfied residual plots should be structureless. The possible causes for the observed heteroscedasticity in the residual plots are explained in the text.

For the fold test, *t*-test and SAM procedure, log-ratios were calculated and subsequently normalized for dye related biases by performing intensity based rescaling (LOWESS fit with smoothing parameter *f* set to 30%, [226]). The results of this normalization are depicted in Figure 3.2.

When using ANOVA, different models can be devised to normalize colour flip designs. Three such models were described in section 3.1.2, all of which were evaluated to select the most appropriate one for the colour flip experiment under study. Figure 3.3 depicts the ANOVA tables and their corresponding residual plots that are obtained when fitting each of the three models to the data. Residual plots were used to check if the models were appropriate and if the assumptions were satisfied. The sum of squares values (SS) describe for each effect its contribution to the global variation in the experiment. From Figure 3.3 it is clear that the contribution of array and gene effects had the highest impact. Given the biological question underlying the experiment, only a limited number of genes were expected to exhibit a differential expression between wild-type and knock-out. The marginal influence of the *GC* effect, as compared to the influence of the other effects, seemed to support this. Because of the unrealistic assumption of individual spot effects in the second model (i.e. assuming that each gene is spotted by a different pin represented by the *AG* effect), the number of measurements available to estimate the spot effects were too low (only two values available per AG effect). The model 'over-fitted' the data as is also shown by the extremely low contribution of the residual error to the global variation in this model and the strange behaviour observed in the residual plot. This observation underlines the danger of arbitrarily omitting and including effects and emphasizes the importance of choosing a realistic model. The third model, probably best adapted to the technological reality of the process, seemed to perform best and was selected for further analysis, as it had less obtrusive trends in the residual plot (compared to the second model), and showed a smaller error variance than when data were normalized with the first model.

All model fits however, suffered from another problem, which is illustrated for the third model in Figure 3.4. Residuals were far from normally distributed and showed an apparent slight heteroscedasticity (a non-constant variance of the residuals) at low expression levels. The observed residual behaviour can either be caused by not satisfying the underlying constant variance assumption of the data (second assumption), or in the worst case because the data are not adequately modelled by a linear model (first assumption). By plotting the residuals against the estimated values for the individual combinations of effects (see Figure 3.5 A) it was clear that the observed heteroscedasticity did not only result from non constant variance in the dataset (presence of additive error in the low expression range) but that non-linear effects occurred in the data. As explained in section 2.2.2.3 of

chapter 2, microarray data often show a strong non-linear dye bias across the intensity range. Such behaviour prohibits readily using linear ANOVA models on non-linearized data. To minimize their influence, non linear dye effects were removed by performing a LOWESS fit (smoothing parameter *f* set to 30%) prior to ANOVA. The results of fitting the third ANOVA model on LOWESS modified data are depicted in Figure 3.5. The impact of the LOWESS normalization is reflected by the zero contribution of the dye and condition effects in the ANOVA table. From the residual plot it is clear that linearization by LOWESS could not completely remove the heteroscedasticity in the residuals. However, when residuals were plotted for each array and dye separately as shown in Figure 3.5 B, it is clear that non-linear tendencies have sufficiently been removed when performing LOWESS prior to ANOVA. The estimated parameters and rescaled residuals of the third model fitted to LOWESS normalized data were used for bootstrap analysis and selection of differential expression.

C **LOWESS + MODEL (3.3)**

Source	SS	df	MS
G	133176.7	3784	35.2
C	0.0	1	0.0
A	22432.4	1	22432.4
D	0.0	1	0.0
R(G)	2164.7	3785	0.6
GA	9052.6	3784	2.4
GC	1239.8	3784	0.3
Error	5656	15139	0.4
Total	173722.2	30279	5.7

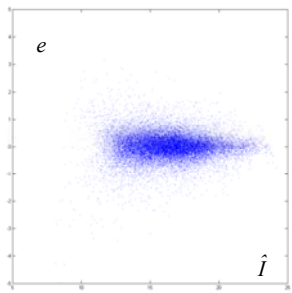


Figure 3.4: **Results of ANOVA model (3.3) fitted to LOWESS nor**malized data. The ANOVA table and the corresponding residual plot of fitting ANOVA model (3.3) to the preprocessed data. In this case, data were log-transformed, genes containing at least 1 zero value were removed and in addition data were LOWESS normalized. Other symbols as in Figure 3.5. The zero entries for condition and dye factors in the ANOVA table are a result of the array-by-array LOWESS normalization. The residual plot exihibits a smaller variance compared to those of panels A and C of Figure 3.5, while showing a more homogeneous distribution (no artifacts) the the one depicted in panel B of Figure 3.5.

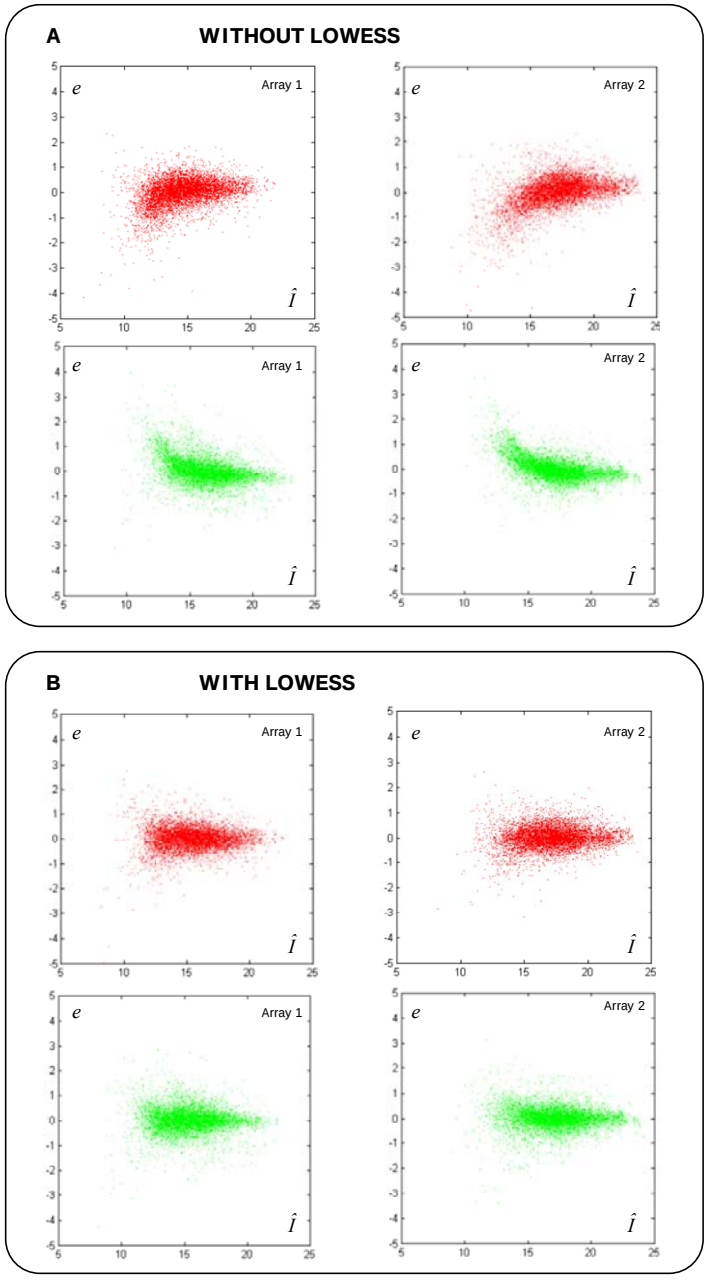


Figure 3.5: **Non-linear effects.** Influence of non-linear effects in the data on the residual plots of the ANOVA model. Panel A: residual plots for the application of ANOVA model (3.3) on the data plotted separately for each array and each dye combination. Data were log-transformed, genes containing at least 1 zero value were removed. Panel B: residual plots for the application of ANOVA model (3.3) on the LOWESS normalized data plotted separately for each array and each dye combination. Data were preprocessed as before, but an additional LOWESS normalization prior ro fitting the ANOVA model allowed removal of strong nonlinear dye effects.

Table 3.2: Overview of the number of statistically differentially expressed genes, as identified by the methods tested.

Method	Number of genes called significant
Fold test	110
t -test	106
SAM	106
ANOVA 95%	163
ANOVA 99%	71

Note: Parameters for each method were chosen as such that each method withheld approximately the same number of genes.

3.3.3 Comparison of the different methods

In this section, the output of the fold test, the *t*-test, SAM and the ANOVA-bootstrap method are compared. Since the number of genes called significantly differentially expressed depends on the specific parameter setting of each method (threshold for fold test, *p*-value for *t*-test, delta slider for SAM, significance level for bootstrap confidence intervals), parameter settings were chosen such that each method predicted approximately the same number of genes as being significant (Table 3.2). All methods were performed on the data preprocessed as outlined in section 3.3.2, and yielded the following results:

- After normalizing the data with ANOVA model (3.3), 163 genes were identified as potentially differentially expressed based on a 95% bootstrap confidence interval, and 71 genes based on a 99% bootstrap confidence interval.
- Using a two-fold threshold for the fold test, 110 genes were selected as being differentially expressed.
- Using the paired *t*-test of Baldi and Long [12] on our dataset resulted in 186 genes with an individual *p*-value lower than 0.01 and 106 genes with a *p*-value lower than 0.005. Only 3 genes in our dataset passed the significance test after correction for multiple testing (assuming an experiment wide false positive rate of 0.25). Therefore, the single step adjusted *p*-values, as implemented in the Cyber-T software are seemingly too conservative, decreasing the

power of the statistical test (ability to detect real positives), and were omitted from these analysis.

- Using a paired test and a value for the delta slider of 0.93, 106 genes were considered as differentially expressed by SAM, with a median number of false positives of 7 (Figure 3.1).

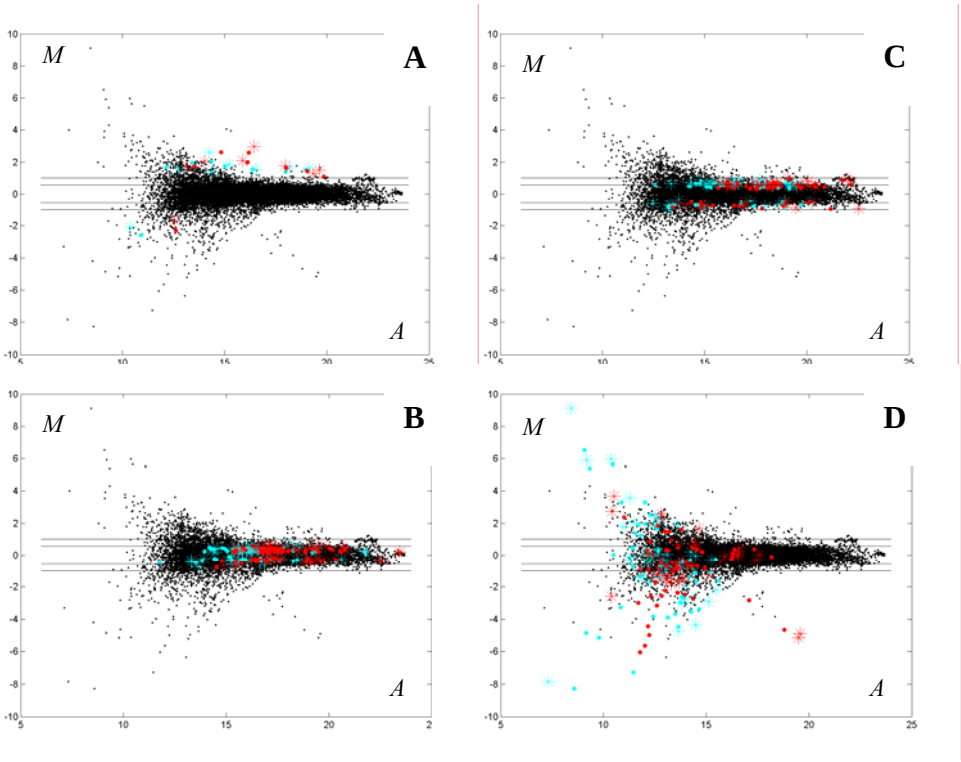


Figure 3.6: Detailed representation of distinct groups of differential genes. Average log-intensities *A* are plotted against LOWESS normalized log-ratios *M* for both arrays in a single figure. Black dots: normalized expression level of all 3785 genes. Dots colored otherwise indicate the expression levels of the genes showing the profile of the corresponding groups. Cyan and red: expression levels as measured on the first array or second array respectively. +: right spots; *: left spots. Dashed lines indicate 1.5 fold and 2 fold levels of over and under-expression respectively. A: genes detected by all methods (group 1 in Table 3.3). B: genes only detected by the *t*-test (group 10 in Table 3.3). These measurements were very consistent but probably too close to zero (not differentially expressed) to be biologically relevant. C: genes only detected by *t*-test and SAM (group 11 in Table 3.3). These measurements were consistent and sufficiently different from zero (differentially expressed) to be detected by resampling methods such as SAM. D: genes detected by the fold test and ANOVA-based methods only (group 9 in Table 3.3). Due to their high average expression value these genes were considered as being significant, but the consistency of these genes was remarkably low. Most of the data points were located in the region of low average intensity. In this range the ratio becomes a poor estimator of differential expression. Due to the heteroscedasticity in the data, bootstrap based confidence intervals systematically underestimated the variation at low average intensity levels and failed to reject these potentially false positives.

Table 3.3: Overview of the performance of the different methods that were tested.

	Genes	Under-expressed	Range	Over-expressed	Range	Range *p*-value	Fold test	*t*-test	SAM	ANOVA 95%	ANOVA 99%
Reliable genes											
Group 1	8	1	> -2	7	> 2	< 0.005	+	+	+	+	+
Group 2	3	0	> -2	3	> 2	< 0.005	+	+	+	-	+
Group 5	6	1	-1.71	5	[1.93; 1.94]	< 0.005	-	+	+	-	-
Group 11	36	25	[-1.85, -1.43]	11	[1.37; 1.80]	< 0.005	-	+	+	-	-
Genes of mediocre reliability											
Group 3	18	10	> -2	8	> 2	0.005-0.01: 4 0.01-0.05: 13 > 0.05: 1	+	-	+	+	+
Group 4	3	1	> -2	2	> 2	0.01-0.05: 3	+	-	+	-	+
Group 6	6	4	[-1.97; -1.71]	2	[1.82; 1.85]	0.005-0.01: 3 0.01-0.05: 3	-	-	+	-	+
Group 12	28	4	[-1.76; -1.60]	24	[1.50; 1.72]	0.005-0.01: 19 0.01-0.05: 7	-	-	+	-	-
Genes of low reliability											
Group 10	51	19	[-1.51; -1.09]	32	[1.14; 1.38]	< 0.005	-	+	-	-	-
Group 7	41	21	[-0.50; -0.55]	20	[1.81; 1.97]	0.01-0.05: 5 > 0.05: 36	-	-	-	-	+
Group 8	33	22	> -2	11	> 2	0.01-0.05: 2 > 0.05: 31	+	-	-	-	+
Group 9	45	30	> -2	15	> 2	0.01-0.05: 1 > 0.05: 44	+	-	-	+	+

Note: Genes were grouped as follows: a binary profile was assigned to each gene indicating whether the gene was detected (+) or not (-) by the methods tested, and genes with the same binary profile were grouped. Each group of genes is characterized by a *p*-value range, reflecting the consistency of its replicates and the range of over- or under-expression. Based on these characteristics the performance of the distinct methods was evaluated. Genes: number of genes within a group; Over-expressed: number of over-expressed genes within a group; Under-expressed: number of under-expressed genes within a group; Range: range of differential expression, determined as the maximal and minimal levels of over(under)-expression of the individual genes belonging to that group, levels of over(under)-expression are expressed as fold over-expression (positive values) or fold under-expression (negative values). Range *p*-value: determined as the maximal and minimal *p*-values of the individual genes belonging to that group.

Of the 3785 genes spotted on the microarray slide, 246 genes were detected by at least one of the methods tested. Results are summarized in Table 3.3. Validating these results is difficult and can hardly be done straightforward, as it is of course unknown which genes are actually differentially expressed. To facilitate inter-comparability, each group of genes in Table 3.3 are characterized by their range of average expression ratios and *p*-values (as determined by the *t*-test). Both of these traits were used as guidelines for interpreting the obtained results. The 'average expression ratio' was considered because, in all statistical tests, such a ratio was used as an estimator of the differential expression. When using a fold test, *t*-test, or SAM, this constitutes the average log-ratio $\overline{M}$. When using ANOVA, differential expression is estimated as a difference in *GC* effects, which, in the case of a colour flip design, can be considered as a rescaled log-ratio. The *p*-value, as calculated by the paired *t*-test, can be considered an indication of the consistency of a particular measurement. A lower *p*-value reflects a low variation between the replicate measurements for the ratio estimate of that gene. This means that the better the specific characteristics of genes belonging to a group (higher differential expression level and more consistent measurements), the more reliable the predictions on the genes within that group are assumed to be. Comparing these gene characteristics of the different groups allowed to make conclusions about the performance of the different methods tested, which are outlined below.

Only 8 genes were detected by all methods (see Table 3.3, group 1). Given that the number of differentially expressed genes identified by the individual methods ranges up to 100 and over, this points towards a rather low degree of agreement between the different methods in the prediction of the differentially expressed genes.

Genes that were called differentially expressed merely based on a fold test showed a huge variation across the different replicate measurements. As can be seen in Table 3.3, in group 8 and 9 the high *p*-values reflected this low consistency. These genes would have been rejected by tests that take into account explicitly the within group variation (such as a paired *t*-test or SAM). Indeed, the choice of a constant arbitrary threshold implicitly assumes that the variance among replicates is the same for every gene. This is, however, not the case since the variation on the ratio, as estimator of the differential expression, depends on the variation of the absolute signals that constitute the factors of that ratio. Low absolute expression values in one of the two channels results in unstable, often artificially high ratios. As such, a fixed ratio threshold of 2 gave rise to a high number of false positives, especially in the low expression range where the signal to noise ratio is low. However, as the intensities in both the channels increase, the ratios theoretically become a more reliable estimate of the differential expression. In this region a fixed ratio threshold of 2 might have been too stringent.

Different variants of the fold test have been described hitherto that are based on additional series of filtering steps, e.g. a filtering step removing all genes below a certain signal to noise level. Though likely to give better results than the fold test as described here, these fold tests make use of arbitrarily defined thresholds and are not statistically founded.

From this perspective the paired *t*-test is a better alternative to the fold test. It does not only focus on the extent to which a gene is differentially expressed, but also takes into account the variation across the different measurements used to determine this average differential expression level. Indeed, genes that are retained by the paired *t*-test will per definition behave consistently (only genes with a *p*-value smaller than 0.005). However, what is often observed is that the lower the signal, the more consistent genes tend to behave. This could be observed in our dataset in group 10 (Figure 3.6) that represent all genes retrieved by the paired *t*-test only. Although behaving consistently, these genes were almost not differentially expressed (relative expression value in logarithmic scale close to zero). The observed consistency might have been merely coincidence. These genes were indeed rejected by the resampling based methods (SAM, and ANOVA followed by bootstrap), and are probably irrelevant from a biological point of view (Table 3.3, group 10). Therefore using a paired *t*-test alone will probably result in the retrieval of consistently behaving but not necessarily differentially expressed genes. On the other hand, the paired *t*-test apparently missed a number of presumably real differentially expressed genes in the data set. Judging from Table 3.3, group 3 and group 4 contained genes that were rejected based on the paired *t*-test, but not by the resampling based methods. These genes exceeded a 2 fold expression level. It is, however, not straightforward to judge the relevance of these genes. Although not very consistent, their replicate measurements had the same tendency either being considerably over or under-expressed. Due to the restricted number of available measurements, the power of the *t*-test could have been too low to retain these genes. In contrast, the SAM method is less stringent because it makes no explicit assumptions on the H_0 distribution. Therefore these genes, though missed by the paired *t*-test, are still considered significant by SAM. Another interesting set of genes were those detected by both the paired *t*-test and resampling-based approaches. These genes are grouped in groups 11 and 5, and were only marginally but reliably down- or up-regulated (Table 3.3). It seems that these genes underwent subtle changes in expression level, barely exceeding what can be expected by coincidence (in contrast to the genes detected by the paired *t*-test alone) and are probably, from a biological point of view, most interesting.

The behaviour of the ANOVA model is illustrated by Figure 3.6, depicting group 9 (group 7 and group 8 suffer from the same problem but are not shown). The ANOVA-based bootstrapping approach assumes a constant

confidence interval identical for all genes. The size of the confidence interval is estimated based on a fixed residual distribution of the model fit on the complete dataset. As mentioned previously, if either one of the channels measures a signal close to zero (reflected by a low average expression level) the expression ratio (difference in *GC* effects) becomes an unreliable estimator of the differential expression. Indeed, in our data set, gene expression values close to zero in either one of the channels often resulted in relative high but inconsistent expression ratios (*p*-values of the paired *t*-test range from 0.04-0.55). For these genes, the constant confidence interval was a serious underestimation of the variation on these measurements. Groups 7, 8 and 9 (Table 3.3) are therefore most likely to contain predominately false positives. On the other hand, for genes that were only slightly differentially expressed, the constant confidence intervals based on the constant residual variance were probably too stringent to retain these genes. This resulted in a failure of the ANOVA based bootstrapping test to detect genes with more subtle alteration in expression level such as those present in group 11. Finally, in group 6 and 12, genes were grouped that were only detected by SAM (Table 3.3). These genes all behaved rather consistent (64% of the genes have *p*-value lower than 0.01, Table 3.3) and deserve further investigation.

3.4 Discussion

The goal of this chapter was to evaluate the use of ANOVA for preprocessing microarray data by identifying differentially expressed genes and comparing these to results from ratio based approaches. Apart from gaining insight into the workings of ANOVA models for normalizing microarray data (and identifying differential genes), a better understanding of the particular advantages and disadvantages of the different ratio based selection procedures resulted from this research.

With regards to the fold test, paired *t*-test, and SAM, the following conclusions could be made from our observations. Each of the methods differs in the required assumptions on the variance of the data and on the distribution of the residuals under the H_0 hypothesis. Therefore, the method for which the underlying assumptions are best satisfied will give the most reliable results, i.e. the reliability of the methods is dependent on the data set, as is often the case with statistical tests. The paired *t*-test could certainly be used as a more statistically founded alternative of the fold test. However, due to the few replicate measurements, it had the tendency to retrieve many consistently behaving ratio estimates seemingly too close to 0 to be called differentially expressed. Moreover, because of the restricted number of replicates, the paired *t*-test also has a rather low power. Of all methods

tested on our dataset, SAM clearly outperformed the other methods because the underlying assumptions were probably best satisfied.

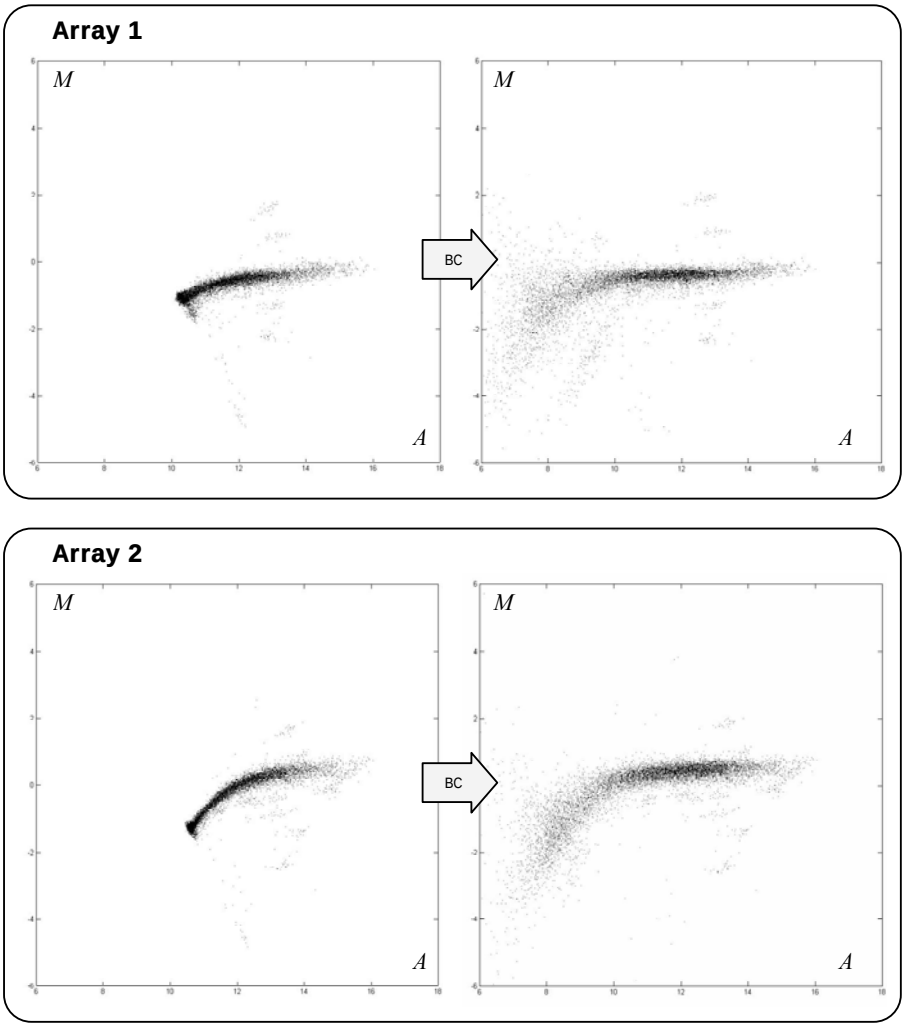


Figure 3.7: **Local background correction of the data.** Effect of performing a local background correction normalization on the data illustrated by MA-plots for both arrays. The upper two plots correspond to one array, the lower two correspond to the other. Raw intensities are shown in the left plots, background corrected intensities are shown in the right plots. Although apparently increasing the measurement range, performing a local background correction is the main reason for observing an increased error variance at lower intensity levels.

Several conclusions can be drawn with regards to ANOVA normalization, and the use of its residual distribution for identifying differentially expressed genes. The ANOVA based bootstrap method clearly underperformed in identifying differentially expressed genes. The assumption of a constant residual variance is obviously an oversimplification viewing the nonlinear trends in the data and the additive error in the low expression range. This oversimplification renders the use of the fitted residual distribution (with normal assumptions or by bootstrapping) for reliable identification of differentially expressed genes impossible. Performing different transformations prior to ANOVA could help to alleviate the problem of heteroscedastic residuals. In order to allow bootstrap analysis despite the unequal variance in residuals, Kerr *et al.,* 2001 proposed an adapted bootstrap procedure [109]. Instead of assuming a constant error for all measurements, the residual distribution was considered either gene-specific or at least intensity-specific. Another option could be to perform a non-local background correction or no background correction at all. At the time this research was conducted, local background correction was widely considered as an obligatory step in the normalization of microarray data. Since then, many objections have been uttered [32,66,82,115,135,201] (see also section 2.2.2.1 of chapter 2) and the matter is considered less clear cut. As shown in Figure 3.7, for the data set used, performing a local background correction, although apparently increasing the measurement range, is the main reason for observing an increased error variance at lower intensity levels. Neither of these approaches however, will work when heteroscedasticity is caused by a superposition of non-linear trends in the residuals for separate combinations of major effects (Figure 3.5 A, e.g. all genes measured with Cy5 on the second array). This was observed in our test examples, and can be attributed to non-linear dye discrepancies. These non-linear tendencies in the data prohibit the use of ANOVA for data normalization. Performing a LOWESS normalization (or a similar intensity based rescaling) prior to the application of an ANOVA model can alleviate such dye biases and can therefore be considered as a required step in any normalization procedure based on an ANOVA model.

From a theoretical point of view, ANOVA is nevertheless a powerful tool for preprocessing microarray data. The simultaneous use of all measurements, not only to normalize the data, but also to provide an estimate of the random experimental noise, can be considered as a major advantage. ANOVA models also provide a means to account for different sources of systematic variations representing the physical realities of a microarray experiment, contrary to the normalization strategies for log-ratios, which often adopt dubious assumptions (GNA) and may be described as ‘what we expect, is what you get’. More importantly, ANOVA models can deliver an estimate of absolute expression and take into account the specifications of each experimental setup. They are therefore better suited to analyze more

complex designs than any ratio based procedure. However, the ANOVA models presented in this chapter can not readily be applied to more complex designs, and extending these models is not a trivial matter. The next chapter will deal the problems and issues that arise when designing ANOVA models for complex microarray experiments.

Chapter 4

Generic ANOVA models

ANOVA models for microarray normalization can not readily be applied to any type of experimental setup of a microarray experiment. This chapter describes the issues that are encountered when attempting to fit published ANOVA models to different experimental designs (section 4.1), and the development of generic (applicable to any experimental setup) ANOVA models for microarray normalization (section 4.2). Section 4.3 is dedicated to the implementation of such a generic model in a user friendly web application.

4.1 ANOVA models and experiment design

In this section we will describe how the parameters of ANOVA models are estimated, as well as how these estimators are influenced by the experimental design of the study. These principles are illustrated by means of three simple, but conceptually different designs (see also chapter 2, section 2.2.2.2): a colour flip design, a reference design, and a loop design (Figure 4.1).

4.1.1 Colour flip

Reconsider model (3.3) outlined in chapter 3, section 3.1.2:

$$I_{ijklm} = \mu + G_i + C_j + A_k + D_l + R(G)_{m(i)} + GA_{ik} + GC_{ij} + \varepsilon_{ijklm} \quad (4.1)$$

Regardless of experiment design, each gene index *i* occurs *m(i)* times (as many as there are replicate spots on an array for that clone) with each combination of a condition, array, and dye, i.e. with each combination of (*j*, *k*, *l*). In the particular case of a two array colour flip design, specifying any

two of array, dye and condition automatically determines the third. With respect to the design variables array and dye, the layout of the tissue varieties forms a 2×2 *Latin square* [40]. A colour flip design is therefore often referred to as a Latin square design (illustrated in Table 4.1).

Table 4.1: Latin square structure of a colour flip design with respect to the main variables dye, array, and condition.

Dye	Array	
	1	2
Cy5	Condition 1	Condition 2
Cy3	Condition 2	Condition 1

For any ANOVA model, estimates of the different factor levels can be calculated in a least-squares sense (see also appendix B). Least-squares estimators (LSEs) minimize the sum of squares of model residuals. For model (4.1), the error sum of squares (*SSE*) is:

$$SSE = \sum_{ijklm} \left(e_{ijklm}\right)^2$$

$$SSE = \sum_{ijklm} \left(I_{ijklm} - \mu + G_i + C_j + A_k + D_l + R(G)_{m(i)} + GA_{ik} + GC_{ij}\right)^2 \tag{4.2}$$

Generally, to fit a linear model it is not necessary to derive the functional form of least-squares parameter estimates, because these parameters can be calculated by matrix inversion (see section B.2, appendix B). Due to the large number of parameters that need to be estimated in the case of microarray data however, this will be computationally infeasible for general matrix inversion programs. A minimum of the *SSE* is therefore calculated analytically by solving the set of normal equations (NEs), which are obtained by taking the partial derivatives of the *SSE* with respect to each the model parameters and setting them to zero:

$$\begin{cases} \dfrac{\partial SSE}{\partial G_i} = 0, \quad \dfrac{\partial SSE}{\partial C_j} = 0, \quad \dfrac{\partial SSE}{\partial A_k} = 0, \quad \dfrac{\partial SSE}{\partial D_l} = 0, \\ \dfrac{\partial SSE}{\partial R(G)_{m(i)}} = 0 \\ \dfrac{\partial SSE}{\partial GA_{ik}} = 0, \quad \dfrac{\partial SSE}{\partial GC_{ij}} = 0 \end{cases} \tag{4.3}$$

In order to solve this system of equations, several constraints need to be taken into account. Adhering to standard ANOVA conventions, for model (4.1) these constraints amount to:

$$\begin{cases} \sum_i G_i = 0, \quad \sum_j C_j = 0, \quad \sum_k A_k = 0, \quad \sum_l D_l = 0 \\ \forall i : \sum_m R_{m(i)} = 0 \\ \sum_i GA_{ik} = \sum_k GA_{ik} = 0, \quad \sum_i GC_{ij} = \sum_j GC_{ij} = 0 \end{cases} \tag{4.4}$$

Solving the set of equations given by (4.3) and (4.2), such that the constraints in (4.4) are respected, leads to parameter solutions for model (4.1) that can be written as:

$$\begin{aligned} \hat{\mu} &= I_{....} \\ \hat{G}_i &= I_{i...} - I_{....} \\ \hat{C}_j &= I_{.j..} - I_{....} \\ \hat{A}_k &= I_{..k.} - I_{....} \\ \hat{D}_l &= I_{...l} - I_{....} \\ \hat{R}(G)_{i(m)} &= I_{i...m} - I_{i...} \\ \hat{G}A_{ik} &= I_{i.k.} - I_{i...} - I_{..k.} + I_{....} \\ \hat{G}C_{ij} &= I_{ij..} - I_{i...} - I_{.j..} + I_{....} \end{aligned} \tag{4.5}$$

where a dotted index '.' indicates to average the logarithm transformed intensities over that index.

As illustrated above, certain factor terms can often be ignored when solving the NEs for a particular variable of the model, e.g. the expression for the LES of the *CG* effect does not depend on whether *GA* effects are included in the model because of the peculiarities of the Latin square design. This of course, is highly dependent on the experimental design of the study. If we assume that the replication variable in model (4.1), which is nested within the gene variable (as indicated in the model), does not interact with any of the other variables, meaning that it only serves to account for replicate measurements in the form of duplicate spots, then there are four main variables. Given these four main factors in the model, theoretically there are sixteen possible effects when we consider interactions of all orders. It turns

out that the Latin square design has a particularly neat structure. Each of these sixteen effects is completely *confounded* with one other effect, meaning one effect is estimable only assuming the other is zero. The pairs of confounded effects for these four main factors in the case of a colour flip design are shown in Table 4.2. Effects that are not completely confounded are called *orthogonal* in the Latin square. Orthogonality arises when a factor is completely balanced with respect to another factor. For example, if every condition in a microarray experiment appears in the design labelled with the red and green dyes equally often, condition is orthogonal to dye. One consequence of orthogonality is that the estimates of the two factors are uncorrelated. A second consequence is that including or excluding one effect in the model does not alter the estimates obtained for the other effect. In general, effects that are neither confounded nor orthogonal are said to be *partially confounded*. The main design problem with ANOVA models for normalizing microarray data should already be apparent now. Spotted microarray technology limits the number of conditions that can be measured on a single array to two. Array and condition will therefore hardly ever be completely orthogonal, as for most experimental designs it is impossible to measure every condition on each array. This becomes all the more clear when we consider other designs than a Latin square colour flip setup.

Table 4.2: Confounding structure of the colour flip design.

Confounding effects
$\mu \sim ADC$
$A \sim DC$
$D \sim AC$
$C \sim AD$
$G \sim ADCG$
$GC \sim ADG$
$GA \sim DCG$
$DG \sim ACG$

Note: This design partitions the sixteen experimental factor effects into eight pairs. The members of each pair are completely confounded, i.e. one member of a pair is estimable only by assuming the other is zero. This results in uncorrelated estimates for all effects not in the same pair. Model (4.1) includes an effect from every pair except the last. Thus it accounts for all data effects except DG and ACG, which are assume to be zero.

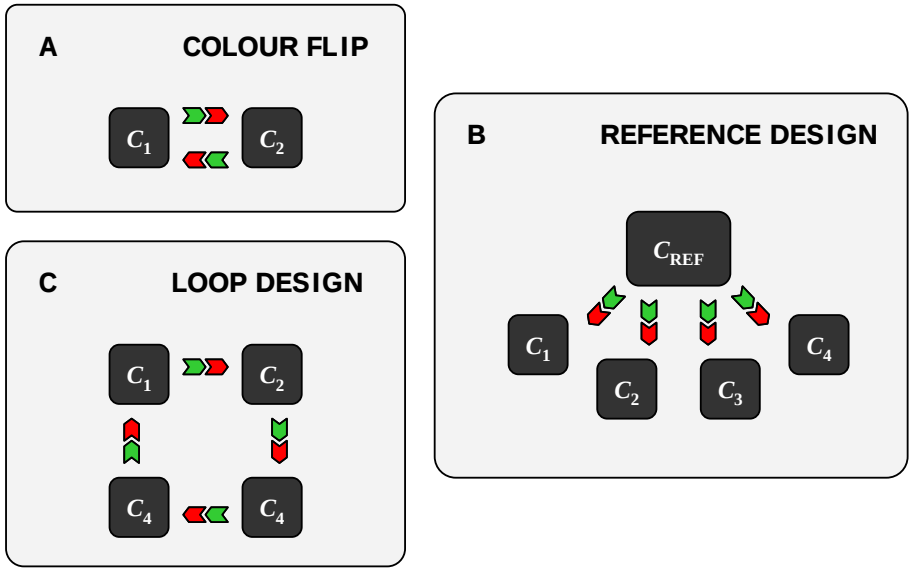


Figure 4.1: **Experimental design.** Schematical representation of the basic experimental designs that are discussed in relation to ANOVA model fitting in section 4.1: A) Colour flip design, B) Reference design (4 conditions and a reference), and C) Loop design (4 conditions). Black boxes represent the different biological conditions. Arrows represent the arrays on which indicated conditions are hybridized, either labeled in Cy5 (red part of arrow) or Cy3 (green part of array).

4.1.2 Reference design

A reference design (see also chapter 2, section 2.2.2.2) is used to compare multiple distinct biological conditions. As illustrated in Figure 4.1, these different test conditions are each paired with the same reference condition on separate arrays. One advantage of the reference design is that it is easily extendable. Additional varieties can be added to the experiment by adding another array on which a new test condition is compared to the reference, given that mRNA samples of this reference are still available. From an experimental point of view, another advantage is that each sample needs only to be labelled with one dye. The reference design however, is not without drawbacks. As pointed out earlier (chapter 2, section 2.2.2.2), more data are collected on the reference condition than any other, and this reference condition will generally be of least interest. Moreover, the choice of reference has a big impact on the quality of measurements of relative expression when working with log-ratios. Genes that demonstrate a low expression level in the reference condition (or no expression at all), will produce unreliable ratios or even missing values.

The latter is less of an issue when working with absolute intensities (e.g. fitting an ANOVA model), but other problems with this design become apparent when one considers model (4.1). First, conditions are completely confounded with dyes because each condition is labelled with only one dye. Thus, one cannot include both condition effects and dye effects in an ANOVA model meant to normalize reference designs. This in itself is not the biggest concern, because these main effects are not of actual interest and in practice, the dye effect will often be irrelevant, as a LOWESS normalization prior to fitting the ANOVA model is usually applied to remove intensity dependent dye biases. Taking in these considerations, a revised version of model (4.1) could be:

$$I_{ijklm} = \mu + G_i + C_j + A_k + R(G)_{m(i)} + GA_{ik} + GC_{ij} + \varepsilon_{ijklm} \qquad (4.6)$$

A more substantial problem with this model is the large cost in degrees of freedom that comes with the additional reference condition. The total amount of measured intensities from a reference design with c conditions (including the reference), g genes, and $r(i)$ spotted replicates for each gene i (so that s is the total number of spots on an array) is:

$$2(c-1)\sum_{i=1}^{g} r(i) = 2(c-1)s$$

The mean and the array, condition, and gene main effects together account for 2(c-1)+(g-1) degrees of freedom. The effects of interest, the *GC* effects, account for (c-1)(g-1) degrees of freedom. The spot effects represented by *GA* and *R(G)* effects, comprise (c-2)(g-1) and (s-g) degrees of freedom respectively. Depending on the number of replicate spots, the degrees of freedom available may be too few too reliably estimate the random error. In the common case were no replicate spots are available (and the *R(G)* effect can be omitted), no degrees of freedom remain to estimate the error term. Remark that this is always the case, irrespective of the number of replicates, when employing a slightly different model where the replicate variable is nested within the interaction variable array×gene (i.e. an *R*(*GA*) effect, see also chapter 3 section 3.1.2, model (3.2)). In such cases, at least one set of effects must be excluded to be able to estimate error and allow statistical inference. If we ignore spot effects all together, a yet further simplified model would be:

$$I_{ijklm} = \mu + G_i + C_j + A_k + GC_{ij} + \varepsilon_{ijklm} \qquad (4.7)$$

As is obvious from the discussion above, the confounding of effects in a reference design is more complex than for a Latin square design. There is no counterpart to the simple confounding structure presented in Table 4.2. As mentioned, conditions are completely confounded with dyes. In addition, since the conditions are not balanced with respect to the arrays, condition

main effects and array main effects are partially confounded (as are genexcondition interactions and genexarray interactions). When effects are partially confounded instead of completely confounded, it is possible to obtain separate estimates for each effect, at the cost of them being correlated. Generally, the estimators have a more complicated functional form because the effects must be 'disentangled'. This usually means less precise estimation, i.e. larger error bars. Failure to account for potentially important effects that are confounded or partially confounded with effects of interest can produce biases in the estimates of the latter. For instance, if there are no replicate measurements, it is impossible to obtain error terms for any of the measurements of the conditions of interest, since each gene is spotted only once on each array. As a result, *GC* effects for the conditions of interest may be biased, and all estimates of the random error term will stem from measurements of the reference condition, and the dye that corresponds to it.

4.1.3 Loop design

The loop design may be considered as an alternative to the reference design (see Figure 4.1). Using the same number of arrays as the reference design, the loop design collects twice as much data on the varieties of interest. Further, notice that for model (4.1) conditions are balanced with respect to the dyes because each condition is labelled once with both Cy5 and Cy3 dyes. This balance means that dye effects are not confounded with condition effects. If one estimates all factor main effects and, in addition, the *GC*, and *GA* and *R(G)* interactions, then at least (in the case of no replicates) g-1 degrees of freedom remain. These degrees of freedom provide information to estimate error variation. Therefore, contrary to the reference design, this design provides a basis for further statistical inference.

A practical drawback of the loop design is that each sample must be labelled with both the Cy5 and Cy3 dyes, effectively doubling the number of labelling reactions. Balancing condition with respect to dyes, and thus also genexcondition with respect to gene×dye, produces data in which gene×dye effects can be detected. According to Kerr *et al*., 2001 [111] this variable could prove very useful, as it would inhibit any anomalous behaviour of genes with respect to dyes biasing estimates of the effect of interest. One could argue however, that the extra degrees of freedom necessary to estimate such *GD* effects, and the fact that such observed anomalies may in fact be contributed to the intensity dependent bias between dyes (see chapter 2, section 2.2.2.3), may not be worthwhile the effort and may possibly lead to overfitting of the data.

From the discussion above, it is clear that an ANOVA model can not readily be applied to any type of experiment design. Careful consideration of the

peculiarities of the experimental design in relation to the retained variables in the model, and the resulting degrees of freedom that are available to estimate the random error, is necessary to assess the appropriateness of any ANOVA model. The matter is further complicated in that the three designs detailed above are hardly the only ones used in microarray experiments. In fact, they often serve as templates or building blocks for larger and more complex designs (e.g. a reference design extended with a colour flip for every array is not uncommon), so that the evaluation of ANOVA models for every other experimental setup becomes a tedious task.

4.2 Generic ANOVA models

As discussed in the previous section, the appropriateness of ANOVA models for normalizing microarray data is highly dependent on the design of the experiment. This is a major drawback for the routine application of such models in the analysis of microarray data. This section describes the construction of generic ANOVA models. These models attempt to balance different trade-offs, so that they can take into account the major sources of variation in a microarray experiment, and yet are able to normalize any type of experiment design. The primary characteristics of these models are discussed below.

The models are ***generic***, i.e. they can be applied to any type of experimental setup, there's no need for deriving different analytical solutions for specific experimental designs. The main problem in creating a generic model lays in the typical characteristics of spotted microarray technology, which limit the number of conditions that can be measured on a single array to two. For most experimental designs it is thus impossible to measure each condition on every array; array and condition will hardly ever be completely orthogonal. The condition variable was therefore abandoned altogether in favour of the array×dye variable. The *AD* factor is confounded with the *C* factor regardless of the experimental design (for a single array×dye combination, only one condition can be measured). As such, condition dependent variation in intensity (e.g. an mRNA sample of one condition may hold a larger quantity of mRNA than that of another condition, resulting in higher intensity measurements) can be accounted for in the estimates of *AD* effects. Using this approach, it will be difficult to determine the actual contribution of individual conditions to the total intensity variation. This is of little concern however, as the *C* effects are not of primary interest.

These days, microarray printing technology can produce arrays that can sport enough probes to represent an entire genome. Up until a few years ago, this was not always so. Very often multiple, different arrays had to be used in order to assay the expression of all genes represented in a clone set. The

incorporation of ***batch*** effects provides support for experiments that employ multiple arrays when the entire set of genes does not fit onto a single slide. As a result, all data points can be analyzed at once (different batches should otherwise require multiple analysis runs) and other parameters, such as dye effects, can be determined across all data points, instead of a batch-wise assessment.

Spotting effects are not modelled on a per spot basis. This will alleviate the overfitting observed with some of the previous models and, although it might increase the random error variance, it assures that regardless of experiment design, a sufficient number of degrees of freedom remain to estimate the random error term. As an alternative to spot effects, a ***pin-group*** effect is introduced into the models. The motivation for the inclusion of this factor can be found in Figure 4.2, which shows the spot placement of a typical microarray. Spots are grouped into smaller sub-grids, referred to as *pin-groups*. Due to the specifics of the spotting process, spots belonging to the same pin-group often share similar printing errors.

Two models were developed that satisfy these features. They differ in the way the pin-group variable is structured with respect to the batch and array variables. In a first model the pin-group factor P is assumed to be nested within batch only, and is thus assumed equal for all arrays belonging to a single batch (i.e. containing the same set of genes). The reasoning behind this choice of pin factor is that microarrays are often printed in series (see chapter 2, section 2.1.1.2), and when intensity variation can be attributed to a certain print group due to spotting errors, a similar variation is often observed for that print group on all arrays of that printing series. The model can be written as:

$$I_{ijklmn} = \mu + B_m + D_l + A_{k(m)} + AD_{kl(m)} + P_{n(m)} + G_{i(n(m))} + GC_{ij(n(m))} + \varepsilon_{ijklmn} \tag{4.8}$$

In this model, I_{ijklmn} denotes the log-transformed intensity of the measurement from the i^{th} gene, j^{th} condition, k^{th} array, l^{th} dye, m^{th} batch and n^{th} pin-group. As before, μ is the overall average signal, D_l represents the effect of the l^{th} dye, $A_{k(m)}$ represents the effect of the k^{th} array, $AD_{kl(m)}$ represents the interaction between the k^{th} array and the l^{th} dye, $P_{n(m)}$ represents the effect of the n^{th} pin-group, $G_{i(n(m))}$ represents the effect of the i^{th} gene, $(GC)_{ij(n(m))}$ represents the interaction between the i^{th} gene and the j^{th} condition, and ε_{ijklmn} represent the error terms which are assumed to be independent and identically distributed with mean 0. The brackets in the subscripts in dictate the nesting structure. The nesting of gene within batch and pin-group may seem somewhat artificial, especially when replicate spots are present, but the accompanying constraints assure an analytic solution for the parameters is always attainable.

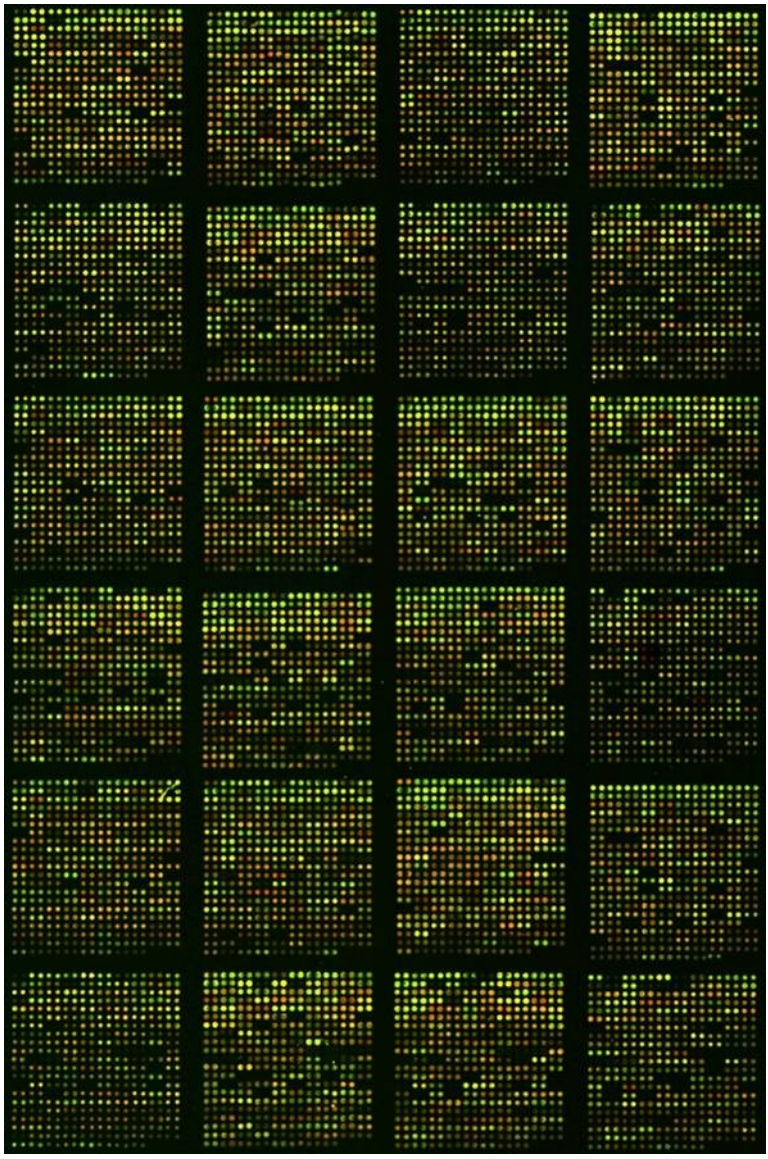


Figure 4.2: Pin-groups. A scanned picture of a hybridized microarray slide that clearly shows the layout of a typical spotted microarray. In this particular case, there are six rows and four columns of pin-groups, each consisting of 21 rows and 20 columns of spotted probes.

Using the constraints

$$\sum_m B_m = \sum_l D_l = 0$$
$$\forall m : \sum_k A_{k(m)} = \sum_k AD_{kl(m)} = \sum_l AD_{kl(m)} = \sum_n P_{n(m)} = 0$$
$$\forall m,n : \sum_i G_{i(n(m))} = \sum_i GC_{ij(n(m))} = \sum_j GC_{ij(n(m))} = 0 \qquad (4.9)$$

the parameter estimators can be written as

$$\hat{\mu} = I_{.....}$$
$$\hat{B}_m = I_{....m.} - I_{.....}$$
$$\hat{A}_{k(m)} = I_{..k.m.} - I_{....m.}$$
$$\hat{D}_l = I_{...l..} - I_{.....}$$
$$\hat{A}D_{kl(m)} = I_{..klm.} - I_{..k.m.} - I_{...l..} - I_{.....}$$
$$\hat{P}_{n(m)} = I_{....mn} - I_{....m.}$$
$$\hat{G}_{i(n(m))} = I_{i.....} - I_{.....} - \underset{i}{\text{avg}}\left[\hat{B}_m + \hat{P}_{n(m)}\right]$$
$$\hat{G}C_{ij(n(m))} = I_{ij....} - I_{i.....} - \underset{ij}{\text{avg}}\left[\hat{D}_l + \hat{A}_{k(m)} + \hat{A}D_{kl(m)}\right] \qquad (4.10)$$

Where 'avg' is the weighted average of the appropriate effects between brackets over all measurements of the subscripted index. For instance, $\underset{i}{\text{avg}}\left[\hat{B}_m + \hat{P}_{n(m)}\right]$ is the sum over all measurements of gene *i* of the corresponding *B* and *P* effects, divided by the number of measurements of gene *i*. The solutions above are the most general case, allowing for replicate spots across several pin-groups (i.e. duplicate spots elsewhere on the array) and even across batches. Although batches usually contain probes representing different genes, this is useful when for instance the same control probes are used on all batches. If all genes are represented only once (no replicate spots), the equations for *G* and *GC* effects simplify to:

$$\hat{G}_{i(n(m))} = I_{i...mn} - I_{....mn}$$
$$\hat{G}C_{ij(n(m))} = I_{ij..mn} - I_{i...mn} - \underset{ij}{\text{avg}}\left[\hat{D}_l + \hat{A}_{k(m)} + \hat{A}D_{kl(m)}\right] \qquad (4.11)$$

The second model (4.12) assumes a different pin-group effect for every pin-group and every array. The pin-group factor itself is omitted, and a *pin×array* interaction variable is introduced to cope with array specific pin-group effects. This is a more general approach that requires only a few more degrees of freedom than the previous model and is more appropriate when all the arrays in the experiment have not originated from the same printing series or assumptions with regards to constant pin-group effects within a particular printing series may not be valid for the experiment under study. Let $P_{n(m)}$ represents the interaction of the n^{th} pin-group and k^{th} array, this second model can then be written as:

$$I_{ijklmn} = \mu + B_m + D_l + A_{k(m)} + AD_{kl(m)} + PA_{nk(m)} + G_{i(n(m))} + GC_{ij} + \varepsilon_{ijklmn} \tag{4.12}$$

Using the constraints

$$\begin{aligned} &\sum_m B_m = \sum_l D_l = 0 \\ &\forall m : \sum_k A_{k(m)} = \sum_k AD_{kl(m)} = \sum_l AD_{kl(m)} = \sum_n PA_{nk(m)} = \sum_k PA_{nk(m)} = 0 \\ &\forall m,n : \sum_i G_{i(n(m))} = \sum_i GC_{ij(n(m))} = \sum_j GC_{ij(n(m))} = 0 \end{aligned} \tag{4.13}$$

parameter estimators for the batch, array, array×dye effects will be equal to those of the first generic model (4.8). The remaining estimators can be written as:

$$\begin{aligned} \hat{P}A_{nk(m)} &= I_{..k.mn} - I_{..k.m.} \\ \hat{G}_{i(n(m))} &= I_{i.....} - I_{.....} - \underset{i}{\text{avg}}\left[\hat{B}_m\right] \\ \hat{G}C_{ij(n(m))} &= I_{ij....} - I_{i.....} - \underset{ij}{\text{avg}}\left[\hat{D}_l + \hat{A}_{k(m)} + \hat{A}D_{kl(m)} + \hat{P}A_{nk(m)}\right] \end{aligned} \tag{4.14}$$

Again, the solutions above are the most general case, allowing for replicate spots across several pin-groups and batches. In the case that no replicate spots are present on the array, the equations for *G* and *GC* effects simplify to:

$$\begin{aligned} \hat{G}_{i(n(m))} &= I_{i...mn} - I_{....m.} \\ \hat{G}C_{ij(n(m))} &= I_{ij..mn} - I_{i...mn} - \underset{ij}{\text{avg}}\left[\hat{D}_l + \hat{A}_{k(m)} + \hat{A}D_{kl(m)}\right] \end{aligned} \tag{4.15}$$

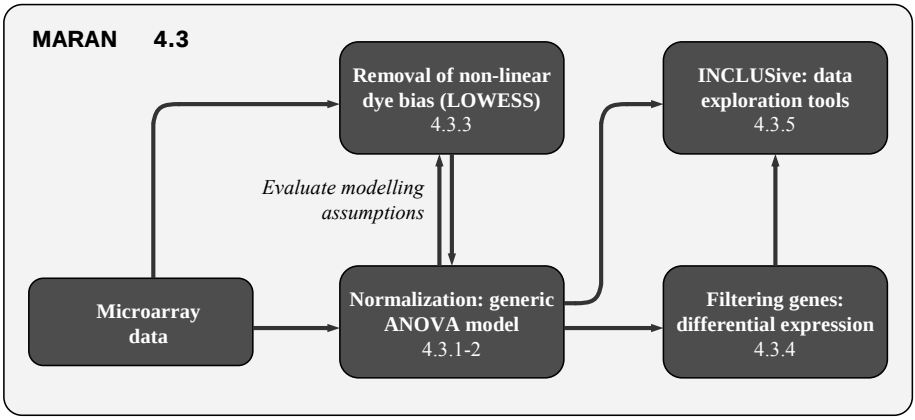


Figure 4.3: **Schematic representation of the MARAN web application.** After the data have been uploaded, they can be normalized by means of a generic ANOVA model, optionally with a preceding LOWESS step to remove nonlinear dye biases. The model and/or LOWESS procedure can be rerun at any time based on an evaluation of model fitting results (evaluation of modelling assumptions). A module for filtering the data and a module that integrates MARAN into INCLUSive (for e.g. clustering, motif detection), are also available.

4.3 MARAN: a web-application for normalizing microarray data

An implementation of model (4.12) was made publicly available (in cooperation with ir. B. Coessens) in the form of a user-friendly web-based application called MARAN [72] (http://www.esat.kuleuven.ac.be/maran). Model (4.12) was chosen over model (4.8) due to the more generic structuring of pin-group effects. The implementation of the generic model is embedded in a larger framework for the normalization of microarray data. Apart from an ANOVA normalization module, additional functionalities are made available to the user. A LOWESS fit procedure [226] is added as a remedial measure for non-linearities. An option for filtering the results by selecting genes with significantly changing expression profiles is also available. Preprocessing results can be sent from the MARAN website to the INCLUSive website [41,198] for further analysis, such as gene clustering and motif detection. After registering, a user can upload one or more data files and perform any of these analysis steps. The data and all result files and images generated during an analysis run can also be stored. A schematic overview of the different conceptual modules of the MARAN web application is given in Figure 4.3. Each of these components is discussed in greater detail in this section.

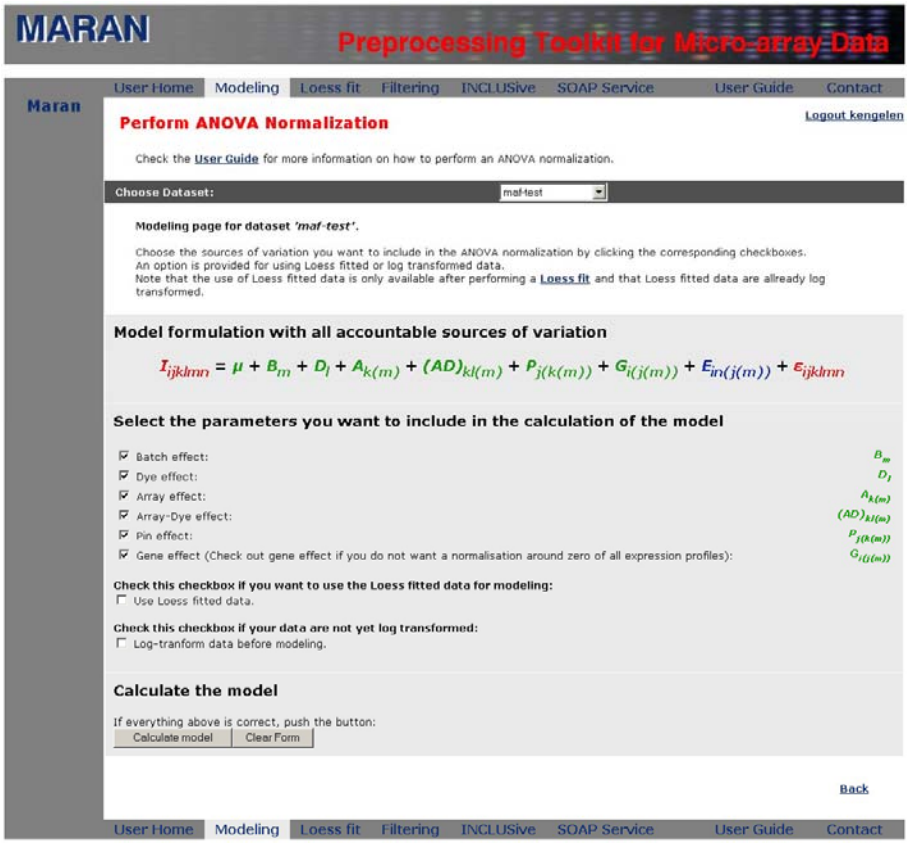


Figure 4.4: **The modelling page of the MARAN web application.** Factors that are not relevant to the experimental design are automatically greyed out. All other effects (except the effect of interest, *GC* effect), can be excluded by the user.

4.3.1 Modelling the data

Modelling the data is fairly straightforward and user-friendly. On the 'Modelling' page (depicted in Figure 4.4), a number of checkboxes represent the different sources of variation taken into account by the model. Depending on the specific design of the experiment, some of these checkboxes may be greyed out, i.e. any of these explanatory variables, that may not be relevant for a specific experimental design, are automatically discarded. For instance, when the total number of genes fits on one array, there will be only one batch, so the 'batch box' will be greyed out. All other

variables (except *GC* effects) can be included or excluded (depending on whether or not the user would like to incorporate the respective sources of variation in the model) by clicking the corresponding checkboxes. For instance, not checking the gene effect will not normalize the expression data with respect to there mean 'basal' expression level (i.e. the gene effect). In some cases, it may be useful to retain this information within the expression values.

On this page, there's also a checkbox for log-transforming the data. If the uploaded data is not log transformed when, we recommend checking this option. Indeed, our model assumes that an additive error (absolute error is independent of measured intensities) is present, while in most cases there's a pronounced multiplicative error (absolute error on the measurement increases with the measured intensity), so that modelling assumptions are not satisfied. Performing a logarithmic transforming the data (multiplicative errors become additive) is therefore often required [132] (see also chapter 2 section 2.2.2.2).

4.3.2 Interpretation of the results

After completion of the analysis, normalized expression values (and all parameters and residuals of the fitted model) can be downloaded from the 'Results' page (see Figure 4.5). An ANOVA table of the fit is shown to allow for interpreting the different effects and their contribution to the total amount of variation (represented by the 'SS' (Sum of Squares) column). Based on the ANOVA table, the normalization can be rerun, omitting effects that are shown to be of little or no significance to observed variation in intensity. Several plots for analyzing the ANOVA modelling assumptions are also included on this page. As explained in section 3.1.1 of chapter 3, these assumptions are twofold: firstly, the data should be adequately described by a linear model. Secondly, the error terms are assumed to be normally distributed with mean zero and constant variance. Information about the heteroscedasticity (non-constant error variance) and normality of the residual distribution can be obtained from the 'Global residual plot' and the 'NQ plot' (Normal Quantile plot of residual values) respectively. Serious heteroscedastic features should be avoided when using the residual distribution for selecting genes with significantly changing expression. It should be noted, however, that deviations from normality, in the form of widened tails, can often be acceptable due to the small amount of data points compared to the number of parameters to be estimated. As discussed in the section 3.1.3 of chapter 3, bootstrap methods are advisable for selecting genes with significantly changing expression when serious heteroscedasticity or non-normality occurs in the residual distribution.

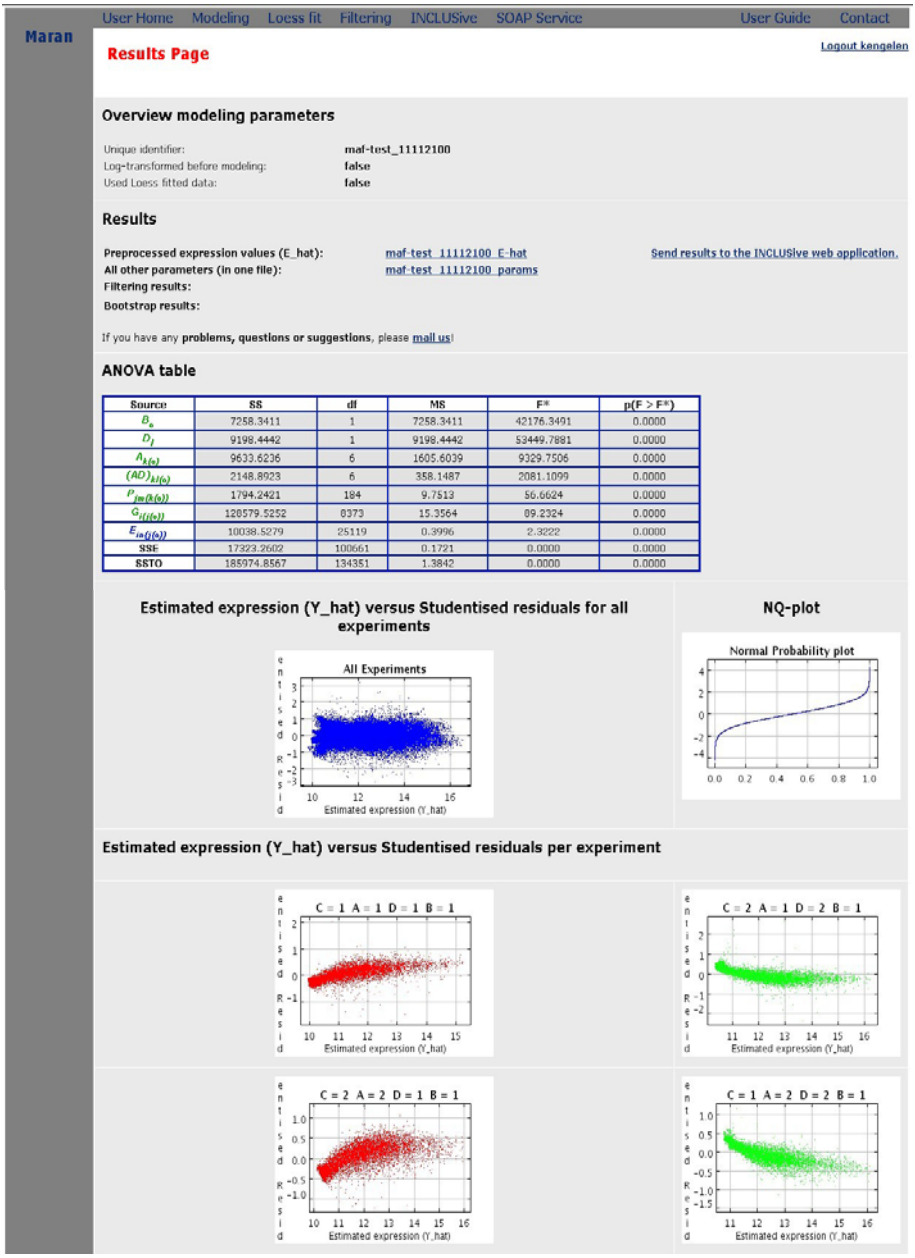


Figure 4.5: Result page of the MARAN web application. Part of the results page after analyzing a data set. All estimated parameters can be downloaded. The page also displays the ANOVA table of the fit and several plots for evaluating the modelling assumptions.

More problematic, however, is an apparent heteroscedasticity caused by a superposition of non-linear trends in the residuals for each combination of major effects, indicating that a linear model is not adequate for describing the data (i.e. the first assumption is not satisfied). The other plots on the 'Results' page are residual plots for each specific array×dye combination. When obvious curvilinear trends are observed on these plots, remedial measures should be taken, as described below.

4.3.3 Remedial measures for nonlinear dye bias

A well-established remedial measure for removing non-linear dye effects in the dataset is the LOWESS fit as described by Yang *et al.*, 2002 [226]. This intensity dependent rescaling (explained in greater detail in chapter 2, section 2.2.2.3) has been made available in the MARAN web application. The 'LOWESS' page can be accessed directly or after inspecting the results of an initial fit.

4.3.4 Filtering the results

After fitting an ANOVA model, the obtained estimates of the error terms can be used for various statistical analysis concerning the ANOVA parameters. Two different methods for selecting genes with significantly changing expression have been made available on the website. Both methods differ in the calculation of confidence intervals for the *GC* effects (the parameters of interest, i.e. the condition-affected change in intensity for each gene), as derived from fitting the ANOVA model. The statistical test for selecting genes with significantly changing expression, based on these confidence intervals, is identical for both methods. For each gene i:

$$\begin{aligned} &H_0: GC_{i1} = GC_{i2} = \ldots = \delta_i \\ &H_a: \text{at least one } GC_{ij} \neq \delta_i \end{aligned} \qquad (4.16)$$

Where δ_i is an undefined value. Basically, this test evaluates, for every single gene, whether a single expression level (i.e. δ_i) exists that could account for each calculated GC_{ij} (j=1,…,c; c being the total number of conditions in the experiment) effect of that gene (indicating that this gene is not differentially expressed).

The first method is valid under the assumption of normally distributed error terms, with mean zero and constant error variance. The null hypothesis for selecting differentially expressed genes is that all calculated *GC* effects for a single gene are sampling instances from a normal distribution (based on rescaled residuals) around a non-specified 'expression' value. Correction for

multiple testing is done by using the Bonferroni correction procedure (see chapter 4 of Neter *et al.* [141]). This correction is done depending on the number of conditions that are present in the experiment, not according to the number of genes (a way too restrictive measure as discussed in chapter 3, section 3.2.2). A selection of genes can be obtained by entering a preferred significance or, when desired, *p*-values for all genes can be downloaded. Although this method should not be applied when there is doubt that the ANOVA assumptions are satisfied, this method is relatively fast and may therefore serve as a preliminary indication of differentially expressed genes.

The alternative method is based on a bootstrap procedure [50,67,68]. It is a *fixed predictor sampling* method, similar to the one described by Kerr *et al*., 2000 [113] and is appropriate when the residuals show serious deviations from normality, but no apparent heteroscedasticity is present. A selection of genes can be obtained by entering a preferred significance. It is not possible however, to obtain p-values for each gene, contrary to the method described above.

4.3.5 Further analysis

MARAN is integrated in INCLUSive [41,198] (created by ir. B. Coessens and Dr. G. Thijs), a web-based suite of algorithms and tools for the analysis of gene expression data and the discovery of *cis*-regulatory sequence elements (http://www.esat.kuleuven.ac.be/inclusive). The complete results file (i.e. all of the estimated *GC* effects), or a selection of genes obtained after filtering, can be used for further analysis with any of the other tools implemented in INCLUSive. These allow for clustering of microarray data (Adaptive Quality Based Clustering or AQBC [45]), functional scoring of gene clusters (based on Gene Onthology (GO) terms [7]), sequence retrieval at a myriad of different public, and detection of known (MotifSampler [195-197]) and unknown (MotifScanner [2]) regulatory elements using probabilistic sequence models and Gibbs sampling. All tools (including MARAN) are available via different web pages and also as web services; several tools are also available through Toucan [2], a stand-alone application for the detection of *cis*-regulatory elements in promoter regions of higher eukaryotes. The web pages of INCLUSive are connected and integrated to reflect a methodology and facilitate complex analysis using different tools (see Figure 4.6).

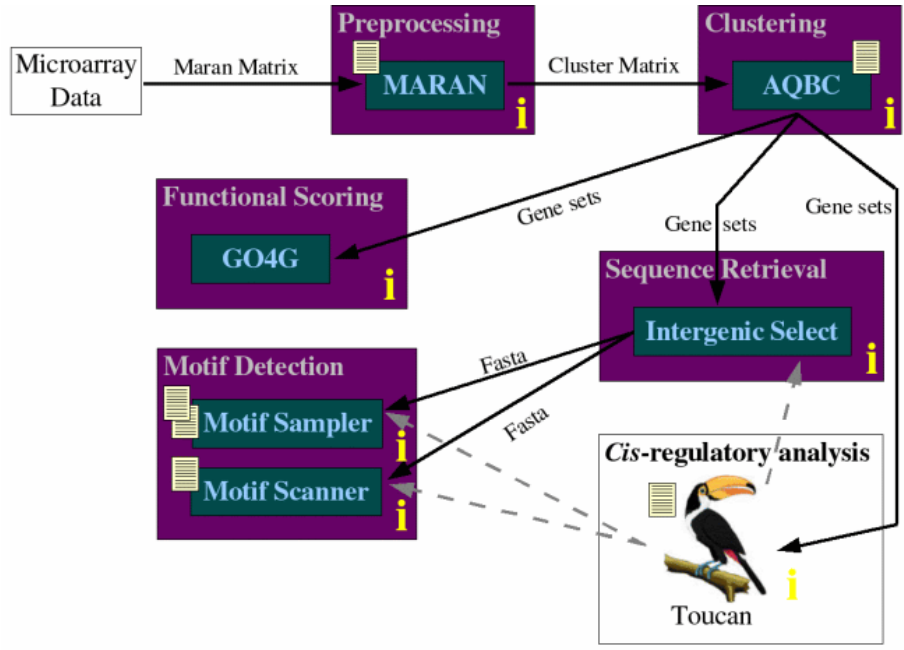


Figure 4.6: **INCLUSive.** Schematic overview of the data flow between the different modules of INCLUSive. The flow supports complex analysis of microarray data, comprising ANOVA normalization, filtering and clustering, functional scoring of gene clusters, sequence retrieval, and detection of known and unknown regulatory elements. All modules are independent of each other and can be used separately. Taken from http://www.esat.kuleuven.be/inclusive.

4.4 Conclusions

In this chapter we have described the development of two generic ANOVA models for normalizing microarray data. The major advantage of these models is that they are independent of experimental design and can readily be applied to any type of experimental setup; there is no need for deriving different analytical solutions, or redesigning the model for different experiments. One of these models was made publicly available in the form of a web-based application for normalizing microarray data, dubbed MARAN.

The benefits of these generic ANOVA models are evident. There are however, some considerations to the general use of ANOVA for normalizing microarray data, which are outlined here in greater detail. In section 4.4.1, we discuss the importance of replicate measurements (in fact a major issue

regardless of the chosen normalization method) and the implications on ANOVA based microarray analysis. In section 4.4.2, we discuss some strange and important complications of the LOWESS normalization, which only became discernable after using ANOVA in combination with a LOWESS fit procedure in the analysis of large scale experimental designs.

4.4.1 Experimental design limitations

Using the generic models, it is technically possible to normalize any type of experimental design. This does not imply however, that the quality of the normalization results is independent of the experimental design. In general, the more replicates are measured for each gene×condition combination, the better the estimated ANOVA parameters will be (which is not necessarily expressed through a low error sum of squares). This follows naturally from the fact that each *GC* effect (the effect of interest, the condition affected change in intensity for each gene) is calculated from all measurements of a single gene×condition combination (after being corrected for the other experimental variations included in the model). As a rule of thumb, we would suggest that each gene×condition combination is measured at least twice as to ensure a residual is obtained for each measurement (this is especially important when the residuals are later to be used as a statistical measure for e.g. selecting genes with significant change in expression). There are no strict regulations as to how these replications should be incorporated in the experimental design. However, in order to avoid partial confounding of effects, it may be wise to ensure that experimental sources of variation are different from one replicate to the next. For instance, a colour flip experiment may be more informative than simply repeating all measurements on a different array (same dye) or multiple spotting on the same array (same array, same dye), and it could help account for gene specific dye biases [14,51-55,112,133,134,139,182,193,203,227] should they occur.

To illustrate how the lack of replicates can lead to bad error estimation (due to the lack of residuals), the data set of Spellman *et al*., 1998 [187] was normalized with MARAN. This experiment is a reference design. Eighteen arrays were used to test eighteen time points of the yeast cell-cycle (labelled in Cy5); each array used the same reference condition (labelled in Cy3). Since there was no multiple spotting, all genexcondition combinations measured with Cy5 were only measured once, while the combination of each gene with the nineteenth condition (i.e. the control condition, always measured in Cy3) were measured seven times each (once on each array). This results in partial residual plots for each array as the one illustrated in Figure 4.7. There are no residuals for any of the measurements in Cy5 (red). Valuable biological interpretations may still be obtained from analyzing the

expression effects (after all, they are normalized with respect to various experimental sources of variation), but using the obtained residuals for further statistical inference may prove detrimental.

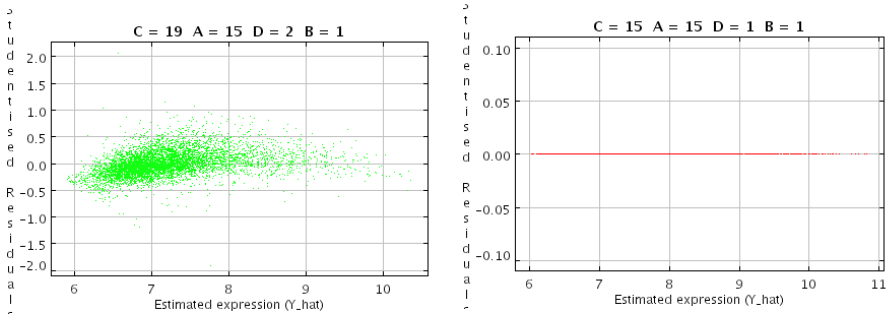


Figure 4.7: Experimental design limitations. Detailed residual plots (MARAN output; estimated intensity in X-axis, residuals in Y-axis) for the measurements of the fifteenth array after analyzing the data set of Spellman *et al.*, 1998 [187]. The experiment is a reference design with 18 different conditions (time points in the *Saccharomyces cerevisiae* cell cycle). *C* indicates condition number (19 being the reference condition), *A* indicates array number, *D* indicates dye number (Cy5 = 1 and Cy3 = 2 by default) and *B* indicates batch number. Due to the characteristics of the reference design, no replicates are available for the conditions of interest, and as a result, no residuals are obtained for any of the Cy5 (red) intensities.

4.4.2 Persistent non-linearities

One of the major problems with ANOVA normalization models (or any linear model for that matter), is that they are unable to cope with typical non-linear dye biases in microarray data (see chapter 2, section 2.2.2.3). Thus, modelling assumptions are seldom satisfied and inferences regarding the normalization parameters (e.g. selecting genes with significantly changing expression), which are based on these residuals, are unreliable. Remedial nonlinear measures, such as performing array-by-array LOWESS fits prior to the linear normalization, should alleviate these nonlinear dye discrepancies (e.g. in chapter 3, section 3.3.2 LOWESS normalized data did not show any nonlinear tendencies in the residuals of the ANOVA model fit). Indeed, a LOWESS normalization is usually performed according to the principles dictated by the GNA (Global Normalization Assumption, see chapter 2, section 2.2.2.3). The GNA assumes that only a limited number of genes on the array alter their expression, and that there is symmetry in the amount of up-regulated versus down-regulated genes. LOWESS normalized

intensities should therefore be devoid of *any* systematic variation between both samples hybridized to the microarray (reflected in log-ratios that are evenly distributed around zero across the entire intensity range), regardless of whether they can be attributed to actual dye effects or other factors such as mRNA quantity and quality, or even biological characteristics that do not adhere to the GNA (e.g. the bulk of genes being up-regulated). Some interesting features were revealed however, when using ANOVA in combination with a LOWESS fit procedure for the analysis of large scale experimental designs, indicating that a LOWESS normalization may not be able to completely alleviate intensity dependent nonlinear tendencies in the data, despite of its harsh assumptions with regards to the distribution of gene expression from one biological condition to the next (i.e. the GNA).

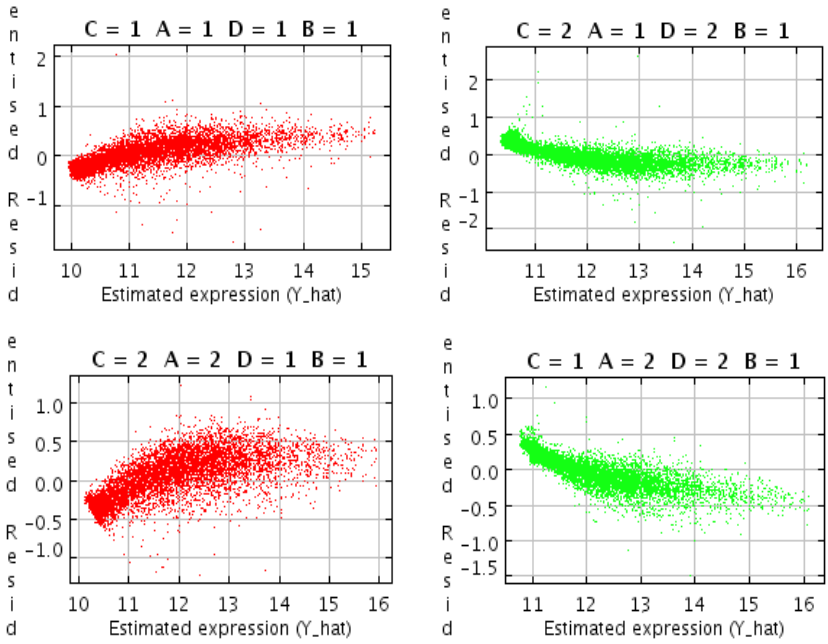


Figure 4.8: **Non-linearities after ANOVA normalization.** **T**he residual plots (MARAN output; estimated intensity in X-axis, residuals in Y-axis) for some of the array and dye combinations in the vitamin D_3 experiment of Verlinden *et al.*, 2005 [213] for the data that was not LOWESS normalized. The figure clearly shows the nonlinear tendencies and their relation with the dye variable, i.e. Cy5 (red) measurements tend to be progressively more overestimated at lower intensities, while Cy3 (green) measurements tend to be more underestimated.

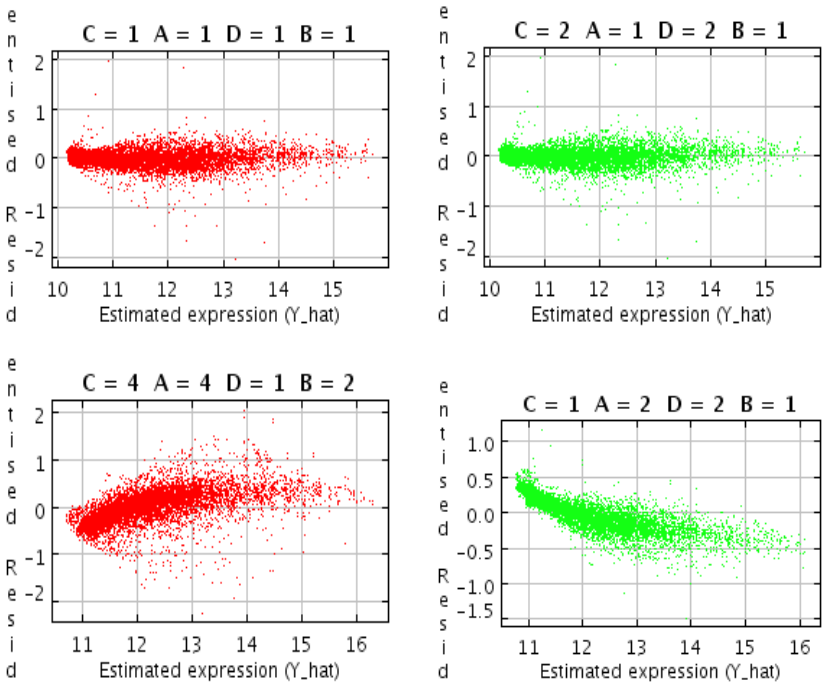


Figure 4.9: **Persistent non-linearities after LOWESS and ANOVA normalization.** The residual plots (MARAN output; estimated intensity in X-axis, residuals in Y-axis) for some of the array and dye combinations (same as in Figure 4.8) in the vitamin D_3 experiment of Verlinden *et al.*, 2005 [213] for the LOWESS normalized data. This figure illustrates how a LOWESS fit can remove this characteristic dye bias for some arrays (upper two plots), but is not capable of removing all the systematic nonlinear variation in the data, as illustrated by the lower two plots.

We will illustrate our general findings with the data set of Verlinden *et al.*, 2005 [213]. This microarray experiment was a study of the potent anti-proliferative effects of 1α,25-dihydroxyvitamin D_3 ($1,25(OH)_2D_3$, the active metabolite of vitamin D_3) that coincide with a hampered G1/S transition of the cell cycle. cDNA microarrays were used to monitor gene expression in MC3T3-E1 mouse osteoblasts at 1, 6, 12, 24 and 36 h after treatment with 1,25(OH)2D3 and compared to non-treated MC3T3-E1 cell lines at matched time points. Hybridizations were performed by pairing the treated samples and control samples for every time point, replicated with a colour flip hybridization. The total mouse clone set consisted of 21,492 cDNA fragments, and was spread across five different slides. The design of this experiment thus consisted of ten different biological conditions, measured on a series of colour flip designs, and with multiple batches. Measurements

were LOWESS normalized (smoothing factor *f* set to 30%) for each array. Both LOWESS normalized and original data from the entire experiment were fitted to ANOVA model (4.12). Figure 4.8 shows the residual plots for some of the array×dye combinations in the experiment for the data that was not LOWESS normalized. Figure 4.9 shows the residual plots for the same array×dye combinations in the experiment for the LOWESS normalized data. Figure 4.8 clearly shows the nonlinear tendencies and their relation with the dye variable (i.e. Cy5 measurements tend to be progressively more underestimated at lower intensities, while Cy3 measurements tend to be more overestimated). As shown in Figure 4.9, a LOWESS fit can remove this characteristic dye bias, but is not capable of removing all the systematic nonlinear variation in the data. This is particularly puzzling as a LOWESS fit is essentially designed to remove all of the non-linear dye bias that can be observed in an MA-plot.

An explanation for this seemingly implausible phenomenon may be found in a saturation of the lower intensity levels. It is generally assumed that the measured intensity of a signal (whether absorption, emission or fluorescence) is directly proportional to the concentration of the compound responsible for this signal (Law of Lambert-Beer [183]). In reality however, this law only holds for a certain intensity or concentration range: saturation occurs in the higher and lower regions so that the actual relationship between intensity and concentration is more like a sigmoid (logarithmic scale). This is arguably true for the fluorescence measurements of microarrays, where scanner characteristics, background signals, and quenching [160] can cause intensity saturation to occur. So for microarrays we may assume two distinct saturation curves (one for Cy5, another for Cy3) that describe the relationship between measured intensity and concentration of fluorescent dye (an indication of the amount of hybridized target). The distance between both curves would then coincide with the divergence, as generally observed in an MA-plot, of the point cloud with respect to the axis of zero log-ratios. This is conceptually illustrated in the left hand panels of Figure 4.10. Intensity dependent normalization methods, such as LOWESS, will merely remove the nonlinearities between the Cy3 and Cy5 intensity measurements, and not between the measured intensity and the dye/target concentration. After performing such a procedure, one is left with new intensities that still show a nonlinear relation to the corresponding concentration (right hand panels of Figure 4.10). These remaining artefacts become apparent when fitting a linear ANOVA model, which takes in account different sources of variation across the entire experiment, to the LOWESS normalized data of complex experiment designs. Indeed, when multiple conditions are measured across multiple arrays, such a model assumes linear relationships between absolute intensities measured for the same biological condition. If only the nonlinear difference between Cy5 and Cy3 intensities is removed, as with a LOWESS procedure, residuals of the

model fit will still show pronounced nonlinear trends that would otherwise remain hidden through the use of log-ratios. The next chapter will delve further into these observations, and will deal with the construction of a normalization method that will better acknowledge the particular nonlinear characteristics of the relation between measured intensity and target concentration.

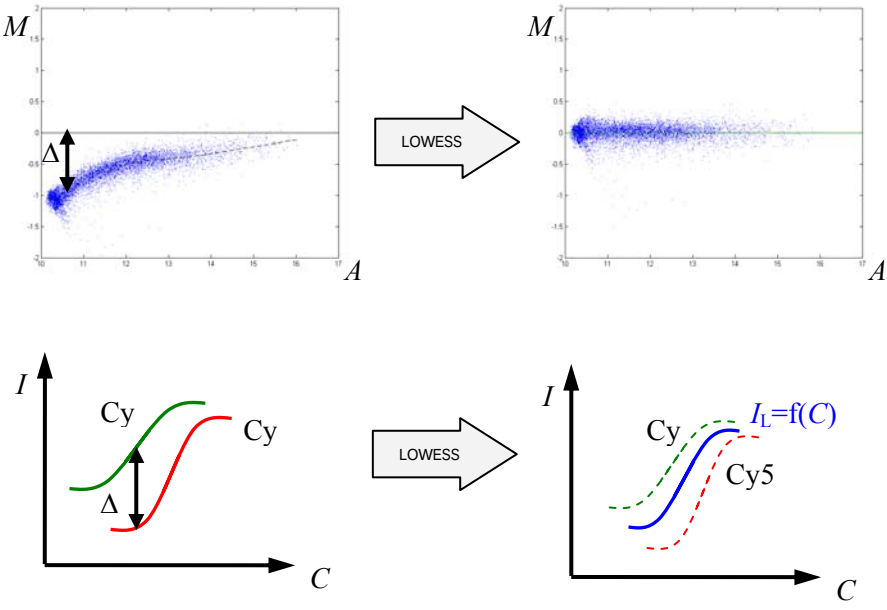


Figure 4.10: An explanation for persistent non-linearities. Two distinct saturation curves (one for Cy5, another for Cy3) describe the relationship between measured intensity (I) and concentration of fluorescent dye (C; an indication of the amount of hybridized target). The distance between both curves would then coincide with the divergence, as generally observed in an MA-plot, of the point cloud with respect to the axis of zero log-ratios (left hand panels). Intensity dependent normalization methods, such as LOWESS, will merely remove the nonlinearities between the Cy3 and Cy5 intensity measurements, and not between the measured intensity and the concentration of labelled target. After performing such a procedure, one is left with new intensities (I_L) that still show a nonlinear relation to the corresponding concentration (right hand panels; indicated by the blue curve in the lower right hand panel).

Chapter 5

A calibration procedure for spotted microarrays

The normalization method we propose in this chapter differs in spirit from previously published normalization strategies [73]. It relies heavily on the intensity measurements of external control spikes (RNA transcripts that are added to the hybridization solution in known concentrations) and is based on a physically motivated calibration model. It is nevertheless a continuation of the research described in the previous chapters as the basic idea remains unchanged. The measured intensities are to be modelled as functions of systematic sources of variation in a physically and experimentally meaningful way, and should allow for the calculation of an absolute value of expression, instead of being limited to the relative nature of intensity ratios. External control spikes turned out to be an essential asset in this respect; a detailed description of their nature and of the insights that can be gained from their employment is given in section 5.1.

The calibration model that is the core of this normalization procedure is presented in detail in section 5.2.1. The model consists of two major components, describing the hybridization of target transcripts to their corresponding spotted probes on the one hand (section 5.2.1.1), and the measurement of fluorescence from the hybridized, labelled target on the other hand (section 5.2.1.2). The parameters of this model and their error distributions are estimated from external control spikes (section 5.2.2), and are used to obtain absolute expression levels for every gene in every biological conditions present in he experiment (section 5.2.3).

Results that were obtained from applying our method to a publicly available data set are discussed in section 5.3. We show that the procedure is capable of adequately removing the typical non-linearities of microarray data, without making any assumptions on the distribution of differences in gene expression from one biological sample to the next (section 5.3.2), and compare our method to results obtained from normalizing the data with a

standard LOWESS procedure prior to fitting an ANOVA model (section 5.3.3). Since our model links target concentration to measured intensity, we further demonstrate how absolute expression values of transcripts in the hybridization solution can be estimated (section 5.3.4 and 5.3.5). Finally, we illustrate the effect of local background correction and the models capacity to deal with negative (background corrected) intensity values (section 5.3.6).

5.1 External control spikes

In the previous chapter, we noted problems with the LOWESS fit procedure in completely removing systematic non-linear trends in microarray data. It was speculated that this problem originated from saturation of the lower intensities, the characteristics of which are determined by the specific dye, and presumably for the same dye by different arrays as well (due to differences in hybridization and labelling reactions). It is impossible to evaluate any hypothesis regarding the relation between concentration and measured intensity based on measurements from biological samples alone, as true concentrations of target in the hybridization solution are unknown. Experimental controls are required in order to extensively address this issue. Analysis of the data obtained from such controls should provide information on the dynamic range, sensitivity, and specificity of the hybridization, and should grant an insight in the reproducibility of the observed expression ratios. Several types of controls can be used for monitoring a microarray experiment, and they can be conceptually subdivided into positive, negative, or spiked controls.

- ***Positive controls*** are designed to verify that the targets are labelled to an acceptable specific activity by Cy3 and Cy5. The corresponding target is added to samples prior to labelling, so that each target population (i.e. Cy3 and Cy5) should generate signals of approximately equal intensity after hybridizing to a positive control element. Often dilutions of the spotting probe solutions series are generated to address the signal strength for a wide range of probe amount.
- ***Negative controls*** are used to assess background signals and the degree of non-specific hybridization. Typically, negative controls are segments of coding DNA derived from organisms have no known (or expected) homologues or paralogues in the species under study, but share approximately the same content of guanine and cytosine as the studied species. For instance, microarrays used in expression analysis of mammalian genes often include negative controls composed of a combination of plant or bacterial coding sequences.

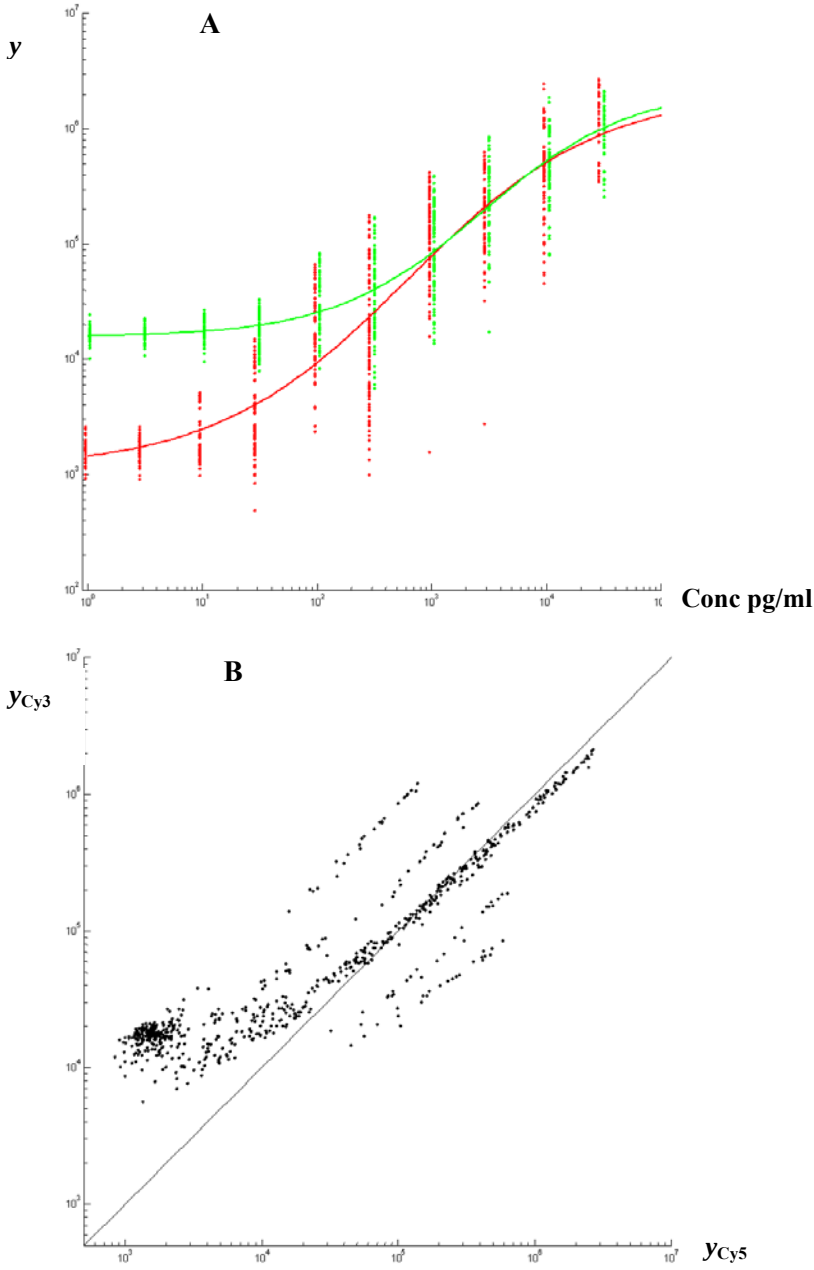


Figure 5.1: **External control spikes.** A) Non-linear relationship (saturation) between measured intensity *y* and corresponding concentrations for all external control spikes with a Cy5:Cy3 ratio of 1:1. B) Measured intensities of all external control spikes (Cy5:Cy3 ratios 1:10, 1:3, 1:1, 3:1 and 10:1). This plot illustrates the relatively small scanner errors, especially compared to the large variation in intensities that is observed in panel A.

- ***External control spikes*** (also *spiked controls*, or simply *spikes*) are made up of sequences that are chosen to not hybridize with any transcripts known to be expressed in the organism under study. They are added at a range of concentrations to Cy3 and Cy5 probes (i.e. a ratio 1:1) to provide data on the dynamic range of the system. Array elements corresponding to these controls should hybridize the Cy3 and Cy5 targets with equal intensity across a range of concentrations. A complementary method is to vary systematically the amount of spiking target added to the two labelling reactions (e.g. in ratios of 1:10, 1:3, 1:1, 3:1 and 10:1) and then to compare the observed signal ratios in the Cy5:Cy3 channels with those predicted from the stoechiometry.

Other types of experimental controls, such as housekeeping genes (genes that are assumed to be constantly expressed), spotted clone pools or spotted genomic DNA, have also been proposed (for an overview, see Kroll and Wölfl, 2002 [118]). However, none of these are able to assess the dynamic concentration range of the intensity measurements, and are thus not appropriate for the research questions at hand.

Several companies sell kits containing DNA samples for spotting, together with matched spike mixes, that can be used to validate and monitor the performance of microarrays (e.g. Lucidea Universal ScoreCard by Amersham, or SpotReport by Stratagene Inc.). Figure 5.1, panel A displays the relation between the measured intensities of a series of external control spikes to their actual concentration in the hybridization solution. The plotted measurements correspond to spikes of a single array which belonged to a data set of 14 arrays (see section 5.3.1), all complemented with the Lucidea Universal ScoreCard (Amersham) external control spikes. This plot confirms the expected behaviour of the intensity saturation characteristics (see 4.4.2). Not only is the saturation clearly present at lower (and higher) intensity levels, but there is a recognizable distinction between the Cy5 and Cy3 intensity signals of the labelled targets. Moreover, this difference between both dyes varies somewhat across arrays (data not shown).

Another prominent feature that can be observed in panel A of Figure 5.1 is large variation in intensity for a single spike concentration. The level of variation seen in this plot is especially remarkable in view of the relatively small scanner errors, as illustrated in panel B of Figure 1. Panel B plots the Cy3 versus Cy5 spike intensities, and shows that ratios of these controls seem highly conserved, in particular at upper intensity levels, an indication of fairly accurate scanner equipment. This apparent contradiction can only be explained by assuming a per spot correlation between the Cy3 and Cy5 intensities, i.e. the large intensity variation observed in panel A is directly caused by variation associated with the printed probes, not the scanner or amount of target in the hybridization solution (previous publications have

already shown that spot related errors have a large effect on the final observed signal [165], and that the influence of the scanner equipment is rather low [161]). Whether the main source of this spot related variation can be attributed to the actual amount of deposited probe DNA, or to a measure of spot quality (e.g. probe density [150], cDNA probe length [189], etc.), the implications are equivalent. Imperfections in the spotting process result in heterogeneous 'spot capacities', in terms of the available quantity of spotted probe, and allow distinct spots to bind different amounts of target from the hybridization solution.

Apart from providing insight into the dynamic range and intensity variation of microarray measurements, there are other advantages to the employment of external control spikes in microarray analysis. They can serve as a useful means of avoiding the Global Normalization Assumption (GNA, see section 2.2.2.3 of chapter 2) in the normalization step of the analysis. Normalization algorithms that do not require this GNA have already been proposed [217,235], but they are bound to other presumptions on the behaviour of gene expression values. A more reliable strategy to avoid making any assumptions regarding the distribution of gene expression is to use external control spikes to estimate normalization parameters. It is important to note that, contrary to external control spikes, none of the other types of experimental normalization controls, such as housekeeping genes, spotted clone pools or spotted genomic DNA (for an overview, see Kroll and Wölfl, 2002 [118]), are able to compensate for unbalanced gene expression changes. Moreover, by using external control spikes, it has been shown that global mRNA changes, resulting in an uneven distribution of expression changes, occur more frequently than what was previously believed [206,208], and that these changes can have a significant impact on the interpretation of data normalized according to the Global Normalization Assumption [158].

In light of the points discussed above, it should come as no surprise that the normalization method described in this chapter relies heavily on measurements of external control spikes. Details of the mathematical models and normalization algorithms are given in the following section 5.2.

5.2 Mathematical models and algorithms

External control spikes have previously been employed for quality control and normalization [10,16,70,80,99,158,208,218], but have seldom [33] been exploited to their full potential. In fact, spikes are genuine calibration points, in that they relate the measured intensity to the actual target concentration in the hybridization solution. Using these calibration points to estimate absolute expression levels instead of expression ratios could greatly simplify

inter platform comparisons and the analysis of large, complex designs comparing multiple biological conditions. However, the large variation in measured intensity, caused by –unknown- spot related errors, prohibits the direct correlation of a measured intensity to a concentration of target in the hybridization solution. The use of external control spikes to estimate absolute expression levels would therefore benefit from a more elaborate normalization procedure.

The proposed normalization procedure itself is straightforward in principle: intensity measurements of external control spikes serve to estimate the parameters of a calibration model. These parameters can then be used to obtain absolute expression levels for every gene in each of the tested biological conditions. The calibration model consists of two components, a hybridization reaction and a dye saturation function. In section 5.2.1a more detailed description of this model is given, along with its corresponding parameters and error distributions. The parameter estimation and normalization procedure are outlined in section 5.2.2 and section 5.2.3 respectively.

5.2.1 A model for microarray intensity measurements

5.2.1.1 Hybridization reaction

To explain these large variations of absolute intensities observed for a single spike concentration, a hybridization component was included in our model to account for spot capacity errors. The relation between the amount of hybridized target (x_s) and the concentration of the corresponding transcript in the hybridization solution (x_0) is modelled by the steady state of the following reaction:

$$X_0 + S \overset{K_A}{\Leftrightarrow} X_S \tag{5.1}$$

For this model, the hybridization constant K_A is assumed to be equal for all spots on a single microarray. Differences in hybridization constants should therefore be interpreted as variations caused by microarray related factors such as temperature, salt concentrations, hybridization time, etc., but do not account for gene specific hybridization efficiencies.

The amount of probe in a spot will often be vastly in excess of the amount of its in target in solution. The hybridization reaction itself is diffuse limited, and stirring has a large effect on overall rate [180]. A practical consequence is that when not properly stirred (e.g. when applied concentrated in a thin film) the solution of target immediately above a spot is rapidly exhausted of

target molecules complementary to the probes in the spot. When stirred properly, the concentration of target across the entire slide will gradually decline until the reaction reaches a steady state. A second assumption underlying our model is that target concentration x_0 is in excess (i.e. x_0 can be considered constant). It may seem somewhat contradictory to the explanation above, but it is a mathematical simplification that ensures that the amount of hybridized target at the end of the reaction depends only on the initial concentration in the hybridization solution. In fact, it is a first order approximation of the actual reaction, in which case the target concentration x_0 would continue to diminish until the equilibrium is reached. The amount of spotted DNA of a spot (s) available for hybridization however, will decrease with an increasing amount of hybridized target x_s ($s = s_0 - x_s$, s_0 being the spot capacity or maximal amount of available probe), so that we can write at thermodynamic equilibrium:

$$\frac{x_s}{x_0\left(s_0 - x_s\right)} = K_A \tag{5.2}$$

The spot capacity s_0 follows a certain distribution around an average spot capacity μ_s: $s_0 = \mu_s + \varepsilon_s$ or $s_0 = \mu_s e^{\varepsilon_s}$ with the spot error $\varepsilon_s \sim N(0,\sigma_s)$. Whichever distribution is more appropriate will depend largely on the type of microarray slide and spotting procedure used. The spot parameters μ_s and σ_s can be considered equal for all measurements of a single array. Finally, we assume that the presence of distinct labels (Cy3 and Cy5) does not influence the hybridization efficiency of the differentially labelled transcripts, i.e.:

$$x_0 = x_{0,Cy3} + x_{0,Cy5} \qquad \text{and} \qquad \frac{x_{0,Cy5}}{x_{0,Cy3}} = \frac{x_{s,Cy5}}{x_{s,Cy3}}$$

$$x_s = x_{s,Cy3} + x_{s,Cy5}$$

In the above equations, it would be more accurate to explicitly model the amount of non-labelled target in the solution, and to include parameters for labelling efficiencies, i.e. to write:

$$x_0 = x_0^* + x_{0,Cy3} + x_{0,Cy5}$$

$$x_s = x_s^* + x_{s,Cy3} + x_{s,Cy5}$$

with x_0^* being the amount of non-labelled target in the hybridization solution, and x_s^* being the amount of non-labelled target bound to the spotted probe. However, the external control spikes are added to the hybridization solution before the actual labelling reaction, and so effects attributed to

labelling efficiency can be accounted for in the dye saturation function, which is described in the following section.

5.2.1.2 Dye saturation function

A second component of the model is the dye saturation function, which describes the relationship between the measured intensity y and the amount of labelled target x_s, hybridized to a single spot on the microarray:

$$y = p_1 x_s e^{\varepsilon_m} + p_2 + \varepsilon_a \tag{5.3}$$

This dye saturation function is a simple linear equation incorporating an additive and multiplicative intensity error (this type of function stems from analytical chemistry [167] and has already been used in other normalization strategies [61-63,98,165,166]). The parameters p_1 and p_2 are the respective slope and intercept of the linear function. The additive and multiplicative errors are both assumed to be independently sampled from normal distributions, represented by $\varepsilon_a \sim N(0,\sigma_a)$ and $\varepsilon_m \sim N(0,\sigma_m)$ respectively. The additive intensity error term is included because of the generally accepted notion that the contribution of slide background is additive with respect to the spot intensity [34] (see section 2.2.2.1 of chapter 2). The validity of this notion can be confirmed by negative control spikes (probes to which no labelled target should bind), which generally show a relatively large variation in intensity. The multiplicative intensity error is included to account for the typical multiplicative error that is present in microarray data, and is usually dealt with by log-transforming the measurements (see section 2.2.2.2 of chapter 2).

In total, there are three different error distributions that are assumed to influence intensity measurements: additive intensity error ε_a, multiplicative intensity error ε_m, and spot capacity error ε_s. Figure 5.2 shows the effect of each error on measured intensity, and illustrates how spot related errors can account for the large variation in intensity observed in panel A of Figure 5.1. In the case of spotted microarray data, the plot would be slightly more complex, as there would be two distinct curves (one for Cy3 and one for Cy5) that are dependent on one and other through the parameters of the hybridization reaction. Indeed, the parameters of the saturation function and the variances of the intensity error distributions are considered specific for all measurements of a single each array and dye combination. The parameters of the hybridization reaction and variance of the spot error on the other hand, apply to all measurements of a single array. As such, Cy3 and Cy5 intensities obtained from the same array element are modelled with different saturation parameters and intensity errors, but will share the same hybridization parameters and spot error. For a single array, the resulting functions that relate measured intensities y_{Cy3} and y_{Cy5} to the amount of

corresponding target $x_{0,Cy3}$ and $x_{0,Cy5}$ in the hybridization solution, can be written as:

$$y_{Cy3} = p_{1,Cy3}\left(\frac{x_{0,Cy3}s_0}{1 + x_{0,Cy3}K_A^{-1} + x_{0,Cy5}}\right)e^{\varepsilon_{m,Cy3}} + p_{2,Cy3} + \varepsilon_{a.Cy3} \tag{5.4}$$

$$y_{Cy5} = p_{1,Cy5}\left(\frac{x_{0,Cy5}s_0}{1 + x_{0,Cy5}K_A^{-1} + x_{0,Cy3}}\right)e^{\varepsilon_{m,Cy5}} + p_{2,Cy5} + \varepsilon_{a,Cy5} \tag{5.5}$$

Where, as mentioned previously, the spot capacity s_0 follows a certain distribution around an average spot capacity μ_s: $s_0 = \mu_s + \varepsilon_s$ or $s_0 = \mu_s e^{\varepsilon_s}$. The differentially labelled targets $x_{0,Cy3}$ and $x_{0,Cy5}$ will compete for the same spotted probe DNA s_0. As shown in the equations above, the intensity measured for the Cy3 channel (y_{Cy3}) is not only dependent on the amount of Cy3 labelled target ($x_{0,Cy3}$), but also on the amount of target labelled with Cy5 ($x_{0,Cy5}$), and *visa versa*.

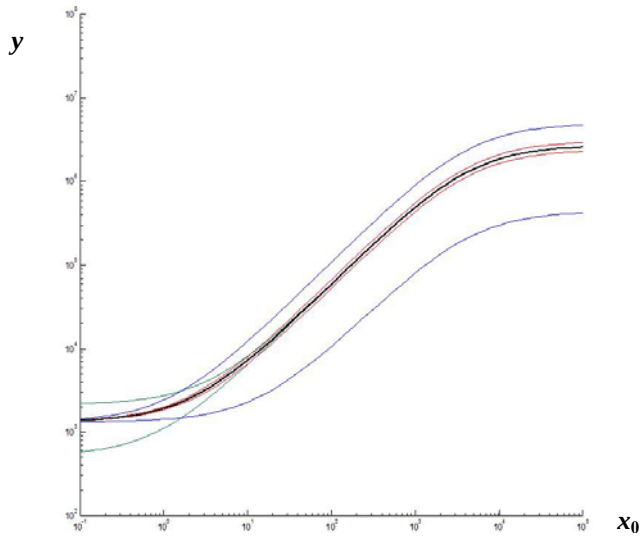


Figure 5.2: **Calibration model.** Illustration of model shape and influence of respective error distributions on measured intensities (one dye only). The thick black curve represents the relation between concentration and intensity if all error contribution were zero. The coloured lines represent 99% confidence intervals for the separate errors in the model: green corresponds to additive intensity error ε_a, red to multiplicative intensity error ε_m, and blue corresponds to spot capacity error ε_s. Model parameters and error variances are based on estimates from actual microarray data (see section 5.3.1).

5.2.2 Parameter estimation

The model parameters are estimated separately for each microarray, based on the measured intensities y of the external control spikes and their known concentration in the hybridization solution x_0. In order to determine these model parameters, it is important to have initial, reliable values for σ_m and σ_a. Estimates for $\sigma_{a,Cy3}$ and $\sigma_{a,Cy5}$ can easily be obtained by computing the standard deviation of the intensities for the negative control spikes whose targets not present in the hybridization solution. Finding a reliable for $\sigma_{m,Cy3}$ and $\sigma_{m,Cy5}$ is less evident. Although the additive intensity error can be neglected, the multiplicative errors are still confounded with the influence of spot errors at high intensity levels. Estimating $\sigma_{m,Cy3}$ and $\sigma_{m,Cy5}$ independently for both channels from these higher intensity replicate measurements is not feasible. Obtaining an adequate approximation is nonetheless possible. In the higher intensity range where the calibration controls (ratio 1:1) exhibit a log-linear behaviour in a y_{Cy3} versus y_{Cy5} plot (Figure 5.3), the main contribution to the observed variation can be assigned to the multiplicative intensity error. Indeed in this range, differences in spot size will obviously nullify themselves and the additive intensity error can be neglected. If we then assume that $\sigma_{m,Cy3}$ and $\sigma_{m,Cy5}$ contribute equally to the observed variation ($\sigma_m = \sigma_{m,Cy3} = \sigma_{m,Cy5}$), a value for σ_m can be obtained (Figure 5.3) by performing an orthogonal regression on the selected data points.

Obtaining a solution for the remaining parameters (dye saturation and hybridization parameters $p_{1,Cy3}$, $p_{1,Cy5}$, $p_{2,Cy3}$, $p_{2,Cy5}$ and K_A respectively; μ_s is kept constant at an arbitrary value) is done in a least squares sense. The error sum of squares that is minimized is that of spot capacity errors, i.e.

$$\min\left(SSE_s = \sum_i \varepsilon_{s,i}^2 \right) \tag{5.6}$$

with respect to $p_{1,Cy3}$, $p_{2,Cy3}$, $p_{1,Cy5}$, $p_{2,Cy5}$ and K_A.

The minimization of SSE_s is done numerically. The individual spot errors, necessary to calculate the SSE_s in every iteration (i.e. for any given set of parameter values), are of course unknown. For every spot on the microarray, they are estimated by comparing the expected intensity (a function of target concentration $x_{0,Cy3}$ and $x_{0,Cy5}$, and a set of parameter values as indicated by (5.4) and (5.5)) to the measured intensity values (y_{Cy3} and y_{Cy5}) for both channels, and scoring the difference based on the estimators of additive and multiplicative intensity variances. More precisely, for each pair of measurements obtained from a single spot, the following object function is minimized with respect to that spots error ε_s:

$$Q_{estim} = Q_{estim}^{Cy3} + Q_{estim}^{Cy5} \tag{5.7}$$

with respect to ε_s and where:

$$Q_{estim}^{D} = \arg\min_{\varepsilon_m, \varepsilon_a} \left(\left(\frac{\varepsilon_m}{\sigma_m \sqrt{2}} \right)^2 + \left(\frac{\varepsilon_a}{\sigma_a \sqrt{2}} \right)^2 \right)_D \tag{5.8}$$

$$D = Cy3, Cy5$$

subject to equations (5.4) and (5.5)

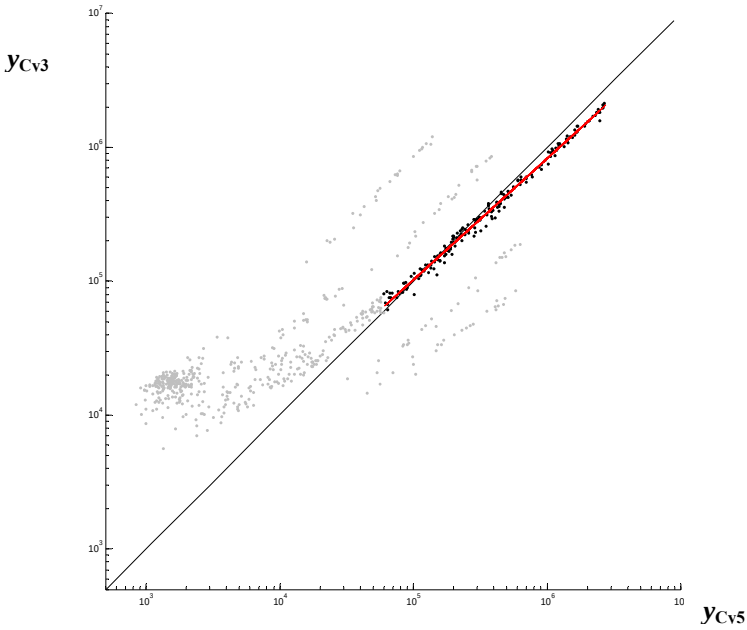


Figure 5.3: **Multiplicative intensity error.** Estimation of multiplicative intensity error is done on a subset of spikes (black dots; all other spikes are indicated by grey dots). Performing an orthogonal regression of Cy5 vs. Cy3 intensities on the selected data points (red line) will yield an error distribution of which the standard deviation is an estimate of $\sqrt{2}\sigma_m$.

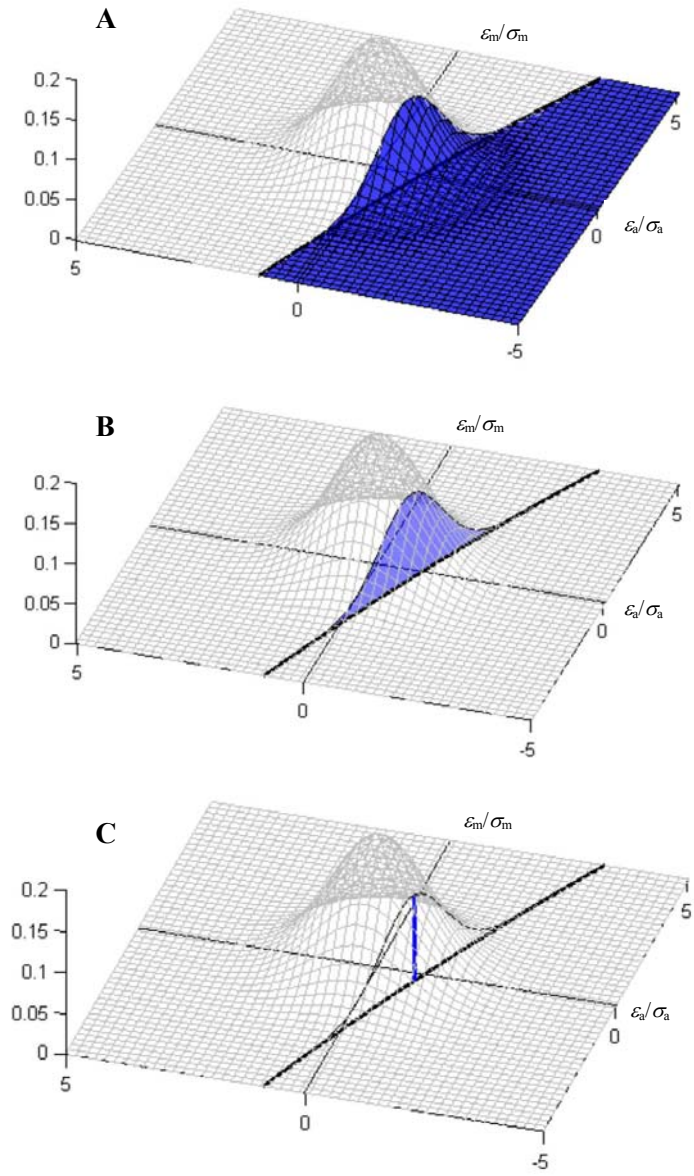


Figure 5.4: Motivation of cost function term (5.8). The grey surface corresponds to the probability density function for observing a certain intensity (a combination of additive and multiplicative error in the plane) for a single x_0, given a set of model parameters. The black curve in all three panels represents the deviation of a measured intensity from the expected intensity (as dictated by the model parameters), which can be explained by an infinite number of combinations of an additive and multiplicative intensity error. A *p*-value for observing such a measurement would be calculated as the blue volume in panel A, while a likelihood would be calculated as the blue area in panel B. Cost function term (5.8) is related to these as it is the negative logarithm of the maximum probability density function (i.e. the negative logarithm of the height indicated by the blue line in panel C).

The object function (5.7) is composed of two terms, one for the Cy5 intensity, and another for the Cy3 intensity. Both of these terms consist of a minimization as shown in equation (5.8) and are related to the probability of observing the measured Cy3 and Cy5 intensities for a given spot error ε_s and a given set of model parameters, in which case the expected intensities (and the amounts of hybridized target) can be calculated (equations (5.4) and (5.5)) since target concentrations of spikes are known. The relation between this probability and cost function term (5.8) is illustrated in Figure 5.4. The deviation of a measured intensity from the expected intensity could be explained by an infinite number of combinations of an additive and multiplicative intensity error, represented by the black curve in all three panels. Calculating a *p*-value for observing a certain intensity deviation to address its likeliness would require solving the following double integral, corresponding to the volume indicated in panel A of Figure 5.4:

$$P\left(\Delta \geq \left|y - p_1 x_s - p_2\right|\right) = \frac{1}{2\pi\sigma_m\sigma_a}\iint_A e^{-\left(\frac{\varepsilon_m}{\sigma_m\sqrt{2}}\right)^2} e^{-\left(\frac{\varepsilon_a}{\sigma_a\sqrt{2}}\right)^2} d\varepsilon_m d\varepsilon_a$$

with $A \equiv \left|y - p_1 x s e^{\varepsilon_m} - p_2 - \varepsilon_a \geq 0\right|$

Another option would be to calculate the likelihood of observing a certain intensity measurement for a given expected intensity, which would require solving the following integral, corresponding to the area indicated in panel B of Figure 5.4:

$$P(Y = y) = \frac{1}{2\pi\sigma_m\sigma_a}\int e^{-\left(\frac{\varepsilon_m}{\sigma_m\sqrt{2}}\right)^2} e^{-\left(\frac{y - p_1 x_s e^{\varepsilon_m} - p_2}{\sigma_a\sqrt{2}}\right)^2} d\varepsilon_m$$

$$= \frac{1}{2\pi\sigma_m\sigma_a}\int e^{-\left(\frac{\ln(y - p_2 - \varepsilon_a) - \ln(p_2 x_s)}{\sigma_m\sqrt{2}}\right)^2} e^{-\left(\frac{\varepsilon_a}{\sigma_a\sqrt{2}}\right)^2} d\varepsilon_a$$

The minimization in (5.8) is related to both (but far less computationally expensive) in that it corresponds to the negative logarithm of the maximum probability density value (i.e. a combination of a single ε_a and ε_m) for observing a measured intensity for a given expected intensity, as illustrated in panel C of Figure 5.4, i.e.:

$$-\ln\left(\underset{\varepsilon_a,\varepsilon_m}{\operatorname{argmax}}\left(P(Y = y)\right)\right) = \underset{\varepsilon_a,\varepsilon_m}{\operatorname{argmin}}\left(\left(\frac{\varepsilon_m}{\sigma_m\sqrt{2}}\right)^2 + \left(\frac{\varepsilon_a}{\sigma_a\sqrt{2}}\right)^2\right)$$

subject to equations (5.4) and (5.5), i.e.

$$-\ln\left(\underset{\varepsilon_a,\varepsilon_m}{\operatorname{argmax}}\left(P(Y=y)\right)\right)=\underset{\varepsilon_m}{\operatorname{argmin}}\left(\left(\frac{\varepsilon_m}{\sigma_m\sqrt{2}}\right)^2+\left(\frac{y-p_1x_se^{\varepsilon_m}-p_2}{\sigma_a\sqrt{2}}\right)^2\right)$$

$$=\underset{\varepsilon_a}{\operatorname{argmin}}\left(\left(\frac{\ln\left(y-p_2-\varepsilon_a\right)-\ln\left(p_2x_s\right)}{\sigma_m\sqrt{2}}\right)^2+\left(\frac{\varepsilon_a}{\sigma_a\sqrt{2}}\right)^2\right)$$

The parameter estimation procedure for an entire microarray is illustrated in Figure 5.5. Panel A shows initial parameter settings (red and green curves), while panel B shows the final parameter settings. The grey dots in Figure 5.5 depict the relation between measured intensity and amount of hybridized target under the assumption of equal spot sizes (i.e. all ε_s are zero). Most of these are localized in regions of high intensity error and are therefore very unlikely. However, by allowing errors on individual spot capacities, and thus altering the amount of hybridized target per spot for both dyes ($x_{s,Cy3}$ and $x_{s,Cy5}$), a good correspondence between intensities and saturation curves can be obtained for both channels, and across the entire measurement range (indicated by the black dots). The parameters of the intensity error distributions, σ_m and σ_a, determine the allowed spread of measurements around the Cy3 and Cy5 saturation curves. It is notable how well the Cy3 and Cy5 intensities, and the relationships between them, can be explained by our model. For instance in the example given, at lower intensities, Cy3 intensities are persistently higher than Cy5 for equal amounts of hybridized target, while the opposite is true for higher levels, a trend that is nicely reflected by the fitted model. Notice also that, while the ratios between Cy3 and Cy5 intensities are highly conserved (at least at higher intensity levels), absolute intensities may vary to a large extent for transcripts with the same x_0 due to spot inhomogenities.

5.2.3 Normalization: estimation of absolute expression levels

The obtained parameter values can be used to estimate a single $x_0(i,j)$ (i.e. the absolute expression level of a single gene i in a single biological condition j) based on all measurements that were obtained for this combination of gene and condition. Although each array and dye combination is attributed with its own set of parameters, the normalization can be considered a global one. Namely, for each combination of a gene and a tested biological condition, a single absolute expression level of target is estimated, irrespective of the number of microarray slides, or the number of replicate spots on a slide, on which this gene condition combination was measured. In this sense, the

results format of this normalization is comparable to the *gene×condition* interaction factor effects in the models of chapter 3 and chapter 4, or similar factors in other ANOVA models.

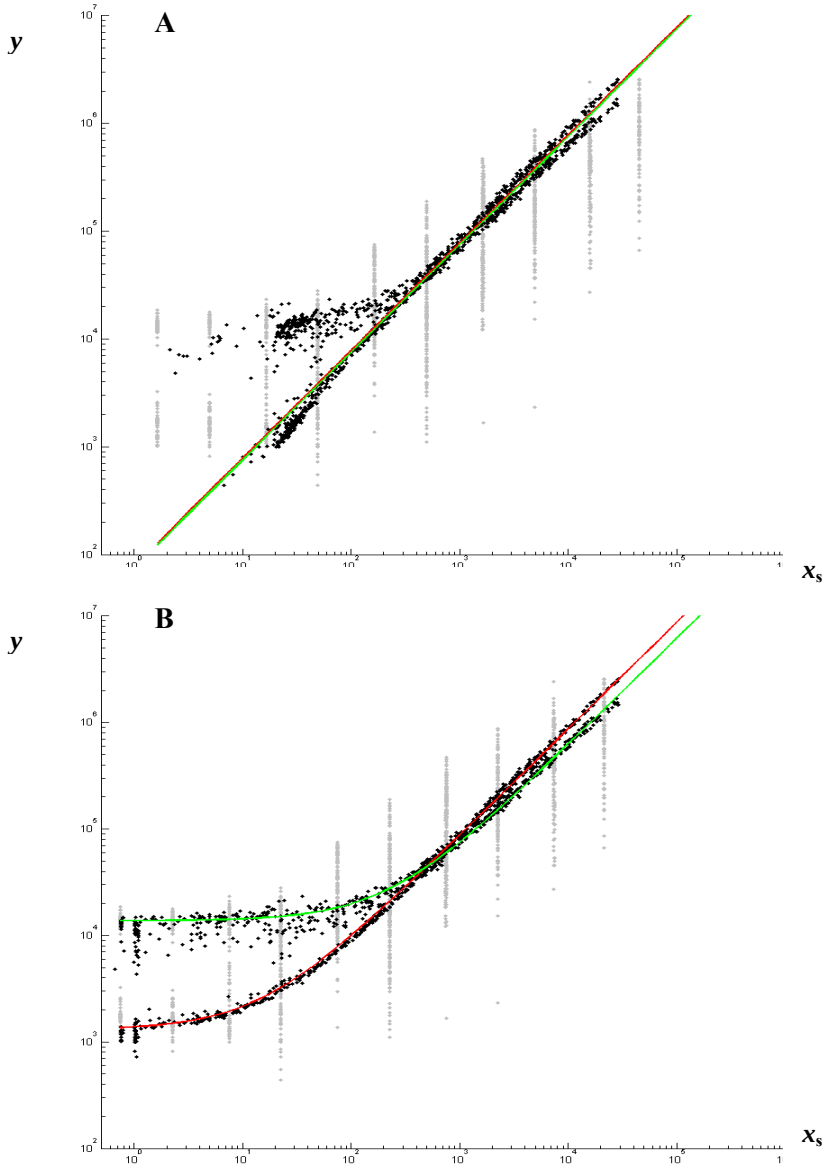


Figure 5.5: **Parameter estimation.** At given parameter values (red and green curve), spot errors are obtained by estimating the amount of hybridized target x_s for the measured intensities y of the external control spikes (black dots). Grey dots depict the amount of hybridized target, assuming equal spot capacities (i.e. all $\varepsilon_s = 0$). Panel A: initial parameter settings in this example are chosen to give a linear relation across the entire measurement range and equal for Cy3 and Cy5. Panel B: final parameter settings (convergence to a minimal SSE_s).

Although this procedure can be applied to any design, its complexity does depend on the used experimental setup. For a single gene, it requires the estimation of expression values for all the biological conditions at once. These $x_0(i,C)$ can be estimated by minimizing the following object function (an extension of the one used to estimate the model parameters):

$$Q_{norm} = \sum_{C} \sum_{S_j} Q_{norm}^{S_j(k)} \tag{5.9}$$

with respect to $x_0(i,C)$ and where:

$$Q_{norm}^{S_j(k)} = \left(\underset{\varepsilon_m, \varepsilon_a}{\arg\min} \left(\left(\frac{\varepsilon_m}{\sigma_m \sqrt{2}} \right)^2 + \left(\frac{\varepsilon_a}{\sigma_a \sqrt{2}} \right)^2 \right) + \left(\frac{\varepsilon_s}{\sigma_s \sqrt{2}} \right)^2 \right)_{S_j(k)} \tag{5.10}$$

subject to equations (5.4) and (5.5)

The subscript C indicates the entire set of biological conditions under survey; it applies to all conditions that are present in the experimental design. The set of data points, and the relevant array-dye combinations of parameters, that measure an expression value $x_0(i,j)$, is represented by S_j (a single data point belonging to this set is designated by $S_j(k)$). So for a single gene i, expression values of all of the biological condition present in the experiment are estimated simultaneously (and together with all the relevant spot errors), and in such a way that the total contribution of the three random errors (i.e. the combined spot errors and additive and multiplicative intensity errors for all intensity data points that are a measure of gene i) is minimized as dictated by the cost function (5.9).

5.3 Application and results

5.3.1 Data set

A publicly available data set [96], consisting of 14 hybridizations, was chosen to illustrate the workings, advantages and drawbacks of our normalization method. This experiment was ideally suited to validate our procedure because firstly, it contained the necessary spots for measuring external control spikes, required for estimating the parameters of our model.

Secondly, the experimental design included only a single biological condition (self-self experiments), which allows assessing the performance of our normalization method in removing non-linear tendencies present in microarray data. Lastly, they were outfitted with an additional set of control spikes that could be used to verify to what extent our method was capable of approximating the absolute target concentrations.

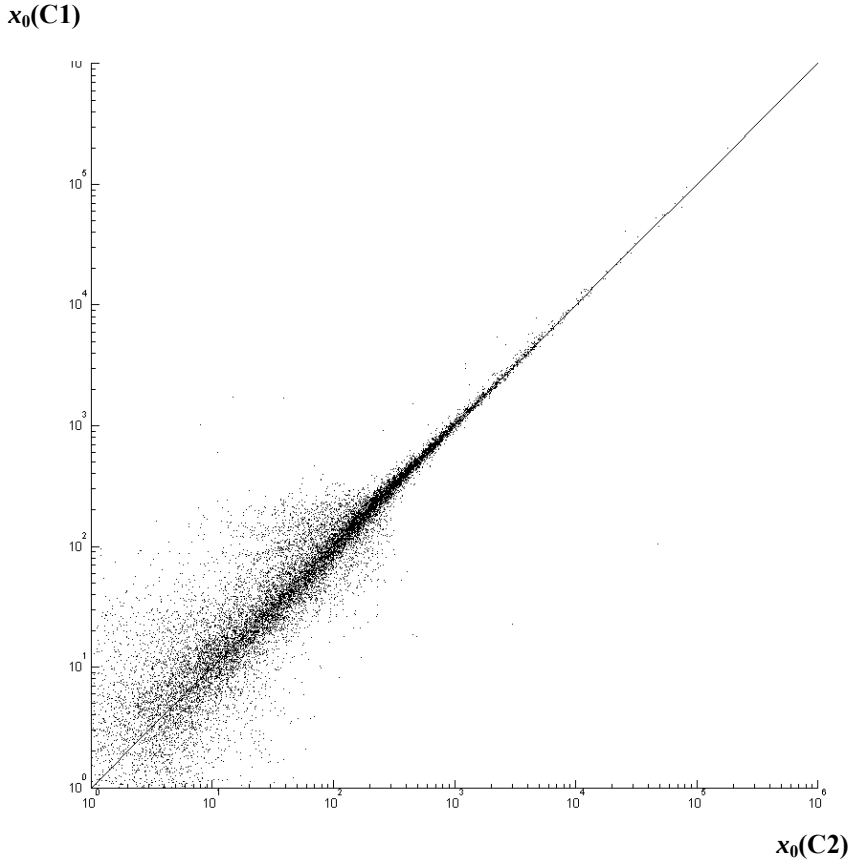


Figure 5.6: **Removal of non-linear artefacts.** Estimated expression levels for C1 are plotted against estimated levels for C2 after normalizing a hypothetical colour flip experiment. C1 and C2 in fact represent the same biological mRNA sample. The centring of data points around the bisector (solid line) indicates that typical microarray non-linearities are adequately accounted for.

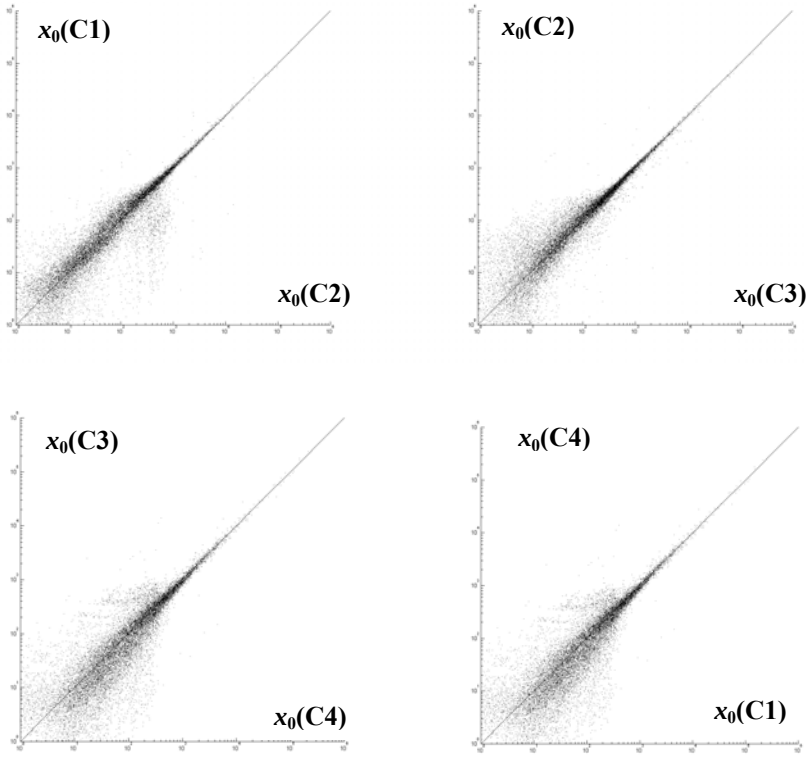


Figure 5.7: Removal of non-linear artefacts. Estimated expression levels are plotted against each other after normalizing a loop design experiment with 4 different hypothetical conditions (designated C1, C2, C3 and C4). These conditions in fact represent the same biological mRNA sample. The centring of data points around the bisector (solid line) indicates that typical microarray non-linearities are adequately accounted for.

5.3.2 Removal of non-linear artefacts

Figure 5.6 illustrates the result of applying our method on a selection of two arrays from the 14-array experiment. As this is a self-self design, the same biological sample was measured 4 times on these 2 arrays (twice labelled with Cy3 and twice with Cy5). For the purpose of our test, we treated this self-self experiment as a colour flip design with two hypothetically different samples (designated C1 and C2). Estimated expression levels x_0 of the approximately 19000 genes are plotted in Figure 5.6 for C1 vs. C2. Because in reality C1 and C2 represent the same biological condition, all estimates being centred along the bisector indicates that our model adequately accounts for the major sources of non-linear variation in the data. The

increased variance of the estimates observed at lower target levels is inherent to microarray technology. This range of expression corresponds to the saturation observed in the lower intensity region, i.e. where the additive error has a significant influence, considerably blurring the relationship between measured intensity y and expression level x_0. Because of these saturation effects, estimates of lower concentration are prone to be less reliable.

As mentioned previously, our method is not bound by experimental design. To illustrate that these results are not only achievable with simple experimental setups, such as a colour flip, we normalized a set of 4 arrays as if it concerned a loop design with 4 different biological conditions (dubbed C1, C2, C3 and C4). A comparison of the estimated expression levels is shown in Figure 5.7.

5.3.3 Comparison to LOWESS+ANOVA

We will illustrate the difference between our method and a LOWESS fit plus ANOVA normalization, by comparing the results of the 4 array loop design described in the previous section. The same 4 arrays were rescaled for intensity dependent dye bias with a LOWESS procedure (smoothing factor f set to 30%), followed by an ANOVA normalization with model (4.13) (see chapter 4, section 4.2). The results are shown in Figure 5.8, where for both approaches, estimated expression levels for C1 are plotted versus those for C3, and estimated expression levels for C2 are plotted versus those for C4. By doing so, these plots directly compare the biological conditions that were never measured together on the same microarray slide.

What is immediately notable is the relatively high error variance our method displays for the lower range of estimated expression levels. Such a feature is completely absent from the LOWESS+ANOVA normalized data. It should be noted however, that panel A and B are plotted at different scales, and that the axis for panel B (our method) show more orders of magnitude than those of panel A (LOWESS+ANOVA). In fact, the orders of magnitude that compromise the LOWESS+ANOVA estimates of expression roughly correspond to the higher range of expression for our method, where estimates are very accurate. An explanation may be found in the fact that our method attempts to estimate actual expression levels and is greatly influenced by the saturation characteristics of the data. An ANOVA normalization on the other hand, is a basic linear rescaling of the measured intensities, and as such, intensities that are saturated will remain saturated after normalization. The spread out, lower range of estimates obtained with our method actually correspond to the dense blob of points at the lowest expression levels for the LOWESS+ANOVA method. When interpreted correctly, both features tell the same tale: these estimates correspond to a

range of intensities that is saturated, and for which it is extremely hard to discern any useful information regarding the absolute levels of gene expression. Another observation that is worthy of mentioning is that the normalized values from the LOWESS+ANOVA method are not completely symmetrical and, contrary our method, do not show an increased accuracy for higher expression levels.

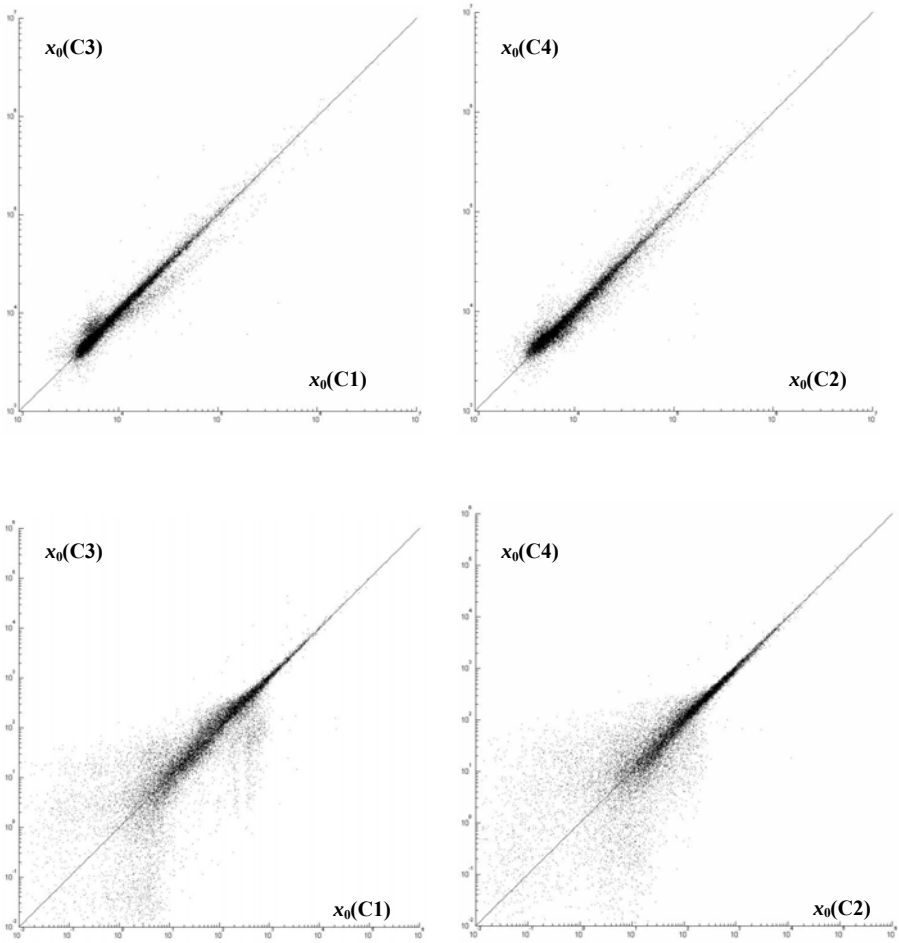


Figure 5.8: Comparison with LOWESS+ANOVA. An illustration of the difference between a LOWESS fit (performed according to the GNA) plus ANOVA normalization (upper two panels), and the calibration method developed in this chapter (lower two panels). Four arrays that were hybridized with labelled target representing the same biological condition were normalized as if it concerned a loop design of 4 different conditions. Estimated expression levels for conditions that were never measured together on the same microarray slide are directly compared in the plots (i.e. estimated expression levels for C1 are plotted versus those for C3, and estimated expression levels for C2 are plotted versus those for C4).

In this particular case, where all microarrays in fact constituted self-self hybridizations, the GNA that serves as the basis for rescaling the log-ratios according to a LOWESS fitted curve is a valid assumption. It could nevertheless be argued that, in order to provide a decent comparison, results of our method should be compared to data that was rescaled according to a LOWESS curve fitted on the external control spikes with a ratio of 1:1. Indeed, if spikes are present on an array, these should be used to perform the rescaling instead of using all measurements. Using spikes to fit the LOWESS curve ensures that the rescaling is independent of the GNA. The results of this analysis are shown in Figure 5.9. These are not markedly different from the ones depicted in panel A of Figure 5.9, and the points discussed above are also applicable in this case. The fact that performing a LOWESS fit on the 1:1 spikes alone is not capable of completely linearizing data illustrates what was speculated in section 4.4.2 of chapter 4: intensity dependent normalization methods, such as LOWESS, will merely remove the nonlinearities between the Cy3 and Cy5 intensity measurements, and not between the measured intensity and the concentration of labelled target.

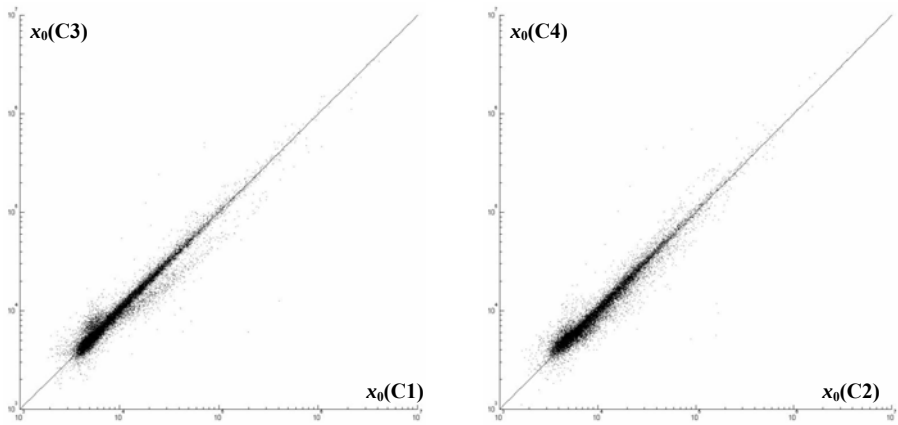


Figure 5.9: **Comparison with LOWESS+ANOVA.** Results of a LOWESS fit (performed on external control spikes) plus ANOVA normalization. These plots show essentially the same data as the upper two panels of Figure 5.8, the one difference being that the LOWESS fit was now performed on the external controls that were spiked in a ratio 1:1, and not on all genes present on the microarray (i.e. not according to the GNA).

Table 5.1: Mixes of the 14 external control spikes.

Spike	Spike Mix 1	Spike Mix 2	Spike Mix 3	Spike Mix 4	Spike Mix 5	Spike Mix 6	Spike Mix 7	Reference Mix
DilA1	10000	0	0.1	1	10	100	1000	100
DilA2	1000	10000	0	0.1	1	100	100	100
DilA3	100	1000	10000	0	0.1	1	10	100
DilA4	10	100	1000	10000	0	0.1	1	100
DilA5	1	10	100	1000	10000	0	0.1	100
DilA6	0.1	1	10	100	1000	10000	0	100
DilA7	0	0.1	1	10	100	1000	10000	100
DilB1	10000	0	0.1	1	10	100	1000	100
DilB2	1000	10000	0	0.1	1	10	100	100
DilB3	100	1000	10000	0	0.1	1	10	100
DilB4	10	100	1000	10000	0	0.1	1	100
DilB5	1	10	100	1000	10000	0	0.1	100
DilB6	0.1	1	10	100	1000	10000	0	100
DilB7	0	0.1	1	10	100	1000	10000	100

Note: These spike mixes were added tot the hybridization samples, prior to labeling. From the total of 14 arrays, 7 were hybridized with the respective spike mixes labeled in Cy5, each time against the reference mix labeled in Cy3. The remaining 7 arrays were hybridized with the respective spike mixes labeled in Cy3, each time against the reference mix labeled in Cy5. Concentrations are given in copy number per cell. *DilB6* was omitted from analysis due to quality issues [4].

5.3.4 Evaluation of absolute expression level estimates

Although we have shown that our method is capable of estimating absolute expression levels that respect true ratios between the different conditions compared, the previous experiment does not reveal anything about the accuracy of these absolute estimates, i.e. it does not show to what extent these absolute expression levels approximate the actual concentrations of target in the hybridization solution.

To verify the accuracy of estimated target concentrations, they should be compared with their actual concentrations in the hybridization solution. Doing this for the entire population of transcripts is impossible, as for most of the genes this concentration is unknown. However, the data set contains an additional set of non commercial spikes for which the absolute concentrations in the hybridization solution are known. The extracted mRNA samples were complemented with fourteen external controls at amounts of 10^4, 10^3, 10^2, 10, 1, 0.1 or zero copies per cell. In all fourteen hybridizations, these controls were compared with a unique reference RNA, capable of binding to all of the 14 spike probes, always added at a concentration of 100 copies per cell. The experimental design for these control spikes is summarized in Table 5.1. Results obtained after performing our normalization are shown in Figure 5.10 (one spike was omitted from analysis because of quality issues [4]). Because the estimated target concentrations, expressed in pg/ml, were not directly comparable to the units of copy number per cell, a linear rescaling of these values by a factor that set our estimate of the unique reference RNA to '100' (copies per cell) was

performed. Figure 5.10 shows that, except for the lowest concentrations, estimated values correspond fairly well to the true target concentrations as present in the hybridization solution. As explained above, also here estimates of the lowest concentrations show a higher error variance.

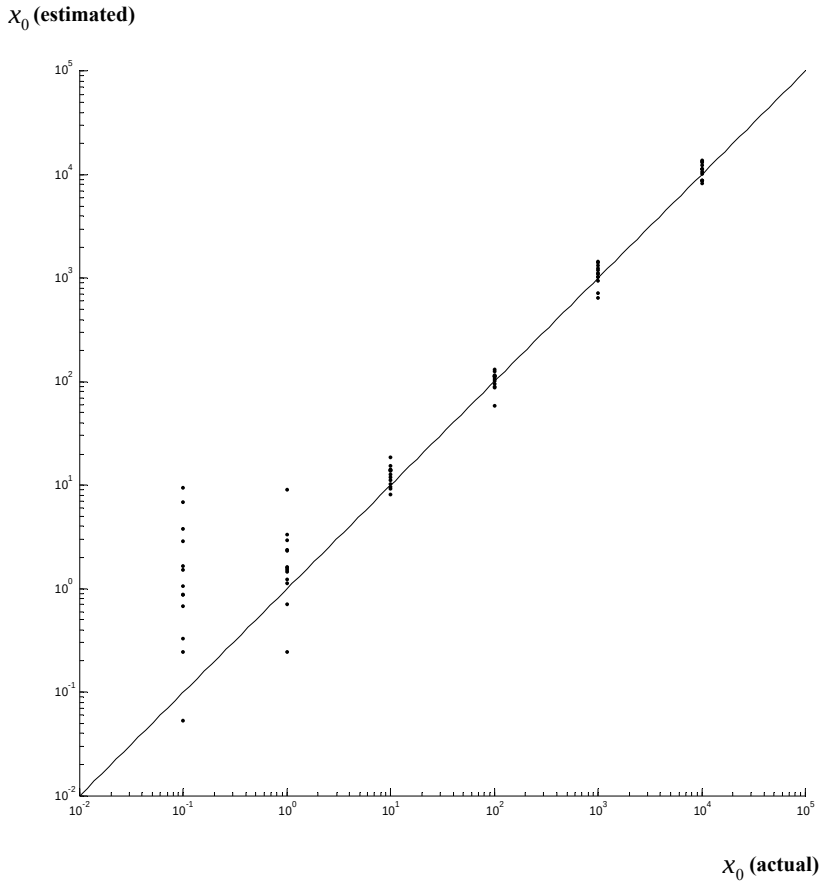


Figure 5.10: Evaluation of absolute expression level estimates. Estimated mRNA concentrations (copy number per cell) for all of the 13 controls are plotted against the actual, spiked concentrations. The solid line depicts the bisector. Execpt for the lowest concentrations, estimated values correspond well to the true target concentrations as present in the hybridization solution.

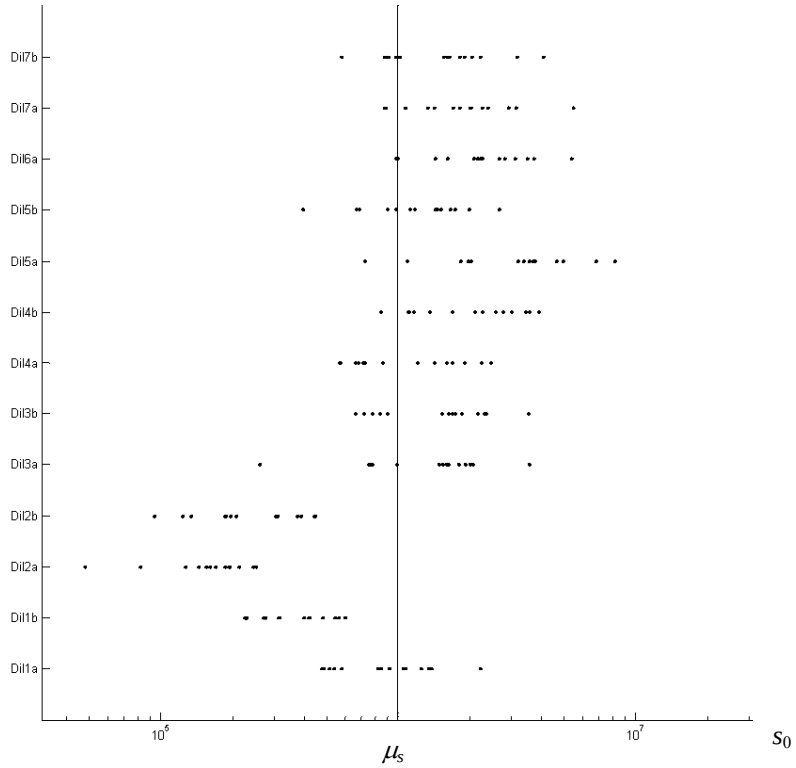


Figure 5.11: Consistent spot errors. Estimated spot capacities s_0, corresponding to the 14 microarrays of the experimental design, are plotted for each of the 13 external controls, revealing consistent (per spike), and across-array spot errors. The solid line represents the mean spot capacity.

5.3.5 Comparison of estimated concentrations between genes

Although Figure 5.10 shows that concentrations can be accurately estimated, there are several gene-dependent factors that could influence the obtained results, possibly hampering the comparison of estimated concentrations between different genes. Gene specific hybridization efficiencies for instance, are not taken into account by our model. 'Consistent spot errors' are another factor for which it is theoretically impossible to compensate. Microarrays are usually spotted in series: experimental errors that influence the DNA probe solutions used for spotting will affect an entire set of microarrays in a similar way. This type of 'consistent spot error' will manifest itself on individual spots across multiple microarray slides, contrary

to e.g. variations related to the spotting pins themselves, which would also affect multiple spots on a single array. The particular setup of the 13 external controls, used for assessing the accuracy of estimated expression levels, can provide some insight. Because the universal reference RNA can hybridize to all the probes of these spikes, it couples the spot errors of all probes during the estimation of target concentrations. As a consequence of this coupling, consistent spot errors could partially be compensated for, as illustrated in Figure 5.11. For certain spikes (e.g. *Dil2a*), estimated spot capacities were persistently above or below the average capacity, a feature that was only detectable through the presence of the universal reference RNA. As a result, estimated target concentrations can be subject to gene specific rescaling, hampering the comparison of these concentrations between genes. They can nevertheless be interpreted as absolute values of expression when comparing different concentrations for a single gene.

5.3.6 Influence of local background corrections

In our model the combination of the additive intensity error ε_a and intercept of the dye saturation function p_2 can be regarded as an elementary model for the entire slide's background. Having a single background for all spots is different from the spot specific background corrections performed during standard microarray analysis, which estimate a spot specific background from pixels corresponding to the area of the glass slide surrounding the spotted probe (see chapter 2, section 2.2.2.1). This background model is by no means a restriction concerning the use of background corrected values; our normalization can be applied to both raw and background corrected intensities. Moreover, our method is perfectly capable of working with negative intensity values that may arise when measurements are lying below background.

Whether or not using background corrected measurements is advisable, depends largely on the data quality. The effects of background correction on estimated parameters and intensity data are illustrated in Figure 5.12. Performing a spot specific background correction prior to applying the model would ideally result in the lower saturation limit of our model (p_2) becoming zero. Local background intensities however, are only a crude approximation of the true background (for a detailed discussion, see section 2.2.2.1 of chapter 2), so in reality, the estimate for p_2 will indeed be lower, but never reaches a zero level. Moreover, the standard deviation estimate for the additive intensity error will be significantly larger than the estimate for non-background corrected intensities. In general, a trade off can be observed: background corrected measurements have a larger linear range, but at the expense of increased measurement errors for lower concentrations. This effect is not only discernable in the estimated parameters, but is often a

prominent feature when comparing MA-plots for raw versus background corrected intensities (see e.g. Figure 3.7 in chapter 3).

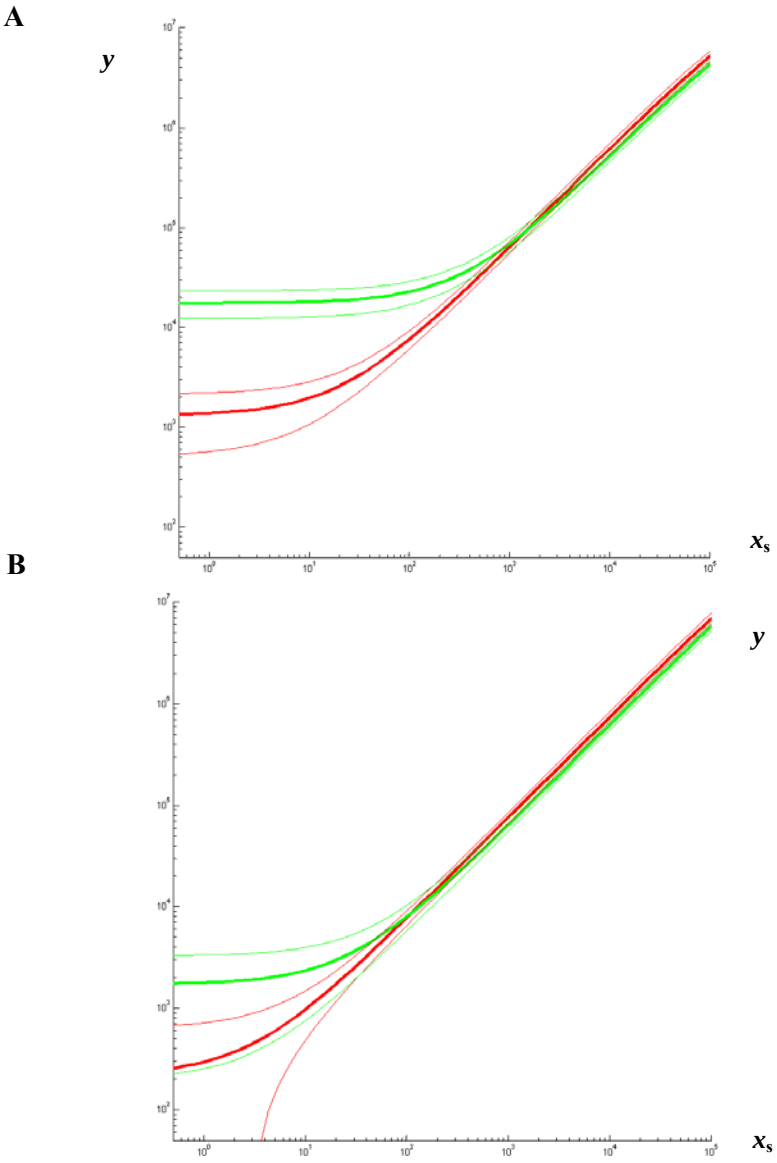


Figure 5.12: Effect of background correction. A) Model parameters (thick line) and 99% confidence interval for intensity errors (thin lines), estimated from raw, non-background corrected data (red = Cy5; green = Cy3) . B) Model parameters and 99% confidence interval for intensity errors, estimated from background corrected data. Compared to panel A, an increased linear range, as well as an increased error variance, can be observed for lower intensity measurements.

5.4 Discussion

In this chapter we presented an approach for normalizing spotted microarray data, using external control spikes to fit a calibration model. This model incorporates parameters and error distributions representing both the hybridization of labelled target to complementary probes, and the subsequent measurement of fluorescence intensities. External control spikes serve to estimate the model parameters. The obtained parameters values are then employed to estimate absolute levels of expression for the remaining genes. For each combination of a gene and a tested biological condition, a single absolute target expression level can be estimated, taken the specificities of the design. Incorporation of external control spikes is thus an absolute requirement for any data set to be normalized according to the methods described in this chapter. While the data that served as an example to illustrate the workings of procedure (see section 5.3.1) was outfitted with a commercial set of control spikes, any set of external controls that relates target concentration to measured intensity across the entire measurement range, can serve to supply the input calibration data. It is important to realize however, that the amount of controls and their overall quality will naturally have a large influence on the final, normalized data, because they will determine the accuracy of the estimated model parameters.

The calibration model in itself is fairly basic, in that, with the exception of spot size errors, it is aimed at capturing the global characteristics of an experiment and their overall influence on intensity measurements, generalizing on hard to quantify local sources of variation. The combination of the additive intensity error ε_a and intercept of the dye saturation function p_2 for instance, can be regarded as a global model for the entire slide's background.

The array specific hybridization constant K_a, another global factor, obviously does not account for transcript specific hybridization efficiencies. Therefore, care should be taken when interpreting the estimated expression levels as actual concentrations or when comparing estimated expression levels between genes. On the other hand, probe sequences for spotted microarrays are often specifically selected to have properties that obviate large differences in transcript specific hybridization effects (contrary to Affymetrix GeneChip arrays, where the short oligonucleotide probes show large differences in hybridization efficiency [89,90]). Besides these gene specific hybridization effects, comparison of estimated expression levels between genes is also complicated by 'consistent spot errors' across multiple slides. These errors, resulting from experimental inaccuracies in the probe DNA preparation, can arise when microarray slides are spotted in series. Due to the characteristics of microarray technology, they cannot be dealt with model wise.

Although our model is a simplification of physical reality dealing with errors in a global, non-gene specific way, results show that our method is capable of adequately linearizing and normalizing spotted microarray data. An important difference over most existing normalization methods is that our procedure does not rely on any assumptions on the distribution of gene expression levels from one biological sample to the next. Hence, our procedure is particularly well suited to normalize experiments for which the Global Normalization Assumption (GNA) may not be entirely valid, i.e. experiments for which there is no symmetry in the amount of genes that are up-regulated versus down-regulated. Such is typically the case with experiments comparing drastically contrasting biological conditions or with dedicated spotted microarrays, containing only a limited number of spotted probes, representing genes involved in the studied biological process.

In contrast to other normalization methods that use spikes to circumvent the Global Normalization Assumption [208], our procedure computes absolute expression levels, avoiding the use of ratios. Moreover, for the described experiment, the estimated absolute expression levels approximate the actual concentrations fairly well. Some caution is nevertheless advised when interpreting estimated concentrations as such. This is only problematic as far as comparing expression levels between different genes; the points discussed above have little or no consequence if a comparison is made between estimated target levels across biological conditions for a single gene.

Our method offers a novel approach to normalizing spotted microarrays that combines the advantages of approaches that attempt to estimate absolute expression levels [33,57,113], and methods that perform data linearization using the ratio distribution (e.g. LOWESS). The procedure offers independence of assumptions concerning the distribution of gene expression (i.e. the GNA) by retaining much of the inherent calibration information of external control spike measurements.

Chapter 6

Conclusions and outlook

The research presented in this PhD thesis dealt entirely with the normalization of data from spotted microarrays. The strategies that were pursued differ in spirit from most accepted techniques. Standard ratio based normalization is heavily bound to assumptions concerning the distribution of gene expression; they are guided by how expression levels are presumed to change across different biological conditions. Ratio normalization methods generally show little interest in the underlying causes of the observed systematic and random variation in microarray data. The underlying idea of our research was to acknowledge –as much as possible- the physical and biological reality of the process and address the normalization problem starting from units of absolute intensities. Instead of being limited to the relative nature of intensity ratios, we attempted to estimate absolute values of expression by modelling the measured intensities as a function of systematic sources of variation in an experimentally meaningful way. The results and observations that culminated from this work are summarized in section 6.1, followed by a short description of some concrete problems that will be studied in the future (section 6.2). An outlook on microarray normalization and spotted microarrays in general, is given in the final section of this dissertation (section 6.3).

6.1 Achievements

Initial research (described in chapter 3) consisted of the evaluation of ANOVA models for microarray normalization and comparing them to ratio based approaches [132]. ANOVA models were the first method to work with absolute intensities, linearly rescaling them to obtain estimates of absolute expression levels. This preliminary research demonstrated that ANOVA based normalization was not without its share of flaws, especially regarding the use of the residual distribution for identifying differentially

expressed genes. Nevertheless, it remained an interesting and potentially powerful tool that deserved further attention. Other notable observations showed that typical non-linear dye biases in the data prohibit the sole use of ANOVA for data normalization. Performing a LOWESS fit (or a similar array-by-array intensity based rescaling) prior to the application of the ANOVA model should therefore be considered as a required step in any ANOVA based normalization procedure.

ANOVA models for microarray normalization can not readily be applied to any type of experimental setup of a microarray experiment. Attempting to fit the published ANOVA models to different experimental designs can be a tedious task, and so further research was directed at the development of generic (applicable to any experimental setup) ANOVA models for microarray normalization (chapter 4). To insure the availability to a wide range of public, such a generic model was implemented in a user friendly web application [72] (http://www.esat.kuleuven.be/maran). Some interesting features were revealed during the course of this research. Results seemed to indicate that a LOWESS normalization may not be able to completely alleviate intensity dependent nonlinear tendencies in the data (despite of harsh assumptions with regards to the distribution of gene expression from one biological condition to the next).

These observations begged for a re-examination of our working hypothesis, and fuelled the research described in chapter 5. External control spikes provided further insight and led to the development of a novel normalizing method for spotted microarray data [73], which itself relies on such external controls to fit a calibration model. The model incorporates parameters and error distributions representing both the hybridization of labelled target to complementary probes, and the subsequent measurement of fluorescence intensities. External control spikes serve to estimate the model parameters. The obtained parameters values are then employed to estimate absolute levels of expression for the remaining genes. For each combination of a gene and a tested biological condition, a single absolute target expression level is estimated, taken the specificities of the design. The results that were obtained from applying our method to a publicly available data set show that the procedure is capable of adequately removing the typical non-linearities of microarray data, without making any assumptions on the distribution of differences in gene expression from one biological sample to the next and thus completely avoiding the GNA. The new procedure performed well compared to a standard LOWESS procedure (prior to fitting an ANOVA model), and might be considered superior in several aspects. More importantly, since our model links target concentration to measured intensity, absolute expression values of transcripts in the hybridization solution can be estimated with fair accuracy.

6.2 Future work

The calibration method described in chapter 5 will be the basis for further investigations. It is to be made widely available as a 'BioConductor Package' (www.bioconductor.org). BioConductor [76] is an 'open source' and 'open software development' project; its goal is providing access to software for the analysis of genomic data. Mostly all of the established methods for microarray normalization (as well as a plethora of tools for further analysis) are already available to the public in the form of BioConductor packages. As such, BioConductor is used across the globe, and on a large scale, for the analysis and interpretation of microarray data.

The implementation of a BioConductor package will coincide with a further elaboration of the semi-physical model (chapter 5, section 5.2.1) that is used in this normalization procedure. The model in itself is fairly basic, in that, with the exception of spot size errors, it is aimed at capturing the global characteristics of an experiment and their overall influence on intensity measurements, generalizing on often hard to quantify local sources of variation. In the models current form, there is plenty of room for improvement by adding parameters and error distribution that account for more local factors of experimental noise, with the intention of diminishing the error variances on the estimated target levels.

As the procedure itself estimates a single expression level based on all available replicates within the experiment, it is imperative that some measure of reliability for these estimates can be obtained as well. Some further research is necessary to find computationally inexpensive ways that adequately quantify the error distribution on the estimated target levels. With the proper test statistics, confidence intervals based on these distributions could be used to identify genes with significantly changing expression, or genes that were measured with a high inconsistency. Also, having a measure of reliability on the estimated expression levels could prove a welcome addition for downstream analysis algorithms used for data exploration (section 2.2.3, chapter 2).

Because of the universal principles that are the basis of this normalization procedure, it could be easily adapted to work with other molecular biological high-througput techniques. A promising candidate is the ChIP-chip technology (see chapter 2, section 2.3), used for identifying binding sites for DNA binding proteins. It relies on the use of spotted microarrays to compare the abundance of DNA molecules in two populations. One sample consists of immunoprecipitated DNA (i.e. DNA representing binding sites), the other sample serves as a negative control. For such a comparison, the GNA is utterly inappropriate, rendering standard normalization strategies useless and adding to the requirement of a large number of replicate

measurements to infer DNA binding sites with any statistical significance. The underlying principles of our model however (section 5.2.1 in chapter 5), are no different in this case than for the more common expression profiling experiments. Incorporating external controls into ChIP-chip experiments in order to normalize the data might result in more accurate predictions of DNA binding sites at fewer expenses (i.e. less microarray hybridizations). Other high-throughput technologies that are essentially different from microarrays, but share similar characteristics (such as differential labelling) might also benefit from a similar normalization approach. A good example is Fluorescence 2D Difference Gel Electrophoresis (2D-DIGE) technique that aims to compare the expression of proteins in different biological conditions [148,149,205,209]. 2D-DIGE uses molecular weight- and pI-matched, spectrally resolvable dyes (Cy2, Cy3 and Cy5) to label protein samples prior to 2D electrophoresis. By using different dyes to separately label proteins isolated from different biological conditions, multiple samples (up to three) can be co-separated and quantitated by three different set of wavelengths.

6.3 Outlook

Red light, green light…

The central issue that stood at the core of this thesis was modelling intensity measurements from spotted microarrays in a semi-physical way (i.e. acknowledging the experimental reality) in order to obtain absolute estimates of target expression levels. We attempted to create a normalization method by not treating microarray data merely as proportions of differential expression, but upholding the view that every measured intensity is a representation of an actual abundance of mRNA, subject to a variety of experimental factors which may –or may not- be mathematically modelled and accounted for.

When spotted microarrays were introduced in the mid 1990s, they empowered researchers with an impressive tool to compare gene expression characteristics on a high-throughput scale. The principles for interpreting the data were remarkably simple: red light indicated over-expression in one sample, green light indicated over-expression in the other sample, a yellow blob could be anything in between, and a black void corresponded to a lack of expression in either condition. Equally noteworthy, these red over green intensity ratios were accurately quantifiable. It soon became clear however, that getting reliable measures for differential expression was slightly more complex due to various experimental biases that are introduced during the course of an experiment. The need for proper normalization strategies thus quickly arose. Normalization methods were generally conceived as *ad hoc*

adjustments of the measured ratios, showing little interest in the underlying causes of the observed systematic and random variation in the data. To this day, little has changed. Elaborate mechanisms for rigorously quantifying different sources of random and systematic noise, and for assessing their influence on the relation between measured intensity and actual mRNA abundance, have yet to be established. Instead, microarray data are still routinely normalized by forcing corrected ratios to comply with an expected pattern of behaviour. This claim does not only hold true for many procedures that aim at removing dye related discrepancies, but also, for instance, for log-ratio variance stabilization techniques. The goal of such techniques is to counter the large variation at lower intensity levels that is often observed in MA-plots, thus facilitating further statistical inferences. What such methods fail to recognize, is that this region of increased ratio variance corresponds to the saturation range of intensities, i.e. where little or no information regarding actual mRNA concentration is retained in the measured intensities. Measured ratios in this region are often hard to reproduce, and replicate experiments show little consistency for these ratios (see e.g. chapter 3), meaning that genes with overall low intensity levels are hard to find significantly differentially expressed. One could dispute the usefulness of such variance stabilization techniques. It may be better to accept that, downwards from a certain point, there is minor knowledge to be gained with regards to actual expression levels.

We feel that the analysis of microarray data could benefit from a more methodical approach to its experimental nature. This however, would require some change in the way microarray experiments are performed and the data are managed. Most notably, there is the absolute requirement for inclusion of experimental controls, necessary to build calibration models and estimate model parameters. Experimental controls should not be limited to the type of external control spikes we relied on in chapter 5, but could be incorporated at different stages during sample processing. Ideally, controls would be added at each step of a microarray experiment to ensure a maximum discernment of systematic and random variation introduced at distinct phases. However in practice, experimental controls consist of external control mixes that are spiked only once, and logic dictates that this is done early during sample processing so that as many steps as possible are monitored.

Regrettably, incorporating external controls into microarray experiments comes with an added cost that not all researchers are willing to pay. This could soon change though, as there are some forces at work, such as the industry led External RNA Control Consortium [11] (ERCC; http://www.affymetrix.com/community/standards/index.affx), that are lobbying for a greater acceptance of external controls in microarray experiments. The ERCC is endeavouring to establish and develop an

affordable, universal set of external RNA controls that can be used across several organisms without fear of cross-hybridization. Providing standards for the incorporation of spikes brings about other problems. At present, the absolute amount and incorporation percentage of labelled material being applied to microarrays is usually not reported (this is not a requirement according to MIAME [29] guidelines). The volumes that are applied are sometimes chosen to have the same total fluorescence value, which does not necessarily correspond to applying equal amounts of target due to the differential labelling. Altering the proportion of Cy3 versus Cy5 sample which are hybridized will naturally affect the concentrations of the spiked controls, ultimately leading to wrong normalizations when not reported properly.

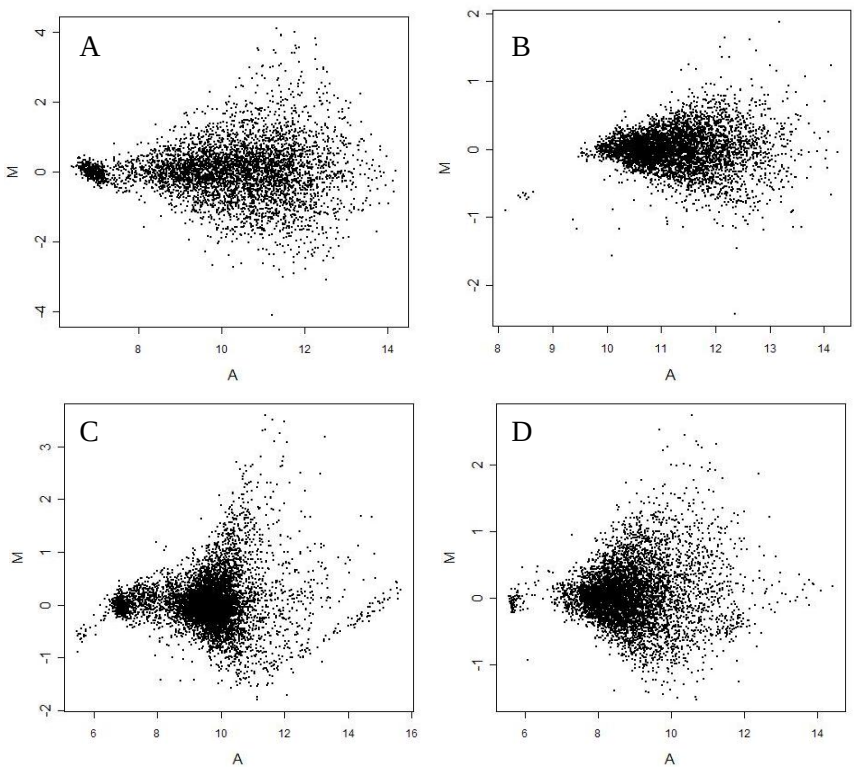


Figure 6.1: Inappropriate use of a GNA based normalization. Microarray data are often normalized according to the GNA regardless of features that reveal its inappropriateness. Such circumstances expose themselves by showing unusually large, often increasing log-ratio variances for higher average intensity levels. In this figure, some examples are shown of ill-suited use of GNA based normalization. A) Khodursky *et al*, 2000 [114] (Stanford Microarray Database Experiment ID 1639). B) Courcelle *et al*, 2001 [42] (SMD-ExpID 1290). C) Bernstein *et al*, 2002 [19] (SMD-ExpID 8589). D) Bernstein *et al*, 2004 [20] (SMD-ExpID 19340).

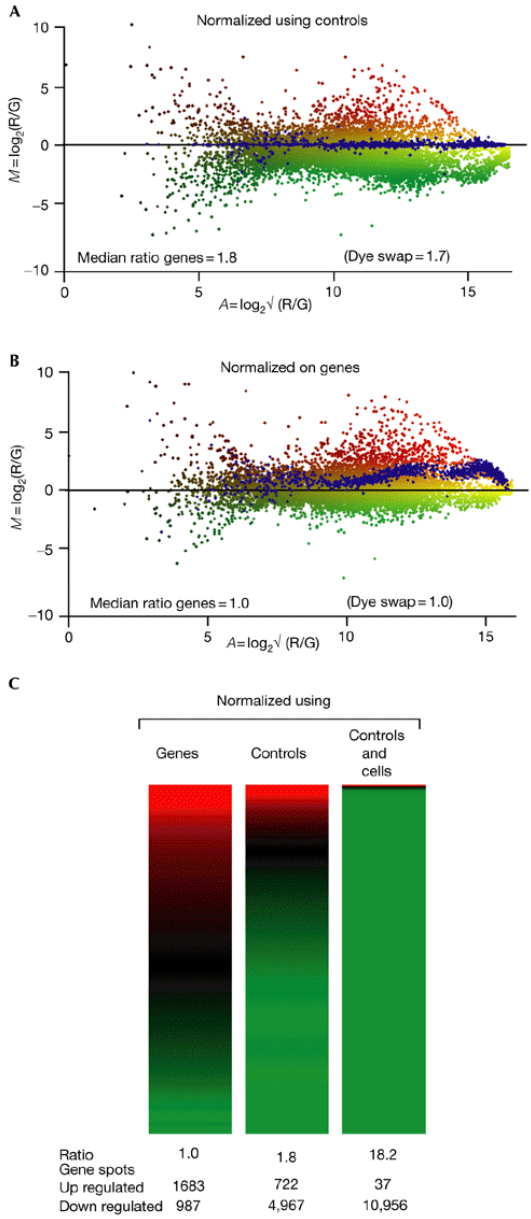


Figure 6.2: Unbalancing changes in global gene expression. Yeast stationary phase culture (R) compared with mid-log phase culture (G). A) MA scatter plot after LOWESS normalization using external controls. B) LOWESS normalization for using all genes, i.e. according to the GNA. The aberrant pattern of external control spots (blue) occurred because messenger RNA levels had not changed uniformly across the entire range of expression levels. C) Comparison of different normalization strategies. The left and middle columns correspond to the graphs shown in (B) and (A), respectively. The drop in total RNA yields per cell (see Methods) can also be taken into account (right column). Taken from van de Peppel *et al*., 2005 [208].

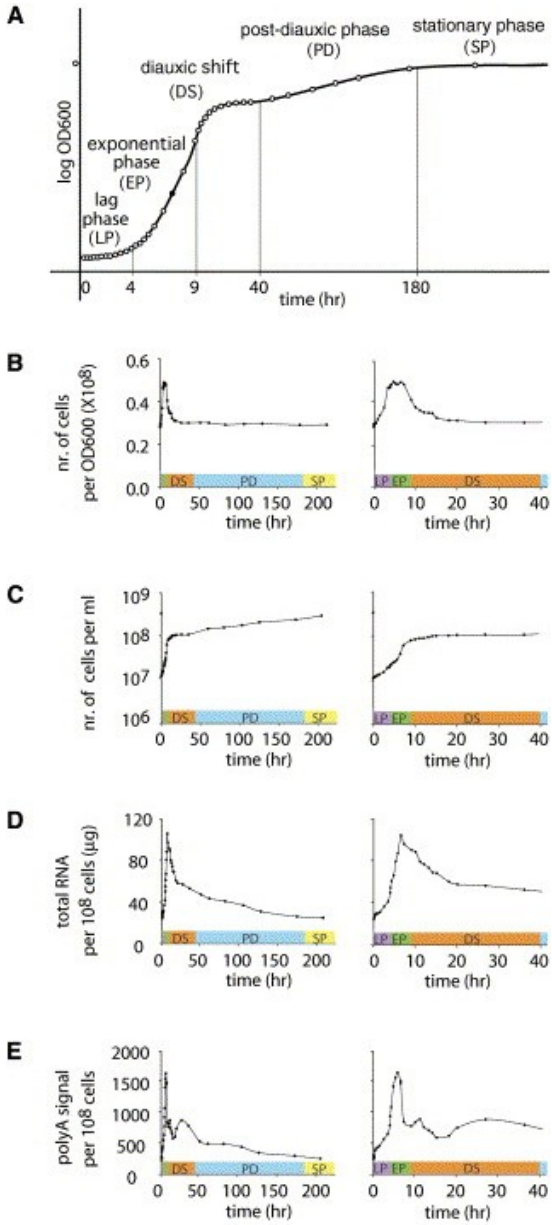


Figure 6.3: Cells, total RNA, and mRNA. A) Schematic representation of a glucose starvation experiment of *Saccharomyces cerevisiae* (circles represent the 39 samples that were analyzed). The OD_{600} is a measure of culture density. B) Cell count per OD_{600} throughout the culture (left) and for the first 40 hr (right). This gives an idea of the difference in average cell size. For an equal amount of cells, a bigger OD_{600} indicates a larger average cell size. The coloured bar represents the various culture periods, with abbreviations according to panel A. C) Number of cells/ml. D) Total RNA per 10^8 cells. E) Average polyA signal, an indication of the amount of mRNA being transcribed. Taken from Radjonic *et al*., 2005 [158].

Despite these issues, the use of external controls has some substantial advantages, not in the least due to its potential to completely avoid any assumptions relating to the distribution of gene expression values, i.e. there is no need for a GNA. The importance of this independence from the GNA should not be underestimated. A LOWESS fit that relies on all genes may be fairly robust for small numbers of outliers (i.e. when only a few genes are differentially expressed) [226], but it is important to stress that, when the GNA is violated, the normalized data would report faulty ratios for a great many genes. It has long been undisputed that the GNA is a 'safe' assumption, thought to be inappropriate only in the most extreme cases, i.e. when comparing radically different biological conditions. Circumstances where the GNA is inappropriate beyond a doubt expose themselves by showing unusually large, often increasing log-ratio variances for higher average intensity levels. Sadly enough in reality, such prominent features are ignored by researchers (some examples from published research articles [19,20,42,114] are shown in Figure 6.1), and data is normalized according to the GNA regardless. Even more worrying, it has recently been shown however that such global, unbalancing changes in gene expression can also occur under more conventional experimental conditions, and are in fact more common than what was previously believed [206-208] (an example is shown in Figure 6.2). The method we describe in chapter 5 could be a convenient tool for further addressing the validity of the GNA when comparing different biological conditions.

… and the eye of the beholder

The main purpose of any transcript profiling experiment is to reveal which genes alter their expression and to what extent. A seemingly trivial question comes to mind: "When is a gene considered differentially expressed?" Or even better: "When is a genes expression considered to be constant across different biological conditions?"

When microarray experiments are performed, usually equal amounts of Cy5 and Cy3 labelled target are hybridized to the array. Alternatively, this amount is varied according to the total fluorescence intensity of the samples. The reasoning behind this rescaling is that it would help alleviate intensity biases caused by differential labelling. Although intuitively plausible, in light of the distinct saturation characteristics of measured intensities from hybridized Cy3 and Cy5 samples, this remedial measure has little relevance. One could state that the general idea is to compare equal amounts of mRNA, regardless of their proportion to the total RNA in the cell. Indeed, a variety of processing steps lies between selecting a biological population of cells and hybridizing its labelled target to a microarray, ranging from isolation of total RNA, to separation, reverse transcribing and labelling of the mRNA (see chapter 2, section 2.1). When comparing merely the mRNA from different biological condition, one is completely oblivious to the fact that

total RNA samples contain different proportions or compositions of mRNA. There is nonetheless an easy way to monitor such differences, namely by spiking external control mixes into total RNA samples, thereby controlling all of the downstream steps. Of course any factor that might alter the initial ratios of spikes controls is completely confounded with the amounts of labelled target that are ultimately applied to the microarray. This brings us back to the point mentioned earlier: there is need for a proper format reporting all relevant experimental steps, including the amounts of hybridized target sample.

When treating differential expression as a unit per amount of mRNA, one does not only dismiss the composition of total RNA in a cell, but also the basic unit of life, the cell itself, is utterly ignored. What happens when one or more of the relevant biological conditions consists of cells in which RNA transcription is significantly impaired or has even grinded to a near halt (e.g. nutrition deprived cells)? How meaningful is it in such circumstances to compare expression levels, normalized with respect to amounts of mRNA or total RNA? A recently published research paper, of which some results are reproduced in Figure 6.3, illustrates these points perfectly [158]. Holstege and colleagues describe an expression profiling experiment of *Saccharomyces cerevisiae* quiescence entry and exit, including intervening events, that covered 9 days of culture. All the time points in this experiment were complemented with ChIP-chip data for the RNA polymerase II protein. Expression data were normalized by performing a LOWESS normalization using external control spikes, which were added to total RNA, thus avoiding GNA-like assumptions and more accurately determining mRNA level changes. Moreover, additional rescaling factors, based on the total RNA yield per cell, were taken into account in the analysis. By incorporating all of these factors into their research, they were able to reach conclusions that one might be hard pressed to find otherwise. They showed that transcription activity is shut down all but entirely in stationary phase, but that the transcription machinery is maintained, largely in an inactive state, and the RNA polymerase II is poised for immediate response by being held upstream of many genes required for changes in environmental conditions and reproliferation.

So how should we define the units to compare gene expression across different biological conditions: as quantities per cell, as quantities per total amount of RNA, as quantities per amount of mRNA, or as something entirely different? Perhaps the better strategy is to incorporate all of these factors into the analysis, as each one of them contains its own share of biologically relevant information. Perhaps it is more rewarding to compare biological conditions as a whole, instead of focussing on significantly differential genes. Studying the distribution of absolute expression levels for a single cellular state in its own right could lead to notable biological

insights. Which genes are expressed *tout court*, at what level, which pathways and functional categories are active in the cell and to what degree, how do they relate to the condition studied, and how (un)likely are all these observations compared to other expression profiles that are available for the same organism? Of course, such analysis would require accurate measurements of absolute mRNA levels across the entire expression range, from the highest activity of transcription to a complete lack of a genes expression. Technology is not yet up to the task, and –for now- differential expression will remain very much in the eye of the beholder.

Appendix A

Locally Weighted Scatter Plot Smoothing

Locally Weighted Scatter Plot Smoothing (LOWESS) is a statistical technique for plotting a smooth curve through a set of data points in a scatter plot (a graph with a predictor variable as its *X*-axis and a response variable as its *Y*-axis). The LOWESS fit procedure, originally proposed by Cleveland (1979) [39], combines much of the simplicity of linear least squares regression with the flexibility of nonlinear regression. At each point in the data set a low-degree polynomial is fit to a subset of the data (localized subset), with explanatory variable values near the point whose response is being estimated. The polynomial is fit using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The value of the regression function for the point is then obtained by evaluating the local polynomial using the predictor variable values for that data point. The LOWESS fit is complete after regression function values have been computed for all of the *n* data points. The procedure is illustrated in Figure A.1.

One of the chief attractions of the LOWESS method is that the data analyst is not required to specify a global function of any form to fit a model to the data. Several of the details and parameters of this method however, such as the degree of the polynomial model and the weights, are flexible. The three most important choices that are available to the user are briefly discussed below:

- **Degree of the local polynomials:** the local polynomials fit to each subset of the data are almost always of first or second degree; that is, either locally linear or locally quadratic. The original LOWESS algorithm relied on a linear polynomial as local regression function. LOESS (without 'w') differs in using a quadratic polynomial for its regression. In practice however, LOWESS and LOESS are often treated as synonyms. Using a zero degree polynomial would turn LOWESS into a weighted moving average.

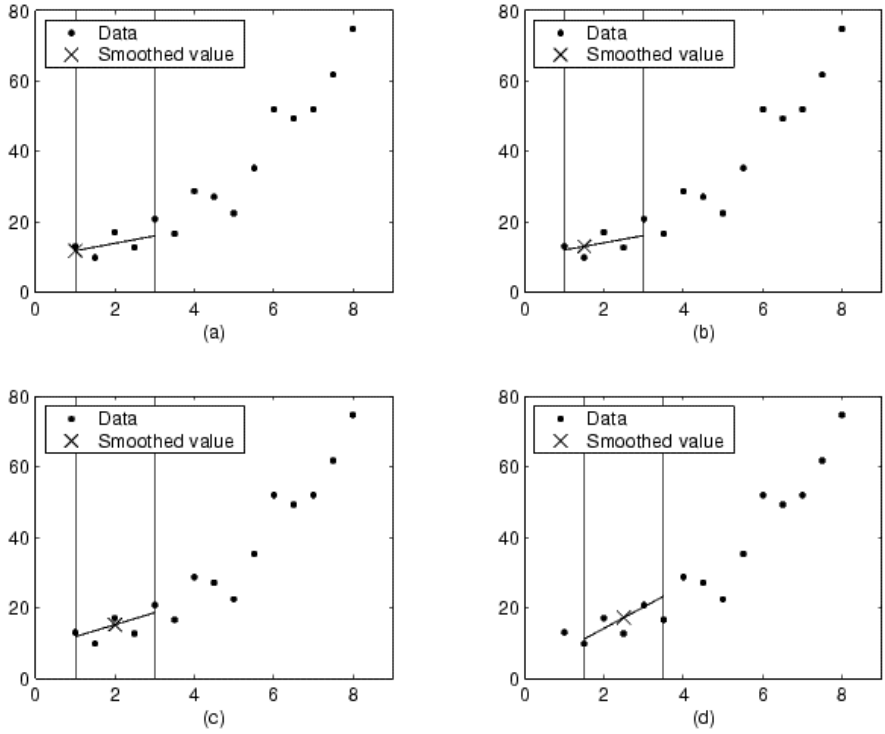


Figure A.1: Illustration of the LOWESS fit procedure. Plot (a) through (d) show the localized subset (vertical lines), local polynomial fit (diagonal lines within localized subset), and smoothed response value (cross) as the algorithm progresses from one data point to the next. The number of data points that are retained in a localized subset (determined by *f*) does not change as the smoothing process progresses from data point to data point. However, depending on the number of nearest neighbours, the regression weight function might not be symmetric about the data point to be smoothed. In particular, plots (a) and (b) use an asymmetric weight function, while plots (c) and (d) use a symmetric weight function. For the LOESS method, the graphs would look similar, except the smoothed value would be generated by a second degree polynomial. Taken from The MathWorks website (http://www.mathworks.com/access/helpdesk/help/toolbox/curvefit/ch_data7.html).

- **Smoothing parameter:** the smoothing parameter f is the proportion of data used in each fit. The value of f is a number between $(d+1)/n$ and 1, with d denoting the degree of the local polynomial. Large values of f produce the smoothest functions that are more robust in response to fluctuations in the data. The smaller f is, the closer the regression function will conform to the data variations.
- **Weight function:** the traditional weight function used for LOWESS is the tri-cube weight function.

$$w_i(x) = \begin{cases} (1 - f_i(x))^3 & \text{for} \quad f_i(x) \le 1 \\ 0 & \text{for} \quad f_i(x) > 1 \end{cases}$$

where

$$f_i(x) = \left| \frac{x - x_i}{\Delta_x} \right|^3 \tag{A.1}$$

In this formulation x is the predictor value associated with the response value to be smoothed, x_i are the nearest neighbours of x as defined by the span (dependent on f), and Δ_x is the distance along the abscissa from x to the most distant predictor value within the span. The weights have the characteristics that the data point to be smoothed has the largest weight and the most influence on the fit and that data points outside the span have zero weight and no influence on the fit.

Appendix B

Analysis of Variance

This appendix serves to give a short overview of the principles of analysis of variance techniques and the notation, as used in this dissertation, to describe the models as well as the procedures for estimating the model parameters. For more detailed information we refer to the book 'Linear statistical models' of Neter *et al.* (1996) [141].

B.1 Principles

Analysis of variance (ANOVA) models are versatile tools for studying the relation between a response variable and one or more explanatory or predictor variables (referred to as *factors*). ANOVA models are a basic type of linear statistical models that share many features with regression models. Like regression models, they are concerned with the statistical relation between one or more predictor variables and a single explanatory variable, the latter of which the nature is always quantitative. ANOVA models however, differ from ordinary regression models in two key respects:

- The explanatory or predictor variables in ANOVA models are often qualitative (manufacturing type, gender, location, etc.). In ordinary regression models, both the predictor and response variable are quantitative.
- If the predictor variables are quantitative, no assumption is made in ANOVA models about the nature of the statistical relation between them and the response variable. Thus, the need to specify the nature of the regression function encountered in ordinary regression analysis does not arise in ANOVA models. Instead, every particular form of a predictor variable (such an instance of a factor is referred to as *factor level*) is attributed with its own parameter.

B.2 Notation

ANOVA models

We will illustrate the notation of the ANOVA models and estimation of the parameters with a simple two-factor study (i.e. two predictor variables). When multiple factors are involved, a single combination of factor levels is referred to as a *treatment*. The mean response for a given treatment in a two factor study can be referred to as μ_{ij}, where i refers to the level of factor A ($i = 1,\ldots,a$) and j refers to the level of factor B ($j = 1,\ldots,b$):

$$\mu_{ij} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} \tag{B.1}$$

This formulation indicates that each μ_{ij} can be viewed as the sum of four component factor effect parameters. Specifically, (B.1) states that the mean response for the treatment where factor A is at the i^{th} level and factor B is at the j^{th} level is the sum of an overall constant μ, the main effect parameter α_i for factor A at the i^{th} level, the main effect parameter β_j for factor B at the j^{th} level, and the interaction effect parameter $\alpha\beta_{ij}$. The interaction effect parameter is the difference between the treatment mean μ_{ij} and the value that would be expected if the factors were additive. If in fact the two factors are additive, all interactions equal zero, i.e. $\alpha\beta_{ij} \equiv 0$.

It is further assumed that the measured responses for each treatment are random selections from a normal distribution with the same variance, and that they are independent from the measured responses for any other factor. As such, we can write:

$$Y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk} \tag{B.2}$$

Where Y_{ijk} is the response value of the k^{th} trial, and n_{ij} the sample size ($k = 1,\ldots,n_{ij}$), for the treatment where factor A is at the i^{th} level, factor B is at the j^{th} level, and the error terms ε_{ijk} are independent $N(0,\sigma)$.

Parameter estimation

Obtaining estimators for the parameters in ANOVA model (B.2) is usually done through least squares or maximum likelihood methods. Both lead to minimizing the residual sum of squares (*SSE*):

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} e_{ijk}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \hat{Y}_{ijk}\right)^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \mu - \alpha_i - \beta_j - \alpha\beta_{ij}\right)^2 \tag{B.3}$$

subject to restrictions for the main effects:

$$\sum_{i=1}^{a} \alpha_i = 0$$

$$\sum_{j=1}^{b} \beta_j = 0 \qquad \text{(B.4)}$$

and subject to restrictions for the interaction effects:

$$\sum_{i=1}^{a} \alpha\beta_{ij} = 0$$

$$\sum_{j=1}^{b} \alpha\beta_{ij} = 0 \qquad \text{(B.5)}$$

The $\hat{Y}_{ijk}$ represent the *fitted values*, i.e. the linear combination of parameters for the corresponding treatment, and e_{ijk} represent the *residuals*, which are defined as the difference between the observed and the fitted values.

When this minimization is performed (which can easily be done analytically by setting the partial derivatives of the *SSE* with respect to each of the parameters to zero, and solving this system of equations), the following estimators for the parameters are obtained:

$$\hat{\mu} = Y_{...}$$

$$\hat{\alpha}_i = Y_{i..} - Y_{...}$$

$$\hat{\beta}_j = Y_{.j.} - Y_{...}$$

$$\alpha\beta_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...} \qquad \text{(B.6)}$$

where a dotted index '.' indicates to average the measurements of the response variable over that index.

Matrix notation

ANOVA models are linear models because they can be stated in the following form (regression model approach):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{(B.7)}$$

The vector $\mathbf{Y}$ is of size $n \times 1$ containing the total of n observations on the response variable. Similarly, the vector $\boldsymbol{\varepsilon}$ is of size $n \times 1$ and represents the error terms. The predictor matrix is of size $n \times p$, and parameter vector $\boldsymbol{\beta}$ is of size $p \times 1$. It should be noted that p is not the total amount of parameters in the model, but the degrees of freedom that are associated with the parameters. In the case of model (B.2), with restrictions (B.4) and (B.5), this

amounts to $p=(a\text{-}1)+(b\text{-}1)+(a\text{-}1)(b\text{-}1)$. Indeed, due to restrictions (B.4), we only need a-1 parameters α_i and b-1 parameters β_j in the regression model, and we can represent α_a and β_b as:

$$\alpha_a = -\alpha_1 - \alpha_2 - \ldots - \alpha_{a-1}$$
$$\beta_b = -\beta_1 - \beta_2 - \ldots - \beta_{b-1} \qquad \text{(B.8)}$$

Similarly for the interaction parameters, if we recognize the restrictions (B.5), we only need $(a\text{-}1)(b\text{-}1)$ parameters $\alpha\beta_{ij}$ in the regression model, and can represent $\alpha\beta_{aj}$ and $\alpha\beta_{ib}$ as:

$$\alpha\beta_{ib} = -\alpha\beta_{i1} - \alpha\beta_{i2} - \ldots - \alpha\beta_{i,b-1} \qquad i = 1,\ldots,a$$
$$\alpha\beta_{aj} = -\alpha\beta_{1j} - \alpha\beta_{2j} - \ldots - \alpha\beta_{a-1,j} \qquad j = 1,\ldots,b \qquad \text{(B.9)}$$

The elements of the predictor matrix **X** can only be 0, 1 or -1. A 1 or 0 indicates whether a factor effects parameter is applicable to an observed response variable or not respectively. The -1 entries of **X** are meant to account for equations (B.8) and (B.9). Matrix X is known as it corresponds to the design of the experiment. A least squares estimator **b** for the parameter vector $\boldsymbol{\beta}$ can be obtained by solving the linear equation (B.7) as:

$$\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \qquad \text{(B.10)}$$

Bibliography

1. Adams R, Bischof L: Seeded Region Growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1994, 16: 641-647.
2. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Research* 2003, 31: 1753-1764.
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A *et al.*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000, 403: 503-511.
4. Allemeersch J, Durinck S, Vanderhaeghen R, Alard P, Maes R, Seeuws K *et al.*: Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. *Plant Physiology* 2005, 137: 588-601.
5. Alwine JC, Kemp DJ, Stark GR: Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America* 1977, 74: 5350-5354.
6. Anderson S, Bankier AT, Barrell BG, Debruijn MHL, Coulson AR, Drouin J *et al.*: Sequence and organization of the human mitochondrial genome. *Nature* 1981, 290: 457-465.
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM *et al.*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 2000, 25: 25-29.
8. Aviv H, Leder P: Purification of biologically-active globin messenger-RNA by chromatography on oligothymidylic-acid-cellulose. *Proceedings of the National Academy of Sciences of the United States of America* 1972, 69: 1408-1412.
9. Bader GD, Heilbut A, Andrews B, Tyers M, Hughes T, Boone C: Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends in Cellular Biology* 2003, 13: 344-356.

10. Badiee A, Eiken HG, Steen VM, Lovlie R: Evaluation of five different cDNA labeling methods for microarrays using spike controls. *BMC Biotechnology* 2003, 3: 23.

11. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J *et al.*: The External RNA Controls Consortium: a progress report. *Nature Methods* 2005, 2: 731-734.

12. Baldi P, Long AD: A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 2001, 17: 509-519.

13. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU *et al.*: Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods* 2005, 2: 351-356.

14. Bartosiewicz M, Trounstine M, Barker D, Johnston R, Buckpitt A: Development of a toxicological gene array and quantitative assessment of this technology. *Archives of Biochemics and Biophysics* 2000, 376: 66-73.

15. Ben Dor A, Shamir R, Yakhini Z: Clustering gene expression patterns. *Journal of Computational Biology* 1999, 6: 281-297.

16. Benes V, Muckenthaler M: Standardization of protocols in cDNA microarray analysis. *Trends in Biochemical Sciences* 2003, 28: 244-249.

17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. *Nucleic Acids Research* 2005, 33: D34-D38.

18. Berk AJ, Sharp PA: Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* 1977, 12: 721-732.

19. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN: Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 2002, 99: 9697-9702.

20. Bernstein JA, Lin PH, Cohen SN, Lin-Chao S: Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 2004, 101: 2758-2763.

21. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S *et al.*: High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research* 2004, 14: 287-295.

22. Bilban M, Buehler LK, Head S, Desoye G, Quaranta V: Normalizing DNA microarray data. *Current Issues in Molecular Biology* 2002, 4: 57-64.

23. Bishop CM: *Neural Networks for Pattern Recognition*. New York: Oxford University Press; 1995.

24. Bishop JO, Morton JG, Rosbash M, Richards M: 3 Abundance classes in HeLa cell messenger-RNA. *Nature* 1974, 250: 199-204.

25. Blais A, Dynlacht BD: Constructing transcriptional regulatory networks. *Genes and Development* 2005, 19: 1499-1511.

26. Boguski MS, Lowe TMJ, Tolstoshev CM: Dbest - Database for Expressed Sequence Tags. *Nature Genetics* 1993, 4: 332-333.

27. Bowtell D, Sambrook J: *DNA microarrays: A Molecular Cloning Manual.* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2002.

28. Boyd KE, Farnham PJ: Coexamination of site-specific transcription factor binding and promoter activity in living cells. *Molecular and Cellular Biology* 1999, 19: 8393-8399.

29. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C *et al.*: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 2001, 29: 365-371.

30. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N *et al.*: ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* 2003, 31: 68-71.

31. Brazma A, Vilo J: Gene expression data analysis. *FEBS Letters* 2000, 480: 17-24.

32. Brown CS, Goodwin PC, Sorger PK: Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98: 8944-8949.

33. Carter MG, Sharov AA, VanBuren V, Dudekula DB, Carmack CE, Nelson C *et al.*: Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biology* 2005, 6: R61.

34. Chen Y, Dougherty ER, Bittner M: Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 1997, 2: 364-374.

35. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L *et al.*: A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 1998, 2: 65-73.

36. Chua G, Robinson MD, Morris Q, Hughes TR: Transcriptional networks: reverse-engineering gene regulation on a global scale. *Current Opinion in Microbiology* 2004, 7: 638-646.

37. Chuang SE, Daniels DL, Blattner FR: Global Regulation of Gene-Expression in Escherichia Coli. *Journal of Bacteriology* 1993, 175: 2026-2036.

38. Churchill GA: Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* 2002, 32 Suppl:490-5.: 490-495.

39. Cleveland WS: Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Society* 1979, 74: 829-836.

40. Cochran WG, Cox GM: *Experimental Designs*. New York: Wiley; 1992.

41. Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, Engelen K *et al.*: INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Research* 2003, 31: 3468-3470.

42. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC: Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli. *Genetics* 2001, 158: 41-64.

43. De Bie T, Monsieurs P, Engelen K, De Moor B, Cristianini N, Marchal K: Discovering transcriptional modules from motif, chip-chip and microarray data. *Pacific Symposium on Biocomputing* 2005, 483-494.

44. De Keersmaecker SC, Marchal K, Verhoeven TL, Engelen K, Vanderleyden J, Detweiler CS: Microarray analysis and motif detection reveal new targets of the Salmonella enterica serovar Typhimurium HilA regulatory protein, including hilA itself. *Journal of Bacteriology* 2005, 187: 4381-4391.

45. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y: Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* 2002, 18:7357-7346.

46. De Smet F, Moreau Y, Engelen K, Timmerman D, Vergote I, De Moor B: Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer* 2004, 91: 1160-1165.

47. De Smet F, Pochet N, Engelen K, Van Gorp T, Van Hummelen P, Marchal K *et al.*: Predicting the clinical behavior of ovarian cancer from gene expression profiles. *International Journal of Gynecological Cancer* 2006, Accepted for publication.

48. Denolet E, De Gendt K, Allemeersch J, Engelen K, Marchal K, Van Hummelen P *et al.*: The effect of a Sertoli cell-selective knockout of the androgen receptor on testicular gene expression in prepubertal mice. *Molecular Endocrinology* 2005, Accepted to appear in February 2006.

49. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M *et al.*: Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* 1996, 14: 457-460.

50. DiCiccio TJ, Efron B: Bootstrap confidence intervals. *Statistical Science* 1996, 11: 189-228.

51. Dobbin K, Shih JH, Simon R: Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *Journal of the National Cancer Institute* 2003, 95: 1362-1369.

52. Dobbin K, Shih JH, Simon R: Statistical design of reverse dye microarrays. *Bioinformatics* 2003, 19: 803-810.

53. Dobbin KK, Kawasaki ES, Petersen DW, Simon RM: Characterizing dye bias in microarray experiments. *Bioinformatics* 2005, 21: 2430-2437.

54. Dobbin KK, Shih JH, Simon RM: Comment on 'Evaluation of the gene-specific dye bias in cDNA microarray experiments'. *Bioinformatics* 2005, 21: 2803-2804.

55. Dombkowski AA, Thibodeau BJ, Starcevic SL, Novak RF: Gene-specific dye bias in microarray reference designs. *FEBS Letters* 2004, 560: 120-124.

56. Dorris DR, Ramakrishnan R, Trakas D, Dudzik F, Belval R, Zhao C *et al.*: A highly reproducible, linear, and automated sample preparation method for DNA microarrays. *Genome Research* 2002, 12: 976-984.

57. Dudley AM, Aach J, Steffen MA, Church GM: Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proceedings of the National Academy of Sciences of the United States of America* 2002, 99: 7554-7559.

58. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. 578, 1-38. 2000. Stanford University, Internal Report.

59. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nature Genetics* 1999, 21: 10-14.

60. Dumanoir S, Speicher MR, Joos S, Schrock E, Popp S, Dohner H *et al.*: Detection of complete and partial chromosome gains and losses by Comparative Genomic Insitu Hybridization. *Human Genetics* 1993, 90: 590-610.

61. Durbin B, Rocke DM: Estimation of transformation parameters for microarray data. *Bioinformatics* 2003, 19: 1360-1367.

62. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 2002, 18 Suppl 1: S105-S110.

63. Durbin BP, Rocke DM: Variance-stabilizing transformations for two-color microarrays. *Bioinformatics* 2004, 20: 660-667.

64. Eberwine J: Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. *Biotechniques* 1996, 20: 584-591.

65. Eberwine J, Spencer C, Miyashiro K, Mackler S, Finnell R: Complementary DNA synthesis in situ: methods and applications. *Methods in Enzymology* 1992, 216: 80-100.

66. Edwards D: Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics* 2003, 19: 825-833.

67. Efron B: Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 1979.

68. Efron B, Tibshirani R: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986, 54-77.

69. Ehrenberg M, Elf J, Aurell E, Sandberg R, Tegner J: Systems biology is taking off. *Genome Research* 2003, 13: 2377-2380.

70. Eickhoff B, Korn B, Schick M, Poustka A, van der BJ: Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Research* 1999, 27: e33.

71. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 1998, 95: 14863-14868.

72. Engelen K, Coessens B, Marchal K, De Moor B: MARAN: normalizing micro-array data. *Bioinformatics* 2003, 19: 893-894.

73. Engelen K, Naudts B, De Moor B, Marchal K: A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics* 2006, Accepted for publication.

74. Finkelstein DB, Gollub J, Ewing R, Sterky F, Somerville S, Cherry JM: *Iterative linear regression by sector: Renormalization of cDNA microarray data and cluster analysis weighted by cross homology*. In: CAMDA'00 (Critical Assessment of Microarray Data Analysis Techniques).

75. Fritz B, Schubert F, Wrobel G, Schwaenen C, Wessendorf S, Nessling M *et al.*: Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. *Cancer Research* 2002, 62: 2993-2998.

76. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S *et al.*: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 2004, 5.

77. Getz MJ, Birnie GD, Young BD, Macphail E, Paul J: Kinetic estimation of base sequence complexity of nuclear poly(A)-containing RNA in mouse friend cells. *Cell* 1975, 4: 121-129.

78. Getz MJ, Reiman HM, Siegal GP, Quinlan TJ, Proper J, Elder PK *et al.*: Gene-expression in chemically transformed mouse embryo cells - Selective enhancement of expression of C-Type RNA tumor-virus genes. *Cell* 1977, 11: 909-921.

79. Ghosh D, Chinnaiyan AM: Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* 2002, 18: 275-286.

80. Girke T, Todd J, Ruuska S, White J, Benning C, Ohlrogge J: Microarray analysis of developing Arabidopsis seeds. *Plant Physiology* 2000, 124: 1570-1581.

81. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999, 286: 531-537.

82. Goryachev AB, MacGregor PF, Edwards AM: Unfolding of microarray data. *Journal of Computational Biology* 2001, 8: 443-461.

83. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M: Interrelating different types of genomic data, from proteome to secretome: 'Oming in on function. *Genome Research* 2001, 11: 1463-1468.

84. Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H: Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mammalian Genome* 1992, 3: 609-619.

85. Halliwell CM, Cass AEG: A factorial analysis of silanization conditions for the immobilization of oligonucleotides on glass surfaces. *Analytical Chemistry* 2001, 73: 2476-2483.

86. Harrington CA, Rosenow C, Retief J: Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology* 2000, 3: 285-291.

87. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L *et al.*: 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 2000, 1: RESEARCH0003.

88. He Z, Wu L, Fields MW, Zhou J: Use of microarrays with different probe sizes for monitoring gene expression. *Applied Environmental Microbiology* 2005, 71: 5154-5162.

89. Hekstra D, Taussig AR, Magnasco M, Naef F: Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research* 2003, 31: 1962-1968.

90. Held GA, Grinstein G, Tu Y: Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the National Academy of Sciences of the United States of America* 2003, 100: 7575-7580.

91. Heller MJ: DNA microarray technology: Devices, systems, and applications. *Annual Review of Biomedical Engineering* 2002, 4: 129-153.

92. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J *et al.*: Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 1997, 94: 2150-2155.

93. Hemminki A, Tomlinson I, Markie D, Jarvinen H, Sistonen P, Bjorkqvist AM *et al.*: Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nature Genetics* 1997, 15: 87-90.

94. Herrero J, Valencia A, Dopazo J: A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 2001, 17: 126-136.

95. Heyer LJ, Kruglyak S, Yooseph S: Exploring expression data: identification and analysis of coexpressed genes. *Genome Research* 1999, 9: 1106-1115.

96. Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J *et al.*: Versatile gene-specific sequence tags for Arabidopsis functional genomics:

transcript profiling and reverse genetics applications. *Genome Research* 2004, 14: 2176-2189.

97. Hodgson G, Hager JH, Volik S, Hariono S, Wernick M, Moore D *et al.*: Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* 2001, 29: 459-464.

98. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002, 18 Suppl 1: S96-S104.

99. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW *et al.*: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* 2001, 19: 342-347.

100. Hughes TR, Shoemaker DD: DNA microarrays for expression profiling. *Current Opinion in Chemical Biology* 2001, 5: 21-25.

101. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC *et al.*: Multiple-laboratory comparison of microarray platforms. *Nature Methods* 2005, 2: 345-350.

102. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001, 409: 533-538.

103. Jansen E, Petit MMR, Schoenmakers EFPM, Ayoubi T, Van de Ven WJM: High mobility group protein HMGI-C: a molecular target in solid tumor formation. *Gene Therapy and Molecular Biology* 1999, 3: 387-395.

104. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* 2001, 29: 389-395.

105. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F *et al.*: Comparative Genomic Hybridization for molecular cytogenetic analysis of solid tumors. *Science* 1992, 258: 818-821.

106. Kallioniemi OP, Kallioniemi A, Sudar D, Rutovitz D, Gray JW, Waldman F *et al.*: Comparative Genomic Hybridization - A rapid new method for detecting and mapping DNA amplification in tumors. *Seminars in Cancer Biology* 1993, 4: 41-46.

107. Kepler TB, Crosby L, Morgan KT: Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* 2002, 3: RESEARCH0037.

108. Kerr MK: Linear models for microarray data analysis: hidden similarities and differences. *Journal of Computational Biology* 2003, 10: 891-901.

109. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ *et al.*: Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* 2002, 12: 203-218.

110. Kerr MK, Churchill GA: Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98: 8961-8965.

111. Kerr MK, Churchill GA: Statistical design and the analysis of gene expression microarray data. *Genetical Research* 2001, 77: 123-128.

112. Kerr MK, Churchill GA: Experimental design for gene expression microarrays. *Biostatistics* 2001, 2: 183-201.

113. Kerr MK, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 2000, 7: 819-837.

114. Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C: DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* 2000, 97: 12170-12175.

115. Kim JH, Shin DM, Lee YS: Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles. *Experimental Molecular Medicine* 2002, 34: 224-232.

116. Kohonen T: *Self-Organizing maps*. Berlin, Germany: Springer-Verlag; 1997.

117. Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T: Improved background correction for spotted DNA microarrays. *Journal of Computational Biology* 2002, 9: 55-66.

118. Kroll TC, Wolfl S: Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic Acids Research* 2002, 30: e50.

119. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.*: Initial sequencing and analysis of the human genome. *Nature* 2001, 409: 860-921.

120. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: Independence and reproducibility across microarray platforms. *Nature Methods* 2005, 2: 337-344.

121. Lashkari DA, Derisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY *et al.*: Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* 1997, 94: 13057-13062.

122. Lennon GG, Lehrach H: Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics* 1991, 7: 314-317.

123. Leung YF, Cavalieri D: Fundamentals of cDNA microarray data analysis. *Trends in Genetics* 2003, 19: 649-659.

124. Lewin B.: *Genes VI*. New York: Oxford University Press; 1997.

125. Lieb JD, Liu X, Botstein D, Brown PO: Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics* 2001, 28: 327-334.

126. Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P: Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *Journal of Biological Chemistry* 2001, 276: 19937-19944.

127. Lukashin AV, Fuchs R: Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* 2001, 17: 405-414.

128. Lyng H, Badiee A, Svendsrud DH, Hovig E, Myklebost O, Stokke T: Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. *BMC Genomics* 2004, 5: 10.

129. Maier E, Meierewert S, Ahmadi AR, Curtis J, Lehrach H: Application of Robotic Technology to Automated Sequence Fingerprint Analysis by Oligonucleotide Hybridization. *Journal of Biotechnology* 1994, 35: 191-203.

130. Majtan T, Bukovska G, Timko J: DNA microarrays--techniques and applications in microbial systems. *Folia Microbiologica* 2004, 49: 635-664.

131. Marchal K, De Smet F, Engelen K, De Moor B: *Computational biology and toxicogenomics.* In Predictive toxicology. Edited by Helma C. M. Dekker; 2004.

132. Marchal K, Engelen K, De Brabanter J, Aert S, De Moor B, Ayoubi T *et al.*: Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays. *Journal of Biological Systems* 2002, 10: 409-430.

133. Martin-Magniette ML, Aubert J, Cabannes E, Daudin JJ: Answer to the comments of K. Dobbin, J. Shih and R. Simon on the paper 'Evaluation of the gene-specific dye-bias in cDNA microarray experiments'. *Bioinformatics* 2005, 21: 3065.

134. Martin-Magniette ML, Aubert J, Cabannes E, Daudin JJ: Evaluation of the gene-specific dye bias in cDNA microarray experiments. *Bioinformatics* 2005, 21: 1995-2000.

135. Martinez MJ, Aragon AD, Rodriguez AL, Weber JM, Timlin JA, Sinclair MB *et al.*: Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays. *Nucleic Acids Research* 2003, 31: e18.

136. Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, Carninci P *et al.*: Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA

arrays. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98: 2199-2204.

137. Monni O, Joensuu H, Franssila K, Knuutila S: DNA copy number changes in diffuse large B-cell lymphoma - Comparative genomic hybridization study. *Blood* 1996, 87: 5269-5278.

138. Moreau Y, De Smet F, Thijs G, Marchal K, De Moor B: Functional bioinformatics of microarray data: from expression to regulation. *IEEE Proceedings* 2002, 30: 1722-1743.

139. Nadon R, Shoemaker J: Statistical issues with microarrays: processing and analysis. *Trends in Genetics* 2002, 18: 265-271.

140. Nakazato H, Edmonds M: Isolation and purification of rapidly labeled polysome-bound ribonucleic-acid on polythymidylate cellulose. *Journal of Biological Chemistry* 1972, 247: 3365-3367.

141. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W: *Applied linear statistical models*. IRWIN, The McGraw-Hill Companies, Inc.; 1996.

142. Oleksiak MF, Churchill GA, Crawford DL: Variation in gene expression within and among natural populations. *Nature Genetics* 2002, 32: 261-266.

143. Orlando V, Paro R: Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell* 1993, 75: 1187-1198.

144. Pan W: A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 2002, 18: 546-554.

145. Panchuk-Voloshina N, Haugland RP, Bishop-Stewart J, Bhalgat MK, Millard PJ, Mao F *et al.*: Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *Journal of Histochemistry and Cytochemistry* 1999, 47: 1179-1188.

146. Papin JA, Hunter T, Palsson BO, Subramaniam S: Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology* 2005, 6: 99-111.

147. Park PJ, Pagano M, Bonetti M: A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pacific Symposium on Biocomputing* 2001, :52-63.: 52-63.

148. Patton WF: Detection technologies in proteome analysis. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 2002, 771: 3-31.

149. Patton WF, Schulenberg B, Steinberg TH: Two-dimensional gel electrophoresis; better than a poke in the ICAT? *Current Opinion in Biotechnology* 2002, 13: 321-328.

150. Peterson AW, Heaton RJ, Georgiadis RM: The effect of surface probe density on DNA hybridization. *Nucleic Acids Research* 2001, 29: 5163-5168.

151. Phillips J, Eberwine JH: Antisense RNA Amplification: A linear amplification method for analyzing the mRNA population from single living cells. *Methods* 1996, 10: 283-288.

152. Phimister B: Going global. *Nature Genetics* 1999, 21: 1.

153. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D *et al.*: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* 1998, 20: 207-211.

154. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF *et al.*: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* 1999, 23: 41-46.

155. Puskas LG, Zvara A, Hackler LJ, Van hummelen P: RNA amplification results in reproducible microarray data with slight ratio biases. *Biotechniques* 2002, 2: 1330-340.

156. Quackenbush J: Computational analysis of microarray data. *Nature Reviews Genetics* 2001, 2: 418-427.

157. Quackenbush J: Microarray data normalization and transformation. *Nature Genetics* 2002, 32 Suppl.: 496-501.

158. Radonjic M, Andrau JC, Lijnzaad P, Kemmeren P, Kockelkorn TT, van Leenen D *et al.*: Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Molecular Cell* 2005, 18: 171-183.

159. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A *et al.*: An assessment of Motorola CodeLink (TM) microarray performance for gene expression profiling applications. *Nucleic Acids Research* 2002, 30: e30.

160. Ramdas L, Coombes KR, Baggerly K, Abruzzo L, Highsmith WE, Krogmann T *et al.*: Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology* 2001, 2: RESEARCH0047.

161. Ramdas L, Wang J, Hu L, Cogdell D, Taylor E, Zhang W: Comparative evaluation of laser-based microarray scanners. *Biotechniques* 2001, 31: 546-550.

162. Randolph JB, Waggoner AS: Stability, specificity and fluorescence brightness of multiply-labeled fluorescent DNA probes. *Nucleic Acids Research* 1997, 25: 2923-2929.

163. Reid JL, Iyer VR, Brown PO, Struhl K: Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Molecular Cell* 2000, 6: 1297-1307.

164. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I *et al.*: Genome-wide location and function of DNA binding proteins. *Science* 2000, 290: 2306-2309.

165. Rocke DM, Durbin B: A model for measurement error for gene expression arrays. *Journal of Computational Biology* 2001, 8: 557-569.

166. Rocke DM, Durbin B: Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* 2003, 19: 966-972.

167. Rocke DM, Lorenzato S: A 2-Component Model for Measurement Error in Analytical-Chemistry. *Technometrics* 1995, 37: 176-184.

168. Roosen J, Engelen K, Marchal K, Mathys J, Griffioen G, Cameroni E *et al.*: PKA and Sch9 control a molecular switch important for the proper adaptation to nutrient availability. *Molecular Microbiology* 2005, 55: 862-880.

169. Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA: Analysis of enzymatically amplified beta-globin and Hla-Dq-Alpha DNA with allele-specific oligonucleotide probes. *Nature* 1986, 324: 163-166.

170. Sanger F, Coulson AR: Rapid method for determining sequences in DNA by primed synthesis with DNA-polymerase. *Journal of Molecular Biology* 1975, 94: 441-448.

171. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL *et al.*: Nucleotide-sequence of bacteriophage-Phi-X174. *Journal of Molecular Biology* 1978, 125: 225-246.

172. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB: Nucleotide-sequence of bacteriophage-Gamma DNA. *Journal of Molecular Biology* 1982, 162: 729-773.

173. Schadt EE, Li C, Ellis B, Wong WH: Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* 2001, Suppl 37: 120-125.

174. Schena M, Davis W: *Technology standards for microarray research. In Microarray biochip technology*. Edited by Schena M. Natick, Massechusetts: Eaton Publishing, Biotechniques Book Division; 2000:1-18.

175. Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science* 1995, 270: 467-470.

176. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America* 1996, 93: 10614-10619.

177. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H *et al.*: Normalization strategies for cDNA microarrays. *Nucleic Acids Research* 2000, 28: e47.

178. Shalon D, Smith SJ, Brown PO: A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research* 1996, 6: 639-645.

179. Sharan R, Shamir R: CLICK: a clustering algorithm with applications to gene expression analysis. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* 2000, 8: 307-316.

180. Shchepinov MS, CaseGreen SC, Southern EM: Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Research* 1997, 25: 1155-1161.

181. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD *et al.*: Experimental annotation of the human genome using microarray technology. *Nature* 2001, 409: 922-927.

182. Simon R, Radmacher MD, Dobbin K: Design of studies using DNA microarrays. *Genetic Epidemiology* 2002, 23: 21-36.

183. Skoog DA, West DM, Holler FJ: *Fundamentals of Analytical Chemistry*. Philadelphia: Saunders College Publishing; 1996.

184. Soille P: *Morphological image analysis: principles and applications*. Berlin: Springer; 1999.

185. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H *et al.*: Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* 1997, 20: 399-407.

186. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H *et al.*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98: 10869-10874.

187. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB *et al.*: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* 1998, 9: 3273-3297.

188. Staden R: Sequence data handling by computer. *Nucleic Acids Research* 1977, 4: 4037-4051.

189. Stillman BA, Tonkinson JL: Expression microarray hybridization kinetics depend on length of the immobilized DNA but are independent of immobilization substrate. *Analytical Biochemistry* 2001, 295: 149-157.

190. Strahl-Bolsinger S, Hecht A, Luo K, Grunstein M: SIR2 and SIR4 interactions differ in core and extended telomeric heterochromatin in yeast. *Genes and Development* 1997, 11: 83-93.

191. Takahashi N, Ko MSH: Toward A whole cDNA catalog - Construction of an equalized cDNA library from mouse embryos. *Genomics* 1994, 23: 202-210.

192. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E *et al.*: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* 1999, 96: 2907-2912.

193. Taniguchi M, Miura K, Iwao H, Yamanaka S: Quantitative assessment of DNA microarrays--comparison with Northern blot analyses. *Genomics* 2001, 71: 34-39.

194. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nature Genetics* 1999, 22: 281-285.

195. Thijs G: Probabilistic methods to search for regulatory elements in sets of coregulated genes. PhD thesis, Faculty of Applied Sciences, K.U.Leuven: 2003.

196. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P *et al.*: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 2001, 17: 1113-1122.

197. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P *et al.*: A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology* 2002, 9: 447-464.

198. Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S *et al.*: INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics* 2002, 18: 331-332.

199. Thomas JG, Olson JM, Tapscott SJ, Zhao LP: An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* 2001, 11: 1227-1236.

200. Tou JT, Gonzalez RC: *Pattern classification by distance functions. In Pattern recognition principles*. Adison-Wesley; 1979:75-109.

201. Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP, Cho KWY: Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Research* 2002, 30.

202. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R *et al.*: Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001, 17: 520-525.

203. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* 2001, 29: 2549-2557.

204. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98: 5116-5121.

205. Unlu M: Difference gel electrophoresis. *Biochemical Society Transactions* 1999, 27: 547-549.

206. van Bakel H, Holstege FC: In control: systematic assessment of microarray performance. *EMBO Reports* 2004, 5: 964-969.

207. van Berkum NL, Holstege FC: DNA microarrays: raising the profile. *Current Opinion in Biotechnology* 2001, 12: 48-52.

208. van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC: Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports* 2003, 4: 387-393.

209. Van den Bergh G, Arckens L: Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics. *Current Opinion in Biotechnology* 2004, 15: 38-43.

210. Van den Bulcke T, Lemmens K, Van de Peer Y, Marchal K: Inferring transcriptional networks by mining 'omics´ data. *Current Bioinformatics* 2006, Accepted to appear in January 2006.

211. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH: Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences of the United States of America* 1990, 87: 1663-1667.

212. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG *et al.*: The sequence of the human genome. *Science* 2001, 291: 1304-1351.

213. Verlinden L, Eelen G, Beullens I, Van Camp M, Van Hummelen P, Engelen K *et al.*: Characterization of the condensin component Cnap1 and the protein kinase Melk as novel E2F-target genes down-regulated by 1,25-dihydroxyvitamin D3. *Journal of Biological Chemistry* 2005, 280: 37319-37330.

214. Vincent L, Soille P: Watersheds in Digital Spaces - An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1991, 13: 583-598.

215. Visakorpi T, Hyytinen E, Koivisto P, Tanner M, Keinanen R, Palmberg C *et al.*: In vivo amplification of the androgen receptor gene and progression of human prostate-cancer. *Nature Genetics* 1995, 9: 401-406.

216. Visakorpi T, Kallioniemi AH, Syvanen AC, Hyytinen ER, Karhu R, Tammela T *et al.*: Genetic changes in primary and recurrent prostate-cancer by comparative genomic hybridization. *Cancer Research* 1995, 55: 342-347.

217. Wang D, Huang J, Xie H, Manzella L, Soares MB: A robust two-way semi-linear model for normalization of cDNA microarray data. *BMC Bioinformatics* 2005, 6: 14.

218. Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B *et al.*: Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biology* 2003, 4: R5.

219. Wei GH, Liu DP, Liang CC: Charting gene regulatory networks: strategies, challenges and perspectives. *Biochemical Journal* 2004, 381: 1-12.

220. Wendisch VF, Zimmer DP, Khodursky A, Peter B, Cozzarelli N, Kustu S: Isolation of Escherichia coli mRNA and comparison of expression using

mRNA and total RNA on DNA microarrays. *Analytical Biochemistry* 2001, 290: 205-213.

221. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P *et al.*: Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 2001, 8: 625-637.

222. Wolkenhauer O, Moller-Levet C, Sanchez-Cabo F: The curse of normalization. *Comparative and Functional Genomics* 2002, 3: 375-379.

223. Wu CFJ: Jackknife, bootstrap, and other resampling methods in regression analysis. *Annals of Statistics* 1986, 14: 1261-1295.

224. Yang MCK, Ruan QG, Yang JJ, Eckenrode S, Wu S, McIndoe RA *et al.*: A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiological Genomics* 2001, 7: 45-53.

225. Yang YH, Buckley MJ, Speed TP: Analysis of cDNA microarray images. *Briefings in Bioinformatics* 2001, 2: 341-349.

226. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J *et al.*: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 2002, 30: e15.

227. Yang YH, Speed T: Design issues for cDNA microarray experiments. *Nature Reviews Genetics* 2002, 3: 579-588.

228. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: Model-based clustering and data transformations for gene expression data. *Bioinformatics* 2001, 17: 977-987.

229. Yin WT, Chen T, Zhou XS, Chakraborty A: Background correction for cDNA microarray images using the TV+L1 model. *Bioinformatics* 2005, 21: 2410-2416.

230. Yoon D, Yi SG, Kim JH, Park T: Two-stage normalization using background intensities in cDNA microarray data. *BMC Bioinformatics* 2004, 5: 97.

231. Yu H, Chao J, Patek D, Mujumdar R, Mujumdar S, Waggoner AS: Cyanine dye dUTP analogs for enzymatic labeling of DNA probes. *Nucleic Acids Research* 1994, 22: 3226-3232.

232. Yu JD, Othman MI, Farjo R, Zareparsi S, Macnee SP, Yoshida S *et al.*: Evaluation and optimization of procedures for target labeling and hybridization of cDNA microarrays. *Molecular Vision* 2002, 8: 130-137.

233. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL *et al.*: An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Research* 2001, 29: e41.

234. Zammatteo N, Jeanmart L, Hamels S, Courtois S, Louette P, Hevesi L *et al.*: Comparison between different strategies of covalent attachment of

DNA to glass surfaces to build DNA microarrays. *Analytical Biochemistry* 2000, 280: 143-150.

235. Zhao Y, Li MC, Simon R: An adaptive method for cDNA microarray normalization. *BMC Bioinformatics* 2005, 6: 28.

236. Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN *et al.*: Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* 2000, 406: 90-94.

237. Zhu ZR, Chao J, Yu H, Waggoner AS: Directly labeled DNA probes asing fluorescent nucleotides with different length linkers. *Nucleic Acids Research* 1994, 22: 3418-3422.

238. Zhu ZR, Waggoner AS: Molecular mechanism controlling the incorporation of fluorescent nucleotides into DNA by PCR. *Cytometry* 1997, 28: 206-211.

Curriculum Vitae

Kristof Engelen was born in Diest, Belgium, on November 2nd, 1977. In 1995, he started his education in applied biological sciences at the K.U.Leuven, where he received the Candidacy diploma in Bioscience Engineering in 1997, and the Masters diploma in Cellular and Biotechnological Engineering in 2000. From September 2000 until May 2001 he worked as a Research Assistant at the CMPG research group (formerly FAJ) under supervision of Prof. Jozef Vanderleyden. From September 2000 until May 2001 he worked as a Research Assistant at the ESAT-SCD research group under supervision of Prof. Bart De Moor. Since January 2002 he has been pursuing his PhD research as a Research Assistant of the '*Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen*' (IWT-Vlaanderen) in the research group ESAT-SCD, under the supervision of Prof. Bart De Moor and Prof. Kathleen Marchal.