RIVM-rapport 340200002/2007

Bioinformatica ten behoeve van genomics

Pennings, J.L.A., Hoebee, B.

Contact:
J.L.A. Pennings
Laboratorium voor Toxicologie, Pathologie en Genetica
Jeroen.Pennings@rivm.nl

Dit technische rapport werd geschreven in het kader van project S/340200: ‘Genomics’.

RIVM, A.van Leeuwenhoeklaan 9, 3721 MA Bilthoven, Nederland

Rapport in het kort

Bioinformatica ten behoeve van genomics

Sinds enkele jaren wordt op het RIVM genomicsonderzoek uitgevoerd. Genomics omvat grootschalig onderzoek naar het erfelijk materiaal (DNA) van organismen. Dit onderzoek levert inzicht op in de manier waarop erfelijke eigenschappen zich vertalen naar het functioneren van een cel, en uiteindelijk een heel organisme. De praktische uitvoering van genomicsexperimenten is recentelijk beschreven in rapport 340200001 “Genomics: Implementatie, toepassing en toekomst”, dat in december 2006 is verschenen.

Dit rapport gaat in op de bioinformatica die het RIVM heeft opgezet en ontwikkeld. Bioinformatica is de wetenschap die methoden uit de informatica gebruikt om biologische data te kunnen verwerken en analyseren. Deze specifieke kennis is nodig om de grote hoeveelheden data die genomicsexperimenten genereren, te kunnen analyseren. De verschillende stappen in de data-analyse, zoals beeldverwerking, kwaliteitscontrole, normalisatie, statistische analyse, patroonherkenning, verlopen succesvol volgens algemeen geaccepteerde methoden. De bioinformatica voor de verdere biologische interpretatie van de resultaten is wereldwijd nog volop in ontwikkeling. In samenwerking met andere instituten wordt dit onderzoeksgebied gevolgd en worden nieuwe ontwikkelingen toegepast.

De komende jaren zullen er via de literatuur meer data van genomicsexperimenten beschikbaar komen. Om die te kunnen vergelijken en te combineren zijn bioinformatica-methoden beschikbaar, die zich de komende jaren verder zullen ontwikkelen. Naast genomicsdata zullen ook steeds meer andere gegevens (bijvoorbeeld eiwit- en metabolietgegevens) beschikbaar komen. Dit biedt mogelijkheden om meerdere soorten data te integreren. Deze aanpak wordt “systems biology” genoemd en is vooral interessant om tot een betere risicoschatting van stoffen te komen. Ook bestaat behoefte aan bioinformatica voor grootschalig eiwitonderzoek (proteomics), dat het RIVM wil gebruiken voor bevolkingsonderzoeken en screeningsprogramma’s van micro-organismen.

Trefwoorden: bioinformatica, genomics, microarray, statistiek

Abstract

Bioinformatics for genomics purposes

Genomics constitutes large-scale research on hereditary material (DNA) of organisms. The genomics research that has been carried out the last few years at the National Institute for Public Health and the Environment (RIVM) has given us insight into the way hereditary information is translated into the functioning of a cell and eventually a whole organism. Practical realization of genomics experiments has recently been described in report 340200001 'Genomics: Implementation, application, and future'.

Since genomics experiments generate a large amount of data, analysis demands specific expertise. The last few years has seen the set-up and further development of the bioinformatics required. The various steps in the data analysis, including image analysis, quality control, normalisation, statistical analysis and pattern recognition, are carried out successfully according to generally accepted methods. The bioinformatics concerned with interpretation of the results is worldwide in full development. This field will be closely followed and new developments applied in cooperation with other institutes.

More genomics experimental data will become available via the literature in the coming years. Bioinformatics methods for comparing and combining these data are available and will develop further in the future. In addition, an increasing number of other kinds of data sets (like protein or metabolite data) will become available, thereby creating possibilities for integration of multidisciplinary data. This approach is called 'systems biology' and is especially interesting for a better risk assessment for chemicals. Furthermore, there will be a need for bioinformatics for proteomics, which the RIVM aims to use for population screening programmes and screening applications on microorganisms.

Key words: bioinformatics, genomics, microarray, statistics

Inhoud

Samenvatting

De afgelopen vier jaar is op het RIVM de bioinformatica ten behoeve van genomics-analyses opgezet. Dit gebeurde in het kader van het SOR-project S/340200: 'Genomics', samen met het implementeren van microarray- en andere genomicstechnieken binnen het RIVM. De voornaamste aandachtspunten binnen het project waren een (tijds)efficiënte analyse van de grote hoeveelheden data en het onderscheiden van daadwerkelijke effecten van artefacten of vals-positieven.
Voor de verschillende stappen in de microarraydata-analyse voor transcriptomics- en CGH-experimenten zijn er protocollen en algoritmes ontwikkeld. De eerste stappen – beeldverwerking op microarrayscans, kwaliteitscontrole, normalisatie, statistische analyse, patroonherkenning – verlopen nu succesvol. Ook is er software geïmplementeerd voor de verdere data-analyse en -interpretatie, zoals het geautomatiseerd koppelen van gennaam aan functie en pathway-analyses. Ook op dit gebied kan worden voorzien in de huidige behoeften. De interpretatie van de verkregen resultaten in biologische termen is op dit moment het voornaamste ontwikkelingspunt van de bioinformatica. Dit geldt zowel voor het RIVM als voor andere onderzoeksinstituten. Dit gebied wordt dan ook nauwlettend gevolgd en nieuwe ontwikkelingen worden toegepast, onder andere via samenwerkingen als het Biomax-platform. De komende jaren zullen meer arraydata (publiek) beschikbaar komen, evenals andersoortige data zoals eiwit- en metabolietgegevens. Dit biedt verdere mogelijkheden tot data-integratie, o.a. op het gebied van infectieziekten en toxicogenomics, waarbij uiteindelijk naar een 'systems biology' aanpak kan worden gestreefd.
Er zijn op dit moment geen knelpunten in de analyse van lopende projecten en het RIVM loopt in de pas met de algemene ontwikkelingen in Nederland en daarbuiten. De genomics-bioinformatica is RIVM-breed beschikbaar en wordt in een ruim aantal projecten gebruikt. Wanneer de komende jaren technieken als proteomics RIVM-breed worden opgezet zullen ook hiervoor bioinformaticatoepassingen moeten worden ontwikkeld en geïmplementeerd.

1. Inleiding

Genomicsonderzoek omvat verschillende methoden voor grootschalig onderzoek aan het genoom van een organisme. Hierdoor is het mogelijk geworden complete genomen van organismen in kaart te brengen (structural genomics), het functioneren en tot expressie komen van genen in respons op bijvoorbeeld een stressor te bepalen (*functional genomics, transcriptomics*), vast te stellen hoe genetische variaties binnen een soort of populatie het functioneren van een organisme beïnvloeden, en genetische determinanten van ziekte te bepalen. Deze nieuwe ontwikkelingen en de daarbij gebruikte 'high throughput' technieken zoals bijvoorbeeld DNA-microarrays hebben geleid tot nieuwe inzichten en onderzoeksmethoden op het gebied van onder andere kanker, infectieziekten, chronische ziekten en toxicologie. Voor de toekomst wordt voorzien dat soortgelijke methodieken op eiwit- en metabolietniveau (*proteomics en metabolomics*) eveneens een belangrijke rol zullen gaan spelen in het wetenschappelijke onderzoek.

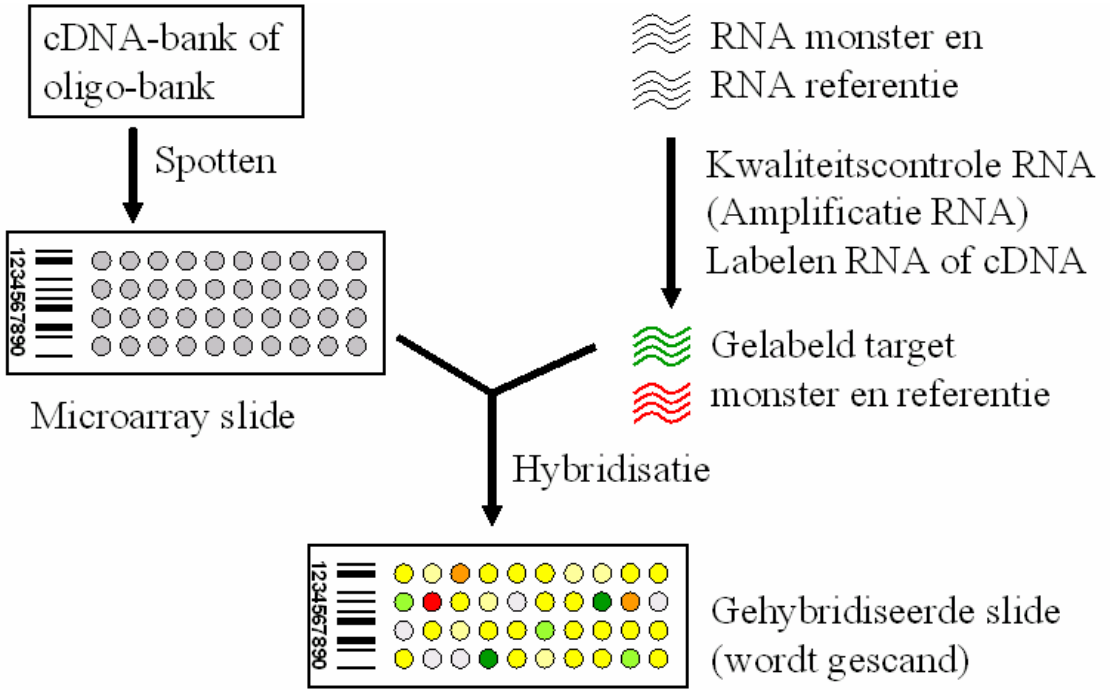
Eind 2001 werd op het RIVM de microarray-unit opgericht om de noodzakelijke infrastructuur en expertise voor het uitvoeren van de genomicstechnologieën in huis te halen, zodat deze aanpak voor het RIVM-volksgezondheidsonderzoek en de stoffen- en geneesmiddelenadvisering beschikbaar zou komen. Het toenmalige LEO en LIO hebben daar toen capaciteit voor vrijgemaakt terwijl de directie van het RIVM de benodigde gelden voor de aanschaf voor apparatuur beschikbaar heeft gesteld. Vanaf 2003 is vanuit het speerpunt 'Vernieuwing Meetmethoden' het SOR-project 'Genomics' (S/340200) opgestart, waarmee de benodigde capaciteit gefinancierd werd. Het doel van dit project is het opzetten en implementeren van genomicsmethodieken voor het RIVM. Binnen het project 'Genomics' zijn twee deelprojecten ondergebracht. Het eerste deelproject is gericht op het spotten van arrays en het uitvoeren van de experimenten. De vorderingen, stand van zaken en toekomstige ontwikkelingen op dit gebied zijn beschreven in het RIVM-rapport 340200001, 'Genomics: Implementatie, toepassing en toekomst'. Het tweede deelproject is gericht op de dataverwerking en bioinformatica, hetgeen het onderwerp is van dit rapport.
High-throughput genomicstechnieken leveren miljoenen datapunten (of enkele gigabytes aan data). Dit maakt een doordachte aanpak van de data-analyse noodzakelijk. Op grond van ervaringen bij andere instituten was het duidelijk dat aan dit aspect specifiek aandacht moest worden besteed, aangezien de analyse van genomicsexperimenten specifieke statistische en bioinformatica kennis vereist. Punten die hierbij spelen, zijn een (tijds)efficiënte analyse van de grote hoeveelheden data, het onderscheiden van daadwerkelijke effecten van artefacten of vals-positieven en het onderscheiden van specifieke van aspecifieke responsen. Na een dergelijke analyse is het aantal datapunten dat relevant is aanzienlijk teruggebracht (tot enige honderden). Voor een verdere interpretatie zijn echter nog aanvullende analyses noodzakelijk die gebruikmaken van een breed scala van specifieke databases en programma's. In dit rapport zal een overzicht worden gegeven van wat er binnen het Genomics-project bereikt is op het gebied van implementatie van bioinformatica binnen het RIVM. De toepassing van deze

bioinformatica heeft binnen de betrokken projecten inmiddels geleid tot een beter begrip van onder andere de gastheer-pathogeen-interactie, toxicologische respons en het ontstaan van kanker en chronische ziekten. Deze resultaten (zullen) worden gerapporteerd vanuit de desbetreffende projecten, in dit rapport zal alleen de implementatie van de bioinformatica vanuit het Genomics-project worden behandeld.
De voornaamste toepassing van genomics en micro-arrays binnen het RIVM ligt in de zogeheten transcriptomics, waarbij grootschalig veranderingen in genexpressie worden gemeten. In hoofdstuk 2 zal besproken worden hoe de data-analyse van transcriptomics-experimenten uitgevoerd wordt. De analyse van andere soorten genomicsdata wordt besproken in hoofdstuk 3 en in hoofdstuk 4 wordt ingegaan op het combineren en vergelijken van experimentele resultaten. Overige bioinformatica komt aan de orde in hoofdstuk 5. In hoofdstuk 6 wordt een overzicht gegeven hoe de bioinformaticakennis uitgewisseld wordt binnen en buiten het instituut. Hoofdstuk 7 bevat de conclusies, waarna in de bijlagen de momenteel gebruikte protocollen worden gegeven.

2. Bioinformatica ten behoeve van transcriptomics-analyse

2.1 Achtergrond

De meest gebruikte toepassing van genomicsonderzoek op het RIVM is transcriptomics. Dit omvat het grootschalig meten van de expressie van genen om te onderzoeken hoe de respons is na toediening van een agens, welke pathofysiologische processen een rol spelen, etc. De praktische uitvoering van transcriptomics staat schematisch weergegeven in Figuur 1. Voor dit soort experimenten maakt men gebruik van een microarray (een 'slide') waarop van duizenden verschillende genen een kleine hoeveelheid DNA (cDNA of oligo) wordt aangebracht, elk op een specifieke positie. Uit een te bestuderen weefsel wordt RNA geïsoleerd en na omzetting tot cDNA of cRNA gelabeld met een fluorescerende dye. Daarnaast wordt referentie-RNA met een andere fluorescerende dye gelabeld. Door de microarrays te hybridiseren met het gelabelde cDNA van zowel het analysemonster als het referentiemonster binden de gelabelde cDNA's aan een complementair gen op de array. Dit leidt tot een fluorescerend signaal op de microarray. Door de microarray na hybridisatie met een confocale laser te scannen ontstaat een beeld met de fluorescentiesignalen van de verschillende dyes.

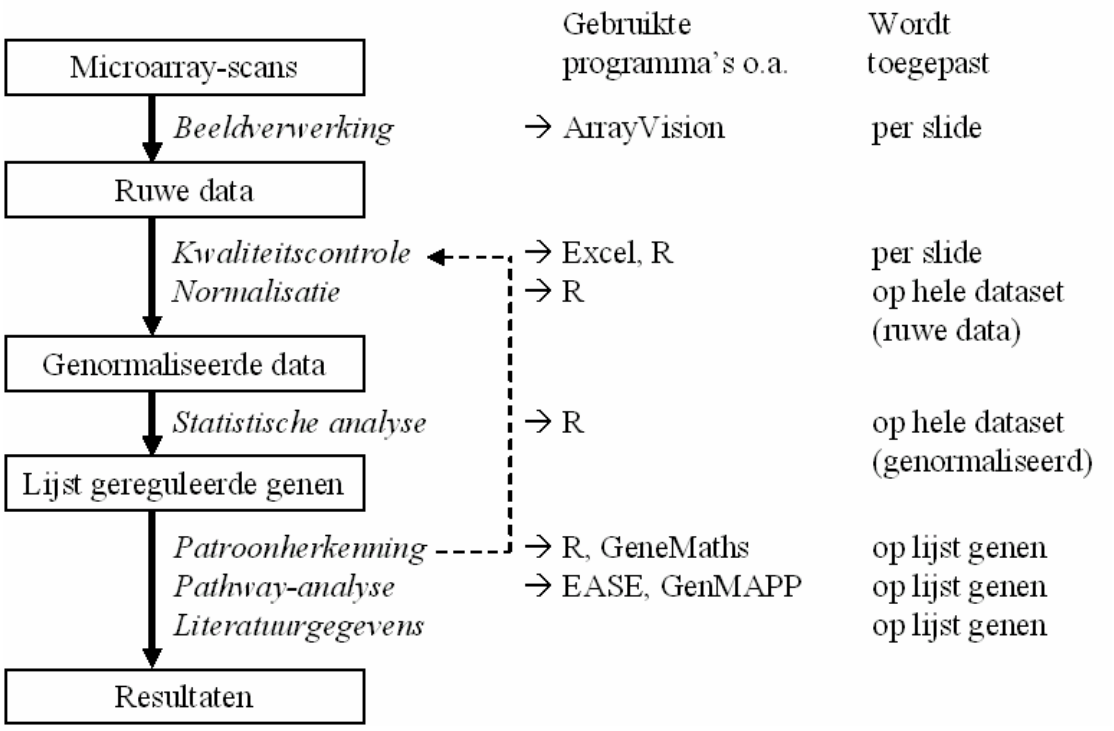


Figuur 1: Schematische weergave van een transcriptomicsexperiment.

Bij een microarray-experiment wordt een aantal monsters (doorgaans 10 tot 70) na labeling gehybridiseerd op een array. Deze monsters zijn afkomstig uit een aantal verschillende

groepen die verschillen in de achterliggende behandeling (bijv. behandeld vs. onbehandeld, geïnfecteerd vs. niet geïnfecteerd) of anderszins verschillend zijn (bijv. tumor vs. normaal, vroeg vs. laat, jong vs. oud). Het uiteindelijke doel van een microarray-experiment is vast te stellen welke genen verschillend tot expressie komen tussen deze groepen. In het verlengde daarvan wordt bepaald welke cellulaire processen (*pathways*) daarbij betrokken zijn en/of in welke pathways regulatie plaatsvindt.
Met name in de beginfase van het project is het opzetten van de bioinformatica gericht geweest op het operationeel maken van de microarray-analyse. Dit omvatte aspecten als beeldverwerking en het verder verwerken van ruwe data. Later is het accent verschoven naar aspecten als grootschalige data-analyse en -interpretatie, oftewel de vervolgstappen in de data-analyse.

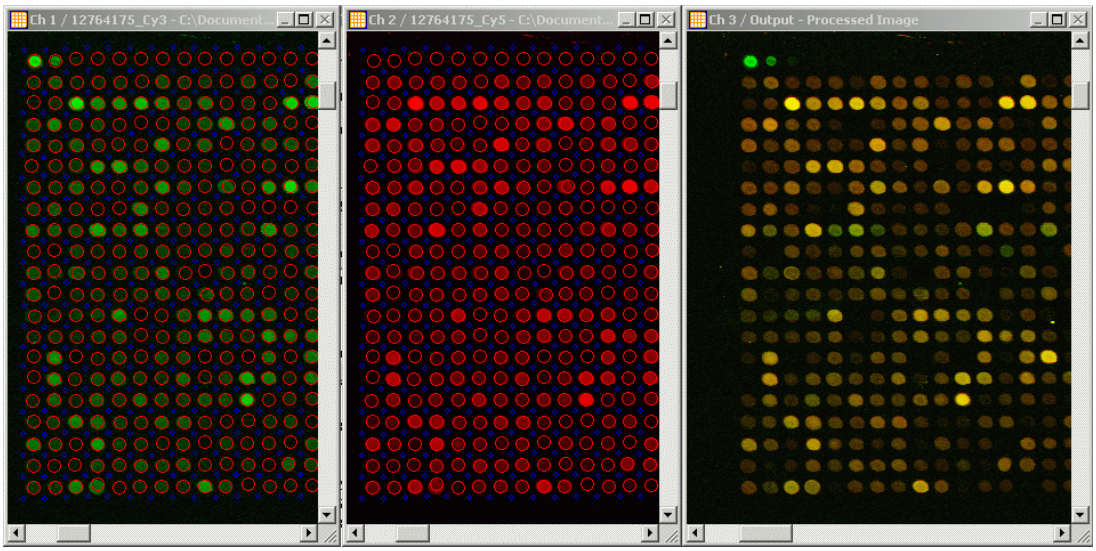
Een typische microarray-analyse bestaat uit een aantal stappen (Figuur 2) waarbij vanuit de microarrayscans via een aantal bewerkingen zoals beeldverwerking, kwaliteitscontrole, normalisatie en statistische analyse de uiteindelijke resultaten worden verkregen.



Figuur 2: Schematische weergave data-analyse van een microarray-experiment.

2.2 Beeldverwerking

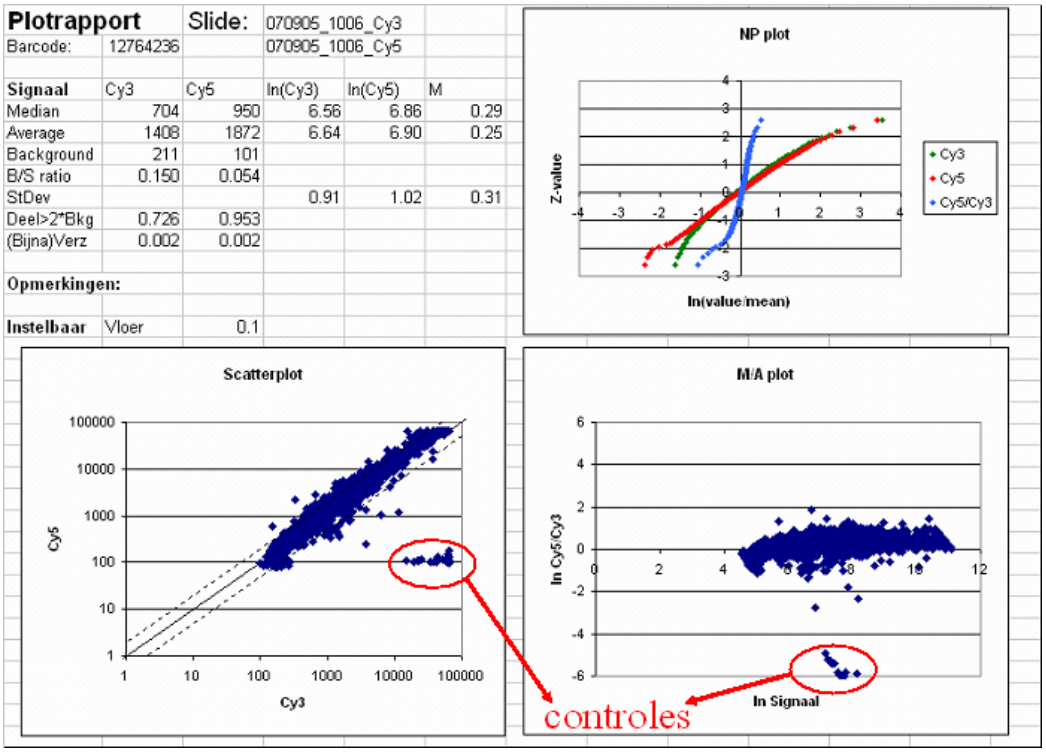
Na het scannen van een gehybridiseerde microarray bij 1, 2 of 3 verschillende golflengten (die overeenkomen met de gebruikte dyes) bestaat de eerste stap van de data-analyse uit de beeldverwerking (‘image extraction’, Figuur 3). Hierbij worden de scans per slide (~30 MB per dye) omgezet in de signaalwaardes per spot (~1 MB per slide). Per microarrayslide worden de scanbeelden die voor de verschillende dyes zijn verkregen, geladen in de daartoe bestemde software. Vervolgens wordt er een raster over de beelden gelegd, zodat de software de juiste positie van de spots kan vinden. Daarna worden de beelden voor de verschillende dyes gecombineerd en voor iedere spot de juiste positie bepaald in de ‘alignment’-stap. Per individuele spot wordt vervolgens een (gemiddelde) signaalwaarde per dye berekend, evenals de waardes voor het achtergrondsignaal en de ruis. Deze laatste twee dienen voor de kwaliteitscontrole (zie paragraaf 2.3). De waarden die de beeldverwerking oplevert, vormen de ruwe data en zijn daarmee de basis voor de volgende stappen in de data-analyse. Het is dan ook van belang om deze stap goed uit te voeren, want wanneer dit niet gebeurt kunnen onbruikbare signaalwaarden worden berekend, of signaalwaarden aan de verkeerde spots worden gekoppeld.
Voor het uitvoeren van deze beeldverwerking heeft het RIVM de beschikking over twee licenties van het ArrayVision pakket (Imaging Research). Voor het gebruik van deze software zijn een handleiding en protocollen beschikbaar (zie bijlagen). Op andere instituten wordt in plaats van ArrayVision ook andere software gebruikt (o.a. GenePix, ImaGene, ScanAlyze), deze zijn qua methodiek en gebruik vergelijkbaar.



Figuur 3: Fragment van een microarrayscan. V.l.n.r.: Cy3-scan, Cy5-scan en de overlay. Gezien het grote aantal genen op de slide kan slechts een deel ervan worden afgebeeld. Op de Cy3- en Cy5-scan is de positie van de spotjes weergegeven door middel van een raster. Verder is zichtbaar dat sommige spotjes een sterker Cy3- dan Cy5-signaal geven. Deze genen komen verhoogd tot expressie in het betrokken monster en zijn in de overlay groen gekleurd.

2.3 Kwaliteitscontrole

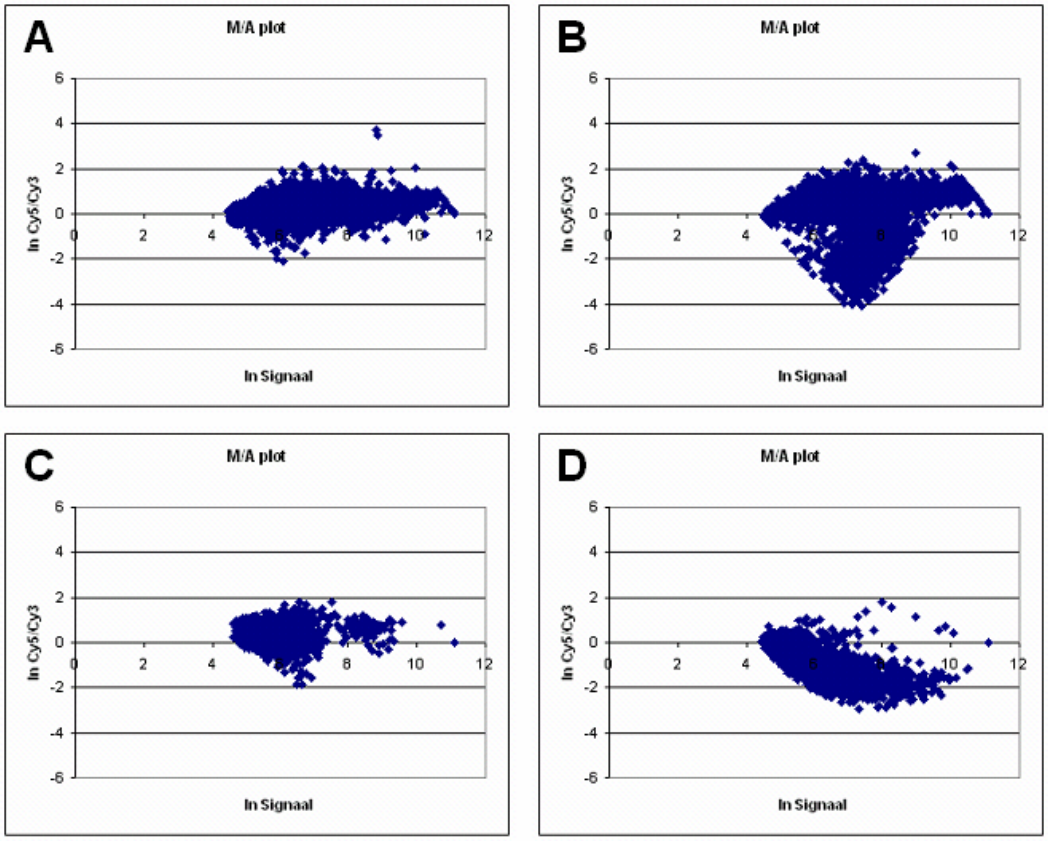
De volgende stap is de kwaliteitscontrole op de ruwe data. Het doel hiervan is om slides die in technisch opzicht mislukt zijn, te herkennen en uit te sluiten van verdere analyse.
De eerste kwaliteitscontrolestap vindt plaats tijdens de beeldverwerking door een visuele inspectie van de scans. Hierbij wordt onder andere bekeken of het fluorescerende signaal en de achtergrond gelijkmatig zijn verdeeld over de slide. Slides waarbij het signaal van meer dan ~10% van de spotjes verloren is gegaan doordat de hybridisatie niet goed is gelukt (bijv. door luchtbellen onder het dekglaasje) of omdat er vlekken of krassen op aanwezig zijn, worden van verdere analyse uitgesloten. Het tweede deel van de kwaliteitscontrole vindt plaats op de ruwe data. De ruwe data per slide worden geanalyseerd in een hiertoe ontwikkeld Excel-bestand (een voorbeeld hiervan staat in Figuur 4). Dit bestand berekent een aantal waarden, zoals gemiddeld signaal, achtergrond en dergelijke. Ook worden in het bestand de ruwe signaalwaarden van Cy3 en Cy5 per spot tegen elkaar uitgezet in een zogeheten scatterplot (Figuur 4, linksonder). Dit wordt ook op een andere manier weergegeven, namelijk in een zogeheten 'M/A-plot' waarin op de x-as het loggemiddelde signaal (de wortel van Cy5 x Cy3 per spot) wordt uitgezet tegen de Cy5/Cy3-ratio per spot (Figuur 4, rechtsonder).



Figuur 4: Kwaliteitscontrole op microarraydata met behulp van een Excel-bestand. In de linkerbovenhoek staan de waarden die betrekking hebben op de wiskundige beschrijving van de data. Linksonder staan de signaalwaarden voor Cy3 en Cy5 per spot tegen elkaar uitgezet. De grafiek rechtsonder geeft hiervan een andere weergave. Rechtsboven staan 'normal probability plots' van de Cy3- en Cy5-waarden en hun ratio.

Deze analyse wordt voor alle slides gedaan, waarna de resultaten van de verschillende slides worden vergeleken om te controleren of de hybridisatie bij de verschillende slides

vergelijkbaar is verlopen (zie de bijlagen voor de te gebruiken protocollen). Verschillen kunnen wijzen op bijvoorbeeld een niet goed gelukte labeling, en wanneer waarden meer dan een orde van grootte afwijken, worden slides verworpen.
Het belangrijkste criterium voor de kwaliteitscontrole is de vorm van de puntenwolk in de M/A-plot (zie Figuur 5). De (internationaal) gangbare aanname voor microarray-analyse is dat de meeste genen door een blootstelling of behandeling geen, of hooguit, weinig verandering in expressie zullen vertonen. Hierop is de normalisatie en verdere analyse gebaseerd. Dit komt tot uitdrukking in een sigaarvormige puntenwolk die in de scatterplot globaal langs de x=y-lijn ligt en in de M/A-plot horizontaal rond de x-as (zie Figuur 4 links- en rechtsonder en Figuur 5A). Wanneer de puntenwolk duidelijke afwijkingen vertoont (zoals in Figuur 5B en C t.o.v. 5A), of wanneer de puntenwolk in de M/A-plot duidelijk niet horizontaal ligt (zoals in Figuur 5D), wordt een slide verworpen. Een dergelijke puntenwolk wijst op een technisch probleem dat een zodanig grote invloed heeft op de ruwe data dat het tijdens de normalisatie niet meer (voldoende) gecorrigeerd kan worden.



Figuur 5: Voorbeelden van M/A-plots. A: goede hybridisatie. B: Cy3-kleurige vlek op de array, te zien als een uitstulping op de puntenwolk. C: slechte labeling van Cy3 en Cy5, dit leidt tot een (korte) puntenwolk die vooral zwakke signalen omvat. D: Cy5-afbraak door ozon in de lucht. Omdat sterkere Cy5-signalen relatief sterker worden aangetast, ontstaat er een trend waarbij de Cy5/Cy3-ratio afneemt bij toenemende signaalsterkte.

Bij het beoordelen van de M/A-plot is op het oog alleen de omtrek van de puntenwolk zichtbaar. Hoe de datapunten binnen de wolk verdeeld zijn, is echter op deze manier niet te zien, al kan deze informatie wel nuttig zijn. Daarom wordt er ook een ‘normal probability plot’ gemaakt (Figuur 4, rechtsboven), waarin de dataverdeling van deze waarden en hun

onderlinge ratio wordt weergegeven. Deze weergave leidt niet tot aanvullende eisen en wordt niet gebruikt voor het goed- of afkeuren van slides, maar geeft inzicht in de technische oorzaken wanneer een slide moet worden afgekeurd.
Het uitvoeren van kwaliteitscontrole is niet op alle instituten gebruikelijk en ook de manier waarop verschilt tussen instituten. Op het RIVM is de kwaliteitscontrole vrij uitgebreid en streng in vergelijking met elders, omdat we het belangrijk vinden dat de gebruikte data betrouwbaar zijn. Wanneer onbetrouwbare slides niet worden verworpen, heeft dit namelijk invloed op de uiteindelijk verkregen resultaten en kan dit leiden tot verkeerde conclusies. Kwaliteitscontrole via visuele inspectie en/of M/A-plots is op veel instituten wel gebruikelijk, maar niet overal worden de minder goede slides met dezelfde stringentie verworpen. De aanvullende controle van microarrayslides via een ‘normal probability plot’ is op het RIVM ontwikkeld en wordt in een enigszins aangepaste vorm, wordt nu ook gebruikt op het RIKILT en de MicroArray Department (MAD) van de Universiteit van Amsterdam.

2.4 Normalisatie

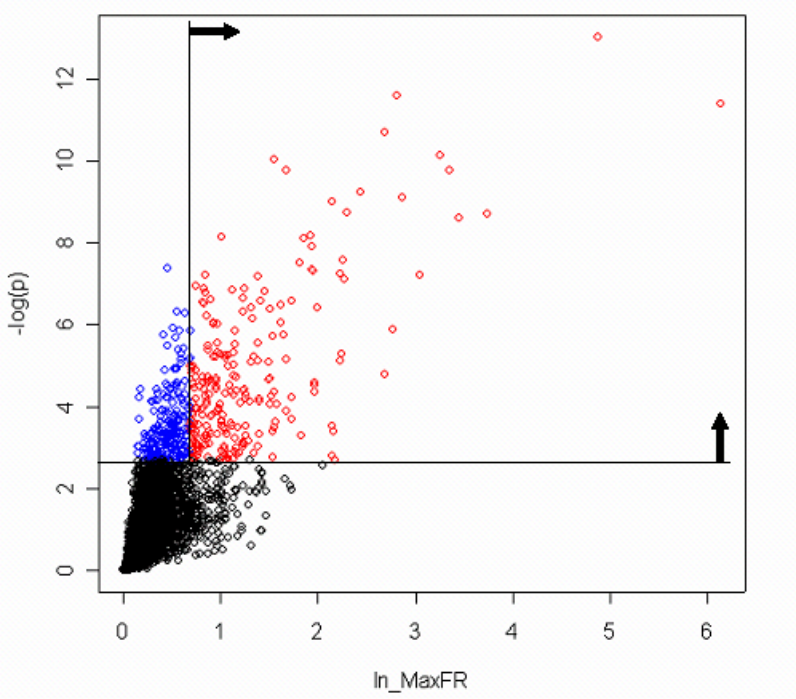
Na de kwaliteitscontrole vindt een normalisatiestap op de data plaats. Het doel van de normalisatiestap is experimentele verschillen (zoals bijv. de hoeveelheid opgebracht monster, labelingsverschillen tussen monsters, etc.) tussen slides te corrigeren en daarmee de data van verschillende slides zo goed mogelijk vergelijkbaar te maken. Zo blijft er in de volgende stap (de statistische analyse) zo weinig mogelijk ruis over en kunnen verschillen in genexpressie beter worden herkend. Men gaat er bij de normalisatiestap vanuit dat op elke slide evenveel RNA is opgebracht en dat de meeste genen geen, of hooguit weinig, verandering in expressie zullen vertonen tussen de verschillende monsters. Bovendien bestaat de referentie op alle slides binnen een experiment uit hetzelfde monster, zodat hier geen verschillen in genexpressie zullen optreden.
De normalisatiestap maakt gebruik van de complete ruwe dataset. Tijdens de stap vinden correcties plaats op de ruwe data, zodat binnen een experiment voor iedere slide eenzelfde gemiddelde signaal wordt verkregen waarbij ook kleinere (achtergrond- of dye-afhankelijke) systematische fouten worden gecorrigeerd. Daarbij worden verschillen in de referentie tussen de verschillende slides onderling verrekend, zodat de beide (Cy3- en Cy5-) waarden worden verwerkt tot één waarde per spotje als maat voor de genexpressie; hierbij worden replicaspotjes gecombineerd.
Voor de normalisatie wordt gebruikgemaakt van het statistische programma R (www.r-project.org/). Hierin zijn algoritmes ontwikkeld die gebruikmaken van zogenaamde ‘quantile normalization’. Dit is qua databewerking vergelijkbaar met het eveneens gangbare LOWESS, maar heeft als voordeel dat het sneller werkt. Daarnaast maakt deze aanpak het mogelijk om microarraydata te analyseren indien één of drie fluorescerende dyes gebruikt worden in plaats van de gebruikelijke twee. Op dit moment wordt dit R-algoritme op het RIVM standaard gebruikt in de data-analyse. Het gebruik van R is ook op andere instituten

gangbaar, waarbij zij net als het RIVM gebruikmaken van een combinatie van publieke en eigen algoritmes.
Incidenteel wordt voor RIVM-projecten gebruikgemaakt van Affymetrix-chips. De praktische uitvoering hiervan vindt buiten het RIVM plaats op arrayunits die over de vereiste apparatuur beschikken. De firma levert hiervoor eigen analysesoftware voor de beeldverwerking, kwaliteitscontrole en normalisatie. Voor de normalisatie wordt naast de MAS5 software van de fabrikant ook gebruikgemaakt van het publieke RMA-algoritme [1]. Dit laatste algoritme wordt internationaal in toenemende mate als de best beschikbare keuze gezien. De verdere analyse verloopt hetzelfde als bij andere arrays.

2.5 Statistische analyse

Na de normalisatie vindt de daadwerkelijke statistische analyse van de data plaats om genexpressie tussen de verschillende groepen (bijv. behandeld – onbehandeld, vroeg – laat, geïnfecteerd – niet geïnfecteerd, tumor – normaal) te vergelijken. Doel hiervan is vast te stellen welke genen verschillend tot expressie komen, waarna met deze genen verdere analyses worden uitgevoerd. Deze stap wordt uitgevoerd op basis van de hele genormaliseerde dataset.
De statistische analyses worden in twee delen uitgevoerd. Als eerste worden voor alle genen de expressieniveaus in de verschillende groepen bepaald. Deze worden vervolgens onderling vergeleken, om te bepalen of deze statistisch significant zijn. Wanneer er in het experiment meer dan twee groepen zijn, worden de verschillende groepen meestal niet onderling paarsgewijs vergeleken maar worden alle groepen onderling tegelijkertijd vergeleken door middel van een one-way ANOVA. Deze aanpak levert namelijk meer statistische power op en de manier waarop de verschillende groepen zich onderling verhouden wordt daarna duidelijk door middel van patroonherkenning (zie paragraaf 2.6). Het vergelijken van de expressieniveaus leidt per gen tot een p-waarde waarop bepaald kan worden of deze aan de significantiecriteria voldoet. Gezien het grote aantal genen dat wordt geanalyseerd, moet hierbij gecorrigeerd worden voor multiple testing. Wanneer dit niet zou gebeuren, krijgt men namelijk voor grote aantallen genen een vals-positief resultaat, op een array met bijvoorbeeld 22.000 genen 220 vals-positieven bij een p-waarde van 0,01. Deze correctie voor multiple testing vindt doorgaans plaats op basis van de zogeheten False Discovery Rate (FDR) [2], die het percentage vals positieve resultaten berekent. Meestal wordt een FDR van 5 of 10% gebruikt, zodat de lijst met statistisch significant verschillende genen respectievelijk 5 of 10% vals-positieve genen bevat.
In de tweede stap wordt voor ieder gen een FoldRatio berekend. Dit is de ratio van het maximale t.o.v. het minimale groepsgemiddelde-signaal, deze geeft de (maximale) mate van verschil tussen de groepen aan. Deze waarde kan gebruikt worden om genen met sterkere of zwakkere effecten van elkaar te onderscheiden. Deze stap wordt toegepast omdat vroeger de interesse vooral uitging naar de sterke effecten. Dit kwam onder meer omdat zwakkere effecten moeilijk of niet worden bevestigd met andere methoden en daardoor moeilijker

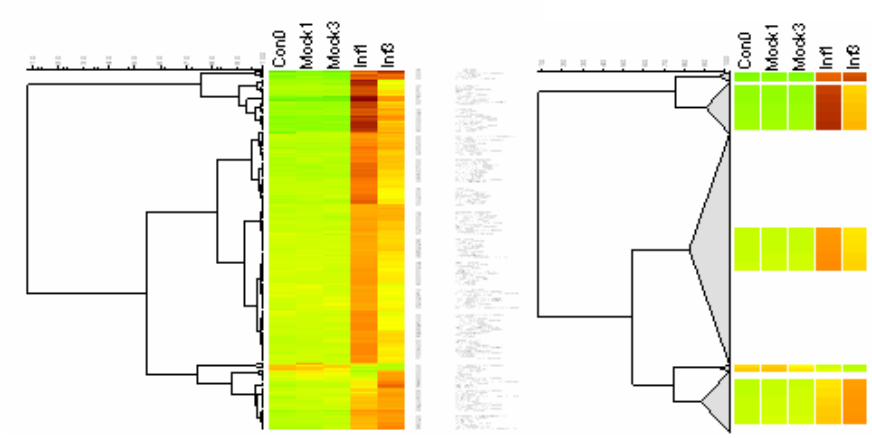
gepubliceerd konden worden. Momenteel is statistische significantie het belangrijkste criterium, maar voor de meeste projecten wordt nog steeds een filtering toegepast op basis van de FoldRatio, waarbij genen die statistisch significant verschillend tot expressie komen maar een relatief klein effect vertonen, worden weggelaten. Enerzijds gebeurt dit omdat men anders soms zoveel gereguleerde genen krijgt dat de downstreamanalyse te complex wordt, anderzijds omdat voor sommige vraagstellingen zwakkere effecten biologisch minder interessant zijn (bijvoorbeeld als biomarker).
De berekende p-waarde en FoldRatio worden tegen elkaar uitgezet in een zogeheten vulkaanplot (Figuur 6), zodat te zien valt in welke mate genen aan de gekozen eisen voldoen. Deze criteria worden voor ieder experiment vastgesteld in overleg met de betrokken onderzoekers, en verschillen afhankelijk van de doelstelling van het experiment. Gangbare criteria zijn echter een FDR van 0,05 of 0,10 en een FoldRatio van 1,5 of 2,0. Op deze manier wordt een lijst verkregen met gereguleerde genen.



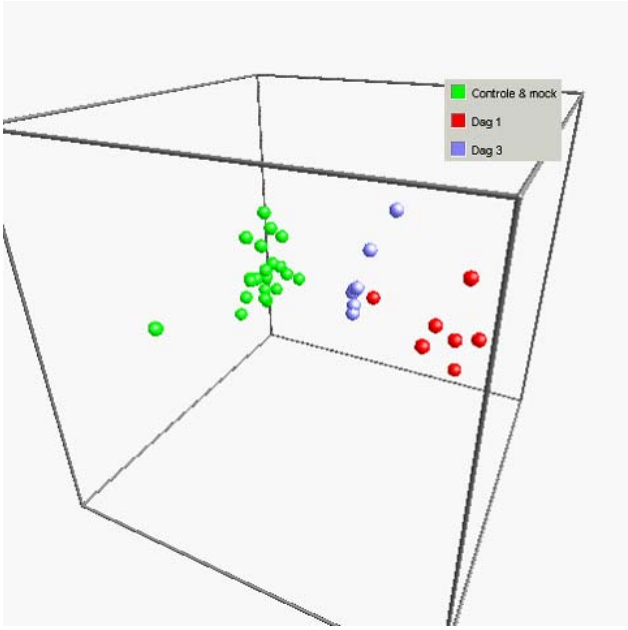
Figuur 6: Voorbeeld van een vulkaanplot. Door middel van kleur is weergegeven welke genen niet significant zijn (zwart) of wel significant zijn en daarnaast een FoldRatio hebben kleiner (blauw) of groter (rood) dan een factor 2. In dit geval worden de roodgekleurde genen gebruikt voor verdere analyse.

2.6 Patroonherkenning

Nadat een lijst significant gereguleerde genen is verkregen, wordt daarop een aantal vormen van patroonherkenning toegepast. Doel van deze patroonherkenningsstap is vast te stellen welke genen vergelijkbare expressiepatronen hebben en daarom mogelijk gezamenlijk gereguleerd worden. Voor de patroonherkenning wordt voornamelijk gebruikgemaakt van hiërarchische clustering (Figuur 7) en Principal Component Analysis (Figuur 8).



Figuur 7: Hiërarchische clustering op microarraydata. Expressiepatronen zijn in een heatmap met een kleurschaal (van laag naar hoog: groen-geel-rood) weergegeven. Links staat het dendrogram, rechts de opdeling in vijf afzonderlijke clusters. Te zien is dat de sterkste effecten plaatsvinden in de Inf1-groep, waarin veel genen verhoogd tot expressie komen.



Figuur 8: Principal Component Analysis (PCA) op microarraydata. De data zijn dezelfde als in Figuur 7, waarbij hier de groepen monsters zijn weergegeven. De assen en schaal zijn in arbitraire eenheden. Er is te zien dat de monsters in de dag 1-groep het sterkst verschillen van de controles.

Hiërarchische clustering geeft overeenkomsten tussen de expressie van genen (of monsters) weer via een boomstructuur waarbij genen met vergelijkbare expressieveranderingen zich op bij elkaar gelegen ‘takken’ bevinden. Hiërarchische clustering wordt meestal gecombineerd met een zogeheten heatmap, waarbij genexpressie wordt weergegeven op een kleurschaal (zie Figuur 7). Principal Component Analysis berekent op de achterliggende datamatrix de wiskundige componenten die de variatie in de data zo goed mogelijk beschrijven. Deze analyse wordt gebruikt om de data weer te geven in een Figuur waarbij geldt dat hoe dichter genen (of monsters) bij elkaar staan, hoe kleiner hun onderlinge verschillen zijn (zie Figuur 8).

Deze technieken worden ook toegepast op de gebruikte arrays om zo nog een aanvullende kwaliteitscontrole uit te voeren op de betrokken arrays en monsters. Op deze manier kunnen eventuele verschillen in de uitvoering van dag tot dag of tussen batches slides worden geïdentificeerd. Dergelijke controles helpen om bronnen van experimentele variatie op te sporen en waar mogelijk te verminderen, om op die manier te kwaliteit van de praktische uitvoering te kunnen verbeteren.
Voor deze analyses zijn op het RIVM twee netwerklicenties GeneMaths (Applied Maths) en een netwerklicentie SpotFire aanwezig. Daarnaast vindt een deel van deze analyse plaats in R op basis van onder andere het DNAMR-package en door ons ‘in huis’ ontwikkelde algoritmes.

2.7 Pathway-analyse

De voorgaande analyse-stappen hebben een lijst met significant gereguleerde genen opgeleverd. Deze lijst geeft echter nog maar een beperkt inzicht in de vraag in welke biologische processen (‘pathways’) er een respons optreedt en welke interacties er optreden na bijvoorbeeld de blootstelling aan een stof of een infectie, terwijl dit meestal het voornaamste onderzoeksdoel is. Om de lijst met genen te kunnen interpreteren en uiteindelijk op een biologisch complexer niveau conclusies te kunnen trekken, moet deze uitgebreid worden met meer informatie over de desbetreffende genen. Hiervoor worden de gennamen (of codes daarvoor) als eerste gekoppeld aan een uitgebreide beschrijving van naam, functie, chromosomale locatie, en, waar van toepassing, een korte samenvatting van literatuurgegevens. Hiervoor wordt gebruikgemaakt van gegevens in publieke databanken, met name de functionele annotatie van het Gene Ontology Consortium (www.geneontology.org) voor wat betreft biologisch proces (bijv. apoptose), moleculaire functie (bijv. kinase) en cellulaire component (bijv. mitochondrion). Deze annotatie vindt plaats met een lokale installatie van het DAVID/EASE-programma (david.abcc.ncifcrf.gov/). Voor het gebruik hiervan is een handleiding aanwezig (zie bijlage).
Ten tweede wordt er op de lijst met gereguleerde genen een pathwayverrijkingsanalyse toegepast. Dit heeft als doel om vast te stellen of de lijst met gereguleerde genen verrijkt is voor genen die bij een gezamenlijk proces (pathway) betrokken zijn. De term ‘pathway’ wordt hierbij in bredere zin gebruikt; hieronder verstaat men alle Gene Ontology categorieën (biologisch proces, moleculaire functie, cellulaire component) en andere groepen genen die bij eenzelfde proces betrokken zijn. Deze pathwayverrijkingsanalyse wordt ook toegepast op delen van de genenlijst waarvan op basis van patroonherkenning (zie 2.6) een overeenkomst in genexpressie is gevonden. Hierbij kan men denken aan genen met verhoogde of juist verlaagde expressie, maar ook aan genen waarvan het expressiepatroon in de tijd een zelfde trend vertoont. Voor deze analyse wordt gebruikgemaakt van de DAVID/EASE -software (david.abcc.ncifcrf.gov/), GoStat (gostat.wehi.edu.au/) en OntoTools (vortex.cs.wayne.edu/ontoexpress/). Deze tools maken gebruik van de Gene Ontology annotatie. Als aanvulling hierop wordt gebruikgemaakt van de KEGG-database (www.genome.ad.jp/kegg/) waarin andere, met name metabole, pathways zijn beschreven.

Daarnaast is er een gezamenlijke licentie MetaCore (www.genego.com). Dit programma bevat een eigen databank met door experts samengestelde pathways en is daarmee aanvullend op de andere software.
Ten derde wordt voor de gevonden pathways nagegaan hoe de gevonden expressieverschillen onderling samenhangen. Wanneer er in bijvoorbeeld een metabole pathway regulatie plaatsvindt, is het interessant te weten of de daarbij betrokken genen allemaal verhoogd of verlaagd tot expressie komen, zodat een uitspraak kan worden gedaan over de metabole route als geheel en vorming of verbruik van metabolieten. Voor deze vorm van pathway-analyse worden in een pathwayschema de desbetreffende stappen voorzien van een kleur of grafiekje; om zo genexpressieverschillen weer te geven. Dit maakt het tevens mogelijk om in het bijbehorende pathway-schema zogeheten gene hubs te herkennen, dit zijn genen die de expressie van een groot aantal genen beïnvloeden en daarmee een centrale rol spelen in het regulatieproces. Een bijkomend voordeel is dat deze vorm van weergave het mogelijk maakt om pathways te herkennen waarin een groot aantal genen een vergelijkbare, maar kleine (en niet significante) verandering in genexpressie vertonen. Indien dit het geval is, kan het gebeuren dat een dergelijke pathway niet wordt gevonden bij de eerdere stappen in de analyse, maar kan deze pathway op basis van bijvoorbeeld literatuuraanwijzingen toch nader worden bekeken. Deze grafische vorm van pathway-analyse heeft als voordeel dat het visueel eenvoudig overkomt, al is men hiervoor afhankelijk van de beschikbaarheid van een goed en actueel pathwayschema. Voor dergelijke pathway-analyses wordt gebruikgemaakt van publieke software als GenMAPP (www.genmapp.org/), KEGG (www.genome.jp/kegg/), BioCarta (www.biocarta.com/) evenals het commerciële pakket MetaCore (www.genego.com).
Voor pathway-analyses worden dus verschillende methoden gecombineerd. Dit levert een vollediger beeld op dan het gebruik van één enkele methode, aangezien de verschillende methoden onderling aanvullend werken. Dit zal in paragraaf 2.8 worden geïllustreerd.

2.8 Vergelijking methoden

De ontwikkelingen op het gebied van microarray-analyse in Nederland zijn zodanig dat momenteel voor de meeste arrayunits de grootste uitdaging ligt bij het maken van de vertaalslag van de resultaten naar de biologische betekenis. Daarom is vanuit het RIVM, samen met de MicroArray Department (MAD) van de Universiteit van Amsterdam, het initiatief genomen voor een landelijk platform waarin ervaringen en software op dit vlak worden besproken. In dit Biomax-platform (Biological Interpretation Of MicroArray eXperiments) zijn bioinformatici van zestien instituten uit Nederland vertegenwoordigd, die tijdens platformbijeenkomsten informatie uitwisselen aan de hand van een gezamenlijke casus-dataset-analyse. Deze meetings worden door het RIVM georganiseerd.
Op de Biomax-bijeenkomst in april 2006 heeeft een aantal instellingen een vergelijking gemaakt tussen hun data-analyses op de hexachloorbenzeen-dataset van Ezendam et al. [3]. In dit experiment zijn ratten via het voer vier weken blootgesteld aan verschillende doses

hexachloorbenzeen, waarna RNA uit verschillende weefsels werd geanalyseerd op Affymetrix-chips. Om het geheel overzichtelijk te houden, werd voor het vergelijken van analyse-methoden alleen gebruikgemaakt van de leverdata.
Vergelijking van de resultaten, die via verschillende analyses zijn verkregen gaf, de volgende bevindingen:

- Beeldverwerking en kwaliteitscontrole: Aangezien de analyse is uitgevoerd op Affymetrix-data (waarbij de beeldverwerking geautomatiseerd werd uitgevoerd) is er geen vergelijking uitgevoerd tussen verschillende methoden voor beeldverwerking. Ook zijn er geen vergelijkingen uitgevoerd tussen methoden voor kwaliteitscontrole.
- Normalisatie en statistiek: Met de momenteel gangbare algoritmes zijn deze eerste stappen in de data-analyse niet langer de meest kritieke stappen. Deze leiden allemaal tot vergelijkbare (tussentijdse) resultaten. Dit aspect is momenteel dus voldoende ontwikkeld.
- Genannotatie: Het koppelen van de gennaam aan de gereguleerde genen introduceerde aanzienlijke verschillen. Dit werd veroorzaakt doordat de verschillende programma's hiervoor verschillende annotatiegegevens gebruikten. De oorzaak hiervan is dat de gen-annotatie voor de array (zoals deze op de Affymetrix-website werd verstrekt) regelmatig verandert, omdat de kennis van het rattengenoom nog in ontwikkeling is. Het is hierom van belang dat software tools altijd de meeste recente beschikbare data gebruiken.
- Patroonherkenning: Dit blijkt voor een experiment met meerdere behandelingsgroepen nuttig om de lijst met gereguleerde genen op te delen in groepen (clusters) met soortgelijke dosis-respons trends. Door bij analyses op pathwayverrijking gebruik te maken van zulke clusters kunnen pathway-effecten worden gerelateerd aan een dosis-respons-trend. Hiermee kunnen extra effecten worden herkend die anders niet worden gevonden.
- Pathway-analyse: De verschillende methoden voor pathway-analyse vinden vrijwel allemaal de voornaamste biologische effecten. Voor subtielere effecten worden daarentegen verschillen gevonden tussen de methoden. Het is echter niet zo dat één bepaald programma of type aanpak duidelijk beter was dan de andere. De verschillende methoden zijn onderling aanvullend zodat meerdere methoden nodig zijn voor een goede interpretatie. De gevonden verschillen blijken zowel verklaarbaar uit de gebruikte criteria maar ook uit verschillen tussen de software.
    - Wanneer een pathway door sommige tools wel en door andere als niet significant werd gezien, kwam het voor dat deze pathway marginaal significant was. Bij enigszins andere significantiecriteria zou deze wel (of juist niet) door de verschillende tools als gereguleerd worden herkend.
    - In sommige gevallen werden verschillen veroorzaakt door een ander onderliggend rekenmodel of algoritme. Wanneer bijvoorbeeld van een bepaalde metabole route één gen zeer sterk is gereguleerd en de andere genen niet of nauwelijks, zal software met een algoritme dat uitgaat van een gemiddelde per pathway deze route als gereguleerd beschouwen, terwijl dit bij andere software niet het geval zal zijn. Welke van deze twee bevindingen het meest relevant is voor de onderliggende biologische verschijnselen varieert echter per pathway en per experiment.
    - Voor pathways die uit slechts een beperkt aantal genen bestaan is het algoritme om verrijking te berekenen van invloed op de resultaten. Wanneer bijvoorbeeld van

een pathway met slechts twee genen één gen gereguleerd is, kan dit vanuit een statistisch oogpunt als een significant effect worden gezien. Vanuit biologisch oogpunt is dit echter allerminst zeker. Dit aspect kan worden ondervangen door een ondergrens te stellen aan het aantal gereguleerde genen of door een andere (conservatiever) verrijkings-algoritme te nemen.

- Textmining: Deze aanpak (zie paragraaf 5.3) is momenteel nog duidelijk in ontwikkeling. Hoewel textmining enkele bekende effecten wist te herkennen, leverde het gebruik van textmining geen aanvullende informatie op ten opzicht van pathway-analyses, noch aanwijzingen voor verder literatuuronderzoek.

Samenvattend kan worden gesteld dat er niet één 'beste methode' bestaat. De eerste stappen in de data-analyse zijn voldoende ontwikkeld en leveren betrouwbare resultaten op. Methoden voor analyses op pathway-niveau zijn nog in ontwikkeling en geven onderling aanvullende resultaten. Voor de toekomst staan nieuwe software- en methodologische vergelijkingen gepland rond de analyse van een andere dataset.

2.9 Dataopslag

Hoewel dataopslag geen onderdeel is van de data-analyse is het wel een belangrijk aspect van de bioinformatica, gezien de grote hoeveelheden (ruwe en verwerkte) data die bij de analyses zijn betrokken.
Voor de opslag van arraydata wordt op dit moment gebruikgemaakt van een daarvoor bestemde netwerkschijf. Hierop worden alle gegevens vanaf de beeldverwerking en de verdere analyses opgeslagen. Van deze schijf worden automatisch en regelmatig backups gemaakt.
Het gangbare format voor opslag van arraydata is als tab delimited text, of soms als Excel-bestand. Dit format wordt ook gebruikt voor uitwisseling met andere instituten (met name. MAD, RIKILT) en publieke arraydata-websites zoals ArrayExpress (www.ebi.ac.uk/arrayexpress/) en GEO (www.ncbi.nlm.nih.gov/geo/). Ruwe scanbestanden van microarray-experimenten worden niet op een netwerkschijf opgeslagen, deze worden gebrand op cd en vervolgens apart bewaard. Nadat de beeldverwerking en kwaliteitscontrole heeft plaatsgevonden worden deze bestanden verder namelijk niet meer gebruikt, terwijl ze wel zeer omvangrijk zijn (~30 Mb). Om deze redenen kiezen de meeste arrayunits ervoor om scanbestanden op een aparte manier op te slaan. Naast arraydata kan een array-experiment nog verschillende soorten andere gegevens opleveren. Deze worden door ieder project afzonderlijk opgeslagen en beheerd.
Voor publicatie wordt door sommige tijdschriften verlangd dat de ruwe (en eventueel genormaliseerde) data publiek beschikbaar worden gesteld. Hiervoor worden de arraydata aangeboden aan de publieke microarraydatabanken ArrayExpress (www.ebi.ac.uk/arrayexpress/) en GEO (www.ncbi.nlm.nih.gov/geo/). Een belangrijke aanvullende eis die tijdschriften en deze databanken stellen, is dat alle gegevens voldoen aan de zogeheten 'Minimal Information About a Microarray Experiment' (MIAME). Dit houdt in

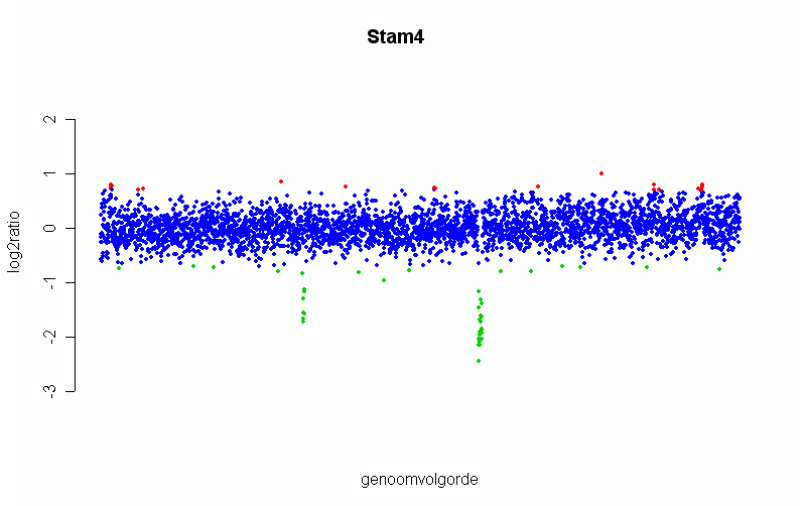
dat naast de arraydata ook informatie moet worden aangeleverd over de gebruikte protocollen en proefopzet, en dat dit volgens een vastgelegde structuur wordt aangeleverd.
Eind 2005 is geïnventariseerd of er bij gebruikers behoefte was aan een database voor gezamenlijke opslag van microarraydata, eventueel aangevuld met andere gegevens. Aangezien het microarray-onderzoek op het RIVM zeer divers is en de meeste gebruikers slechts een beperkt aantal experimenten uitvoeren, is er nu en in de directe toekomst geen behoefte aan een RIVM-brede arraydatabase. Er zal dan ook niet worden geïnvesteerd in een arraydatabase. Over enkele jaren zal opnieuw beoordeeld worden of de situatie zodanig is veranderd dat er wel een database moet worden geïmplementeerd.

3. Bioinformatica ten behoeve van andere genomics-analyses

Naast de microarray-analyses wordt er ook aandacht besteedt aan andere aspecten van de bioinformatica.

3.1 CGH-analyse

Bij Comparative Genomic Hybridization (CGH) wordt bepaald hoeveel kopieën van een gen in een genoom aanwezig zijn. De gebruikte microarrays zijn hierbij in principe vergelijkbaar met arrays die gebruikt worden voor transcriptomics, maar de praktische uitvoering van CGH-experimenten verloopt anders, aangezien wordt uitgegaan van genomisch DNA in plaats van RNA. Op dit moment wordt CGH (ofwel genoomhybridisatie) voornamelijk gebruikt bij LTR voor de typering van kinkhoeststammen. Daarnaast is er interesse in deze techniek voor bacteriële typering (LIS), bioterrorisme (MGB) en het kankeronderzoek (TOX).
Voor de analyse geldt dat stappen als beeldverwerking en kwaliteitscontrole vrijwel gelijk verlopen aan die bij transcriptomicsexperimenten. Het voornaamste verschil in de verdere CGH-analyse is dat de statistische analyse gericht is op de vergelijking met een controle-genoom om te bepalen of er genen zijn waarvan het aantal kopieën verschilt van het controle-genoom. Omdat het verlies (deletie), verwerven of amplificeren van een gen zich vaak niet tot een enkel gen beperkt, wordt in de analyse ook de onderlinge positie van genen op het genoom gebruikt. Voor de normalisatie en statistische analyse zijn dan ook aanvullende algoritmes ontwikkeld (zie bijvoorbeeld Figuur 9). Verdere analyses als pathway-analyses verlopen weer vergelijkbaar met die van transcriptomics.



Figuur 9: Voorbeeld van een bacteriële CGH-weergave. De volgorde van de genen op het genoom staat van links naar rechts weergegeven, met op de y-as de ratio ten opzicht van het controlegenoom. Deleties zijn groen weergegeven, verworven of geamplificeerde genen zijn rood weergegeven.

3.2 Sequentieanalyse

Tijdens het maken van microarrayslides wordt op elk spotje een hoeveelheid DNA van een specifieke oligonucleotide of cDNA-clone aangebracht. De oligonucleotidensets of cDNA-banken waarvan gebruik wordt gemaakt, zijn ontworpen (of verzameld) op basis van de kennis van het genoom zoals die op het moment van ontwerpen van de te spotten sets beschikbaar was. Deze kennis loopt altijd enigszins achter en een regelmatige update van de gennaam en -functie van de op de array gespotte genen is daarom noodzakelijk. Bovendien bevatten de meeste arrays genen waarvan weliswaar een sequentie, maar geen naam of functie bekend is (Expressed Sequence Tags (ESTs)). Door de sequentie en/of het GenBank accessienummer hiervan te vergelijken met de huidige gegevens in publieke of commerciële databanken is het mogelijk voor een aantal van deze clones of ESTs alsnog een gennaam of –functie te bepalen. Een dergelijke situatie doet zich vooral voor bij het gebruik van (weefsel-)specifieke cDNA banken, omdat hier van iedere cDNA-clone in eerste instantie alleen de sequentie bekend is en verdere gegevens daaruit moeten worden afgeleid.
Bij dit soort sequentieanalyses wordt niet alleen gelet op de sequentie van de desbetreffende genen, maar ook op de regio van het genoom waar deze zich bevinden. Zo kan er op worden gelet of er in deze regio nog andere gereguleerde genen aanwezig zijn, en hoe de onderlinge positie van deze genen zich verhoudt tussen soorten (bijvoorbeeld tussen mens en muis). Daarnaast kan worden gecontroleerd of bepaalde genen die bij de ene soort aanwezig zijn, ook bij de andere soort aanwezig zijn en zo ja, of dit in een soortgelijke gencontext is. Dit geeft een beeld van de verschillen tussen species onderling en daarmee over de mate waarin resultaten onderling vertaalbaar zijn. Bij de analyse van prokaryote data (*Bordetella, Neisseria*) wordt tevens gekeken naar de samenhang tussen genregulatie en genoomstructuur, om zo te identificeren welke genomische gebieden gezamenlijk (als operon) worden gereguleerd. Dit geeft aanvullende informatie en blijkt nuttig voor de interpretatie van de resultaten.
Voor de hierboven beschreven sequentieanalyses wordt gebruikgemaakt van de publieke gegevens (o.a. GenBank) die via NCBI (www.ncbi.nlm.nih.gov) beschikbaar zijn. Daarnaast is er een abonnement genomen op de Celera-database voor het humane en muizengenoom. Recentelijk is door de firma de commerciële exploitatie van deze database beëindigd en zijn de desbetreffende gegevens opgenomen in de publieke databanken.

Op het RIVM worden er verschillende studies verricht naar de invloed van genetische variatie op het ontstaan en/of verloop van infectie- en chronische ziekten teneinde risicogroepen in de Nederlandse bevolking te bepalen. Bij dit soort genetische studies wordt gebruikgemaakt van het detecteren van polymorfismes in kandidaatgenen, om zo te bepalen welke genen een rol spelen in de genetische gevoeligheid voor infectie- en chronische ziekten. Binnen enkele van deze projecten (‘Gen-voedingsinteracties’ S/350600, ‘Van gen naar functie’ S/340210) is ondersteuning geboden met het vinden van polymorfismes, met name Single Nucleotide

Polymorphisms (SNP's) om zo relevante kandidaatgenen voor deze studies in kaart te brengen.
Voor het vinden van polymorfismes wordt voornamelijk gebruikgemaakt van de publieke databanken dbSNP (www.ncbi.nlm.nih.gov/SNP/) en de Human Gene Mutation Database (www.hgmd.org/) en vroeger van de gegevens in de Celera-database. Voor het selecteren van kandidaatgenen voor genetische studies wordt daarnaast gebruikgemaakt van de Genetic Association Database (geneticassociationdb.nih.gov/).

Naast het verwerken van sequenties is ook aandacht besteed aan het ontwerpen van PCR-primers ten behoeve van pyrosequencing-analyses. Dit betrof zowel het ontwerpen van sequencing-primers als primers voor multiplex PCR-amplificatie. Hiertoe is zowel gebruik gemaakt van publieke (internet-) software zoals Primer3 (www.bioinformatics.nl/cgi-bin/primer3/primer3_www.cgi, frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi), als van de commerciële Pyrosequencing Assay Design Software. Een tweede toepassing is het ontwerpen van primers voor realtime-PCR-toepassingen. Dit wordt momenteel toegepast in een aantal projecten waarbij realtime-PCR wordt gebruikt voor het meten van genexpressie als validatie of als aanvulling op microarray-experimenten. Hiervoor wordt meestal gebruikgemaakt van het programma Primer Express van de firma Applied Biosystems.

4. Vergelijkingen tussen experimenten

4.1 Inleiding

De in de vorige hoofdstukken beschreven stappen in de data-analyse geven al een behoorlijk compleet beeld bij welke genen en processen expressieveranderingen optreden en hoe deze onderling samenhangen. Voor een verdere interpretatie van de resultaten wordt gebruikgemaakt van kennis die aanwezig is bij de betrokken onderzoekers, aangevuld met literatuurgegevens. De rol van de bioinformatica ligt hier niet zozeer in het uitvoeren van deze interpretatie maar vooral in het helpen ontsluiten van literatuurgegevens en het opwerpen van nieuwe hypotheses. Dit ligt gedeeltelijk in het verlengde van het eerder genoemde koppelen van gennaam aan pathway- en literatuurdata. Daarnaast wordt op een bescheiden schaal textmining toegepast; dit principe zal verderop worden uitgelegd. Een belangrijk aspect bij interpretatie van arraydata is het vergelijken van resultaten met die van andere experimenten. Deze resultaten kunnen zijn verkregen met identieke of juist verschillende agentia, en afkomstig zijn van zowel RIVM-experimenten als literatuurgegevens. Verderop in dit hoofdstuk zal worden uitgelegd hoe dergelijke vergelijkingen kunnen worden uitgevoerd.

4.2 Vergelijkingen tussen vervolgexperimenten

Wanneer binnen een RIVM-project meerdere array-experimenten worden uitgevoerd is het wenselijk om een actueel experiment met de voorafgaande experimenten te kunnen vergelijken. De meerwaarde bestaat hier uit overeenkomsten of verschillen die tussen de experimenten worden gevonden en de nieuwe informatie die dat oplevert.
Bij vergelijkingen binnen hetzelfde project zijn de gegevens doorgaans verkregen met eenzelfde organisme (diermodel) en targetorgaan. Daarnaast is meestal gebruikgemaakt van hetzelfde type array en is de analyse op dezelfde manier uitgevoerd. Deze factoren maken het relatief eenvoudig om resultaten binnen een project te combineren.
Hierbij wordt vergeleken welke genen er zijn gereguleerd en op wat voor manier (inductie, repressie) en in welke mate dit gebeurt. Daarbij kan gebruik worden gemaakt van patroonherkenning (zie paragraaf 2.6) om vast te stellen welke genen in meerdere experimenten op een vergelijkbare manier gereguleerd zijn. Op dergelijke genen kan vervolgens een pathway-analyse worden uitgevoerd (zie paragraaf 2.7). Een andere aanpak is te vergelijken in welke pathways in de afzonderlijke experimenten regulatie plaatsvindt en welke pathways overeenkomen dan wel verschillen tussen de experimenten.

4.3 Vergelijkingen tussen andere experimenten

Naast het vergelijken van vervolgexperimenten kan het interessant zijn minder direct gerelateerde experimenten te vergelijken. Hierbij kan men denken aan het vergelijken van bijvoorbeeld de respons op verschillende stoffen of verschillende pathogenen om zo te onderzoeken welke respons specifiek is voor een bepaald agens of aandoening, maar ook het vergelijken van de effecten in verschillende organen of verschillende dierstammen. In het verlengde hiervan liggen vergelijkingen met experimentele resultaten uit de literatuur.
Bij dit soort vergelijkingen zal er een aantal factoren optreden die het vergelijken en combineren van de data bemoeilijken. Naast eventuele verschillen in het gekozen organisme en orgaan zijn het vooral verschillen in de uitvoering, zoals de proefopzet (onder andere dosis en tijdstip), het gebruikte type array en de data-analyse, die hier een rol bij spelen.
Afhankelijk van de mate waarin de genoemde factoren optreden, is het vaak niet mogelijk om patroonherkenning toe te passen op de genexpressiedata. In plaats daarvan wordt (met behulp van Venn-diagrammen) gekeken naar overlap in genlijsten om zo bijvoorbeeld genen te vinden die overeenkomen dan wel verschillen tussen experimenten. Hiermee kan dan een verdere analyse, zoals pathway-analyse of literatuuronderzoek, worden uitgevoerd. Een dergelijke aanpak werkt beter als voor beide studies hetzelfde type array gebruikt wordt. Hetzelfde geldt wanneer verschillen door data-analyse kunnen worden uitgesloten door op literatuurdata dezelfde analyse uit te voeren als voor de eigen dataset gebruikt is (bijvoorbeeld met dezelfde normalisatie en statistische criteria). Hiervoor is het echter noodzakelijk dat de ruwe data beschikbaar zijn via een publieke database zoals ArrayExpress (www.ebi.ac.uk/arrayexpress/) of GEO (www.ncbi.nlm.nih.gov/geo/).
Voor het vergelijken van pathwayresultaten tussen experimenten leveren de genoemde factoren meestal geen grote problemen op, dit kan dan ook vrijwel altijd worden uitgevoerd. Wanneer er meerdere experimenten zijn die men wil vergelijken, kan men nog een derde soort aanpak toepassen. In plaats van alleen te kijken welke genen bij de afzonderlijke experimenten gereguleerd zijn, kan men kwantitatief bepalen in welke mate genenlijsten onderling overeenkomen, en zo met welke experimenten de grootste overeenkomst bestaat.
Om het uitvoeren van dit soort vergelijkingen te vereenvoudigen is op het RIVM een software-toepassing ontwikkeld genaamd ‘Numerical Overlap Analysis’ of ‘Generic Gene Groups’ (NOAGGG). Hierin wordt een verzameling genen vergeleken met een bibliotheek aan experimentele resultaten. Dit gebeurt in drie stappen: allereerst wordt binnen de software de lijst met genen zoveel mogelijk gestandaardiseerd naar dezelfde gennamen als in de bibliotheek door het gebruik van het NCBI Official Gene Symbol. Daarna wordt tussen de ingevoerde genenlijst en iedere genenlijst in de bibliotheek de overeenkomst berekend, op een manier die vergelijkbaar is met de algoritmes die gebruikt worden voor pathway-verrijkingsanalyses. Daarna worden datasets die aan zelf in te stellen criteria voldoen , gesorteerd op volgorde van overeenkomst.
Op dit moment omvat de NOAGGG-bibliotheek iets meer dan honderd datasets. Het merendeel hiervan betreft RIVM-data die verkregen zijn op muis, rat, en in minder mate *Bordetella*. Daarnaast zijn ook literatuurgegevens over deze organismen en over mens,

zebravis en meningococ opgenomen. Deze applicatie wordt sinds dit jaar in lopende projecten toegepast.

4.4 Mogelijkheden

Naarmate het aantal arraystudies in de literatuur groeit, nemen de mogelijkheden toe om RIVM-resultaten te vergelijken met die van andere laboratoria. Dit geldt vooral op het gebied van stoffen (toxiciteit), kanker en infectieziekten, aangezien dit onderwerpen zijn waarover relatief veel literatuurdata beschikbaar zijn. Wanneer het aanbod aan externe data voldoende kritische massa krijgt, biedt dit zelfs de mogelijkheid om aspecten van het genomicsonderzoek volledig te richten op genexpressiedata uit externe databanken zonder dat deze direct afhankelijk is van eigen RIVM-experimenten. Deze vorm van ‘in silico-research’ zal als eerste toegepast kunnen worden op het gebied van toxicogenomics aangezien op dat gebied de voornaamste initiatieven lopen, zoals de CEBS-database bij het NIEHS (cebs.niehs.nih.gov/microarray/). Daarnaast is een groeiend aanbod beschikbaar aan data op het gebied van kanker, infectieziekten en chronische ziekten, zoals hart/vaatziekten. Hoewel aanbod, kwaliteit en vergelijkbaarheid van dit soort data kritische factoren zullen blijven, zal dit zeker effect hebben op de toepassing van genomics binnen het RIVM. De toekomstige ontwikkelingen op dit terrein zullen dan ook nauwgezet worden gevolgd. De mogelijkheid tot ‘in silico- research’ geldt in wat mindere mate voor geneesmiddelenbeoordeling, daar vertrouwelijkheid van de data hier vaker een belemmering zal vormen voor publieke toegang en uitwisseling.

5. Overige bioinformatica

5.1 Proteomics en metabolomics

Het werk binnen de bioinformatica op het RIVM heeft zich voornamelijk gericht op de analyse en verwerking van genomicsdata. De vraagstellingen binnen het RIVM en de internationale ontwikkelingen waren immers hierop gericht. De technische ontwikkelingen op het gebied van de proteomics zijn vooral het laatste jaar in een versnelling gekomen. Het valt te voorzien dat het RIVM binnenkort nieuwe methoden gaat opzetten voor proteomics en mogelijk metabolomics. Afhankelijk van de gebruikte methodiek en de vraagstelling van dergelijke projecten zal er behoefte ontstaan aan nieuwe kennis om de bijbehorende data-analyse te kunnen uitvoeren. Wanneer het RIVM proteomics- en metabolomicsonderzoek gaat ontwikkelen, zal dan ook capaciteit moeten worden vrijgemaakt voor het ontwikkelen en implementeren van de bijbehorende bioinformatica. Een deel van deze analyse kan in grote lijnen analoog plaatsvinden aan de methodiek die voor transcriptomics wordt gebruikt. Vooral aan de latere stappen (statistiek, patroonherkenning) zal weinig hoeven te worden ontwikkeld. De behoefte aan nieuwe bioinformatica-toepassingen ligt vooral op het gebied van het (grootschalig) voorbewerken en normaliseren van de data. Specifieke software voor het herkennen van individuele eiwitten/peptiden is in toenemende mate beschikbaar, zowel commercieel als publiek. In mindere mate geldt dit ook voor het herkennen van post-translationele eiwitmodificaties, zoals fosforylering en glycosylering. Dit is echter nog niet het geval voor metabolietdata. Methoden voor het combineren van eiwit- of metaboliet-piekenpatronen en -chromatogrammen zijn momenteel in ontwikkeling. Aan wat voor bioinformaticatoepassingen behoefte zal ontstaan, zal echter in belangrijke mate afhangen van de gebruikte methodiek, schaal en vraagstelling van nieuwe projecten op dit gebied. Het is niet aan te bevelen om als RIVM hier het wiel opnieuw uit te willen vinden. Het opbouwen van kennis kan het beste plaatsvinden in samenwerking met andere instituten.

Op dit moment is er binnen het genomicsonderzoek voorzien in de opslag van arraydata (zie paragraaf 2.9). Wanneer op het RIVM nieuwe methoden worden opgezet voor proteomics en metabolomics, zal er ook behoefte ontstaan aan opslagcapaciteit voor dit soort data. De ruwe bestanden met chromatogrammen en piekenpatronen van bijvoorbeeld LC-MS-MS-analyses zijn tamelijk omvangrijk (10-50 MB). In tegenstelling tot microarraydata zijn deze ruwe data ook voor verdere analyses nodig. Hiervoor zal dan ook in opslagcapaciteit voorzien moeten worden. Door een uitbreiding van de bestaande infrastructuur, bijvoorbeeld door middel van een extra netwerkschijf, kan dit echter zonder grote nieuwe investeringen bereikt worden.

5.2 Systems biology

Een onderzoeksveld dat de laatste jaren steeds vaker wordt genoemd is systems biology, in het Nederlands ook wel systeembiologie genoemd. Hierin wordt gestreefd naar een (kwantitatieve, modelmatige) beschrijving van complexe biologische processen op het niveau van de betrokken moleculaire componenten (genen, eiwitten, metabolieten). Daarvoor worden experimentele data gecombineerd met wiskundige modellering. De experimentele data zijn vaak afkomstig van multidisciplinair onderzoek, zoals genexpressiedata, eiwitexpressiedata en metabole data. Naast dit soort metingen aan celcomponenten worden ook andere meer celbiologische gegevens gebruikt, zoals gegevens over celdeling, apoptose of morfologie, die op een ander niveau beschrijven wat voor processen en effecten er plaatsvinden. Deze gegevens worden gecombineerd door hun onderlinge verband mathematisch te beschrijven. Systeembiologie heeft wortels in zowel de moleculaire als de theoretische en fysische biologie [4]. Voor wat betreft het eerste aspect geldt dat door de ontwikkelingen op genomicsgebied de laatste tien jaar de experimentele mogelijkheden enorm zijn toegenomen. Dit heeft geleid tot een groter aanbod aan data en daarmee mogelijkheden tot modelleren. Het modelleren van de betrokken biologische processen steunt op de kennis vanuit de theoretische en fysische biologie.

Op dit moment staat systeembiologie voornamelijk in de belangstelling vanwege de doelstelling om verschillende soorten biologische data te kunnen integreren (zoals genexpressiedata en proteomicsdata), met als uiteindelijk doel een beter inzicht in de betrokken biologische processen. Dit sluit aan op de traditie die er in Nederland met name in de microbiële fysiologie al bestaat, om multidisciplinaire data te verzamelen en te integreren (zie bijv. referentie [5]), zij het dat nu de aandacht vooral wordt gericht op humane en proefdiermodellen.

Voor het RIVM is systeembiologie interessant vanwege de mogelijkheden om verschillende soorten data uit het biomedische onderzoek beter te kunnen combineren (bijv. genexpressie, proteomics, pathologie, literatuurdata). Ook voor het beleidsondersteunende onderzoek van het RIVM kan een systeembiologische aanpak bruikbaar zijn. Op het gebied van bijvoorbeeld toxicologie kan men denken aan modellering van een blootstellingrespons op meerdere typen data. Dit kan uiteindelijk leiden tot een verfijning van de risicoschatting op stoffengebied. Een belangrijke voorwaarde voor dit laatste is het beschikbaar komen van voldoende (betrouwbare) data, zodat daarmee op basis van een model ook kwantitatieve uitspraken kunnen worden gedaan. Kennis op dit gebied sluit nauw aan bij de huidige analyses en wordt opgebouwd via onder andere samenwerking met de MicroArray Department (MAD) en Integrative Bioinformatics Unit (IBU) van de Universiteit van Amsterdam.

Samenvattend biedt systeembiologie nieuwe mogelijkheden om verschillende soorten data te combineren en te modelleren; dit terrein zal zich de komende jaren verder ontwikkelen.

5.3 Textmining

Bij genomicsexperimenten worden vaak enkele honderden gereguleerde genen gevonden. Aangezien het bij de interpretatie hiervan op praktische bezwaren stuit om voor ieder gen afzonderlijk literatuur op te zoeken, ontstaat behoefte aan het geautomatiseerd doorzoeken van literatuur op basis van meerdere zoektermen om zo nieuwe relevante informatie te extraheren. Dit proces wordt textmining genoemd.
De meeste toepassingen van textmining zijn ontwikkeld op basis van het principe dat wanneer zoektermen een onderling verband hebben, zij vaak samen in publicaties voor zullen komen. Voorbeelden van deze aanpak zijn de publieke tools PubMatrix (pubmatrix.grc.nia.nih.gov) en MeSHer (biocomp.dfci.harvard.edu/mesher.html), en het commerciële pakket Collexis (www.collexis.nl/). Een recentere ontwikkeling is het zogeheten natural language processing. Daarbij wordt niet alleen gekeken naar het al dan niet gezamenlijk voorkomen van zoektermen, maar ook naar de tekstuele context waarin dit gebeurt, zodat nauwkeuriger informatie kan worden geëxtraheerd (bijv.: ‘bvgR represses fim3’).
Binnen enkele projecten is gebruikgemaakt van de genoemde drie programma’s. Daarbij bleek dat textmining op dit moment een nuttig middel kan zijn om literatuurgegevens in kaart te brengen, maar nog duidelijk in ontwikkeling is. Voor de verdere ontwikkeling van textmining zullen enkele praktische hindernissen opgelost moeten worden. De belangrijkste hiervan is dat voor een gennaam soms meerdere synoniemen bestaan, en omgekeerd dat soms een gennaam een homoniem kan zijn voor meerdere genen. Naast het ontwikkelen van nieuwe textmining-software en -algoritmes zal ook verdere standaardisering van de (gen-)nomenclatuur belangrijk zijn voor een succesvoller gebruik van textmining.

6. Informatie-uitwisseling

Overleg binnen het RIVM
Voor microarray-gebruikers is er een vierwekelijks overleg opgezet. Dit overleg is bedoeld voor arraygebruikers die zelf direct betrokken zijn bij de experimenten en is vooral gericht op experimentele zaken, zoals de proefopzet van komende experimenten, resultaten, en het plannen van vervolgexperimenten. Hiernaast wordt ook aandacht besteed aan bioinformatica, zoals statistische algoritmes en methoden voor pathway-analyse. Deelnemers aan dit overleg zijn afkomstig van de afdelingen TOX, LTR, MGO, LIS en BMT, evenals het NVI. Een overzicht van de projecten die bij dit overleg betrokken zijn staat gegeven in Bijlage II van RIVM-rapport 340200001, 'Genomics: Implementatie, toepassing en toekomst'.

Voor gebruikers en andere geïnteresseerden is een intranetsite opgezet waarop informatie, protocollen en publicaties van de array- en bioinformatica-unit zijn te vinden. Deze site is te vinden op http://tox/array/.

Externe overlegorganen
Vanuit het RIVM is (samen met de MAD uit Amsterdam) het initiatief genomen voor het oprichten van een landelijk platform, gericht op het uitwisselen van informatie met betrekking tot de biologische interpretatie van arraydata. In dit platform (Biomax) zijn bioinformatici van zestien instituten uit Nederland vertegenwoordigd. Via dit platform bespreken zij ervaringen en verschillende softwaretools. Dit gebeurt aan de hand van een gezamenlijke analyse van een casus-dataset.

Vanuit het RIVM vindt regelmatig overleg plaats met de MicroArray Department (MAD) en Integrative BioInformatics Unit (IBU) van de Universiteit van Amsterdam, onder andere via maandelijkse literatuurbesprekingen. Verder is er vooral regelmatig overleg met bioinformatici van het RIKILT, TNO Voeding en de Universiteit Maastricht. Samen met deze laatste drie instituten werkt het RIVM samen in het Nederlands Toxicogenomics Centrum (NTC), waarbij verder ook het Erasmus MC, het Leids Universitair Medisch Centrum, het Leiden/Amsterdam Center for Drug Research en de Wageningen Universiteit zijn betrokken.

Daarnaast is het RIVM vertegenwoordigd in de volgende nationale samenwerkingsverbanden op genomics- en/of bioinformaticagebied:

- ArrayNL (platform voor microarray-onderzoek in Nederland);
- MicroArray Operators Platform (MAOP), hierin worden technische ontwikkelingen op microarraygebied besproken;
- NVBMB Werkgroep Bioinformatica;
- Gebruikersoverleg NBIC (Nederlands BioInformatica Centrum), voorheen BioASP (Nationale Bioinformatica Applicatie Service Provider).

Tot slot participeert het RIVM in de volgende internationale verbanden:

- ILSI/HESI (International Life Sciences Institute / Health and Environmental Sciences Institute, Washington DC, USA);
- NUGO (European Nutrigenomics Organisation).

7. Conclusies

- Het SOR-project S/340200: 'Genomics' is januari 2003 van start gegaan met als deelprojecten het opzetten van de microarray-unit en de bijbehorende bioinformatica. Het doel was om hiermee de microarraytechniek beschikbaar te maken voor RIVM-projecten en te integreren in het lopende onderzoek. Hierin is het project geslaagd. Dit rapport behandelt de daarbij behorende dataverwerking en bioinformatica.
- Voor de beeldverwerking, kwaliteitscontrole, normalisatie en (grootschalige) statistiek van microarraydata zijn protocollen en algoritmes ontwikkeld en geïmplementeerd. Deze worden gebruikt in de analyse van transcriptomics- en CGH-experimenten. Dit gedeelte van de analyse verloopt succesvol en is dermate voldoende ontwikkeld dat het betrouwbare resultaten oplevert.
- Voor de interpretatie van de resultaten wordt gebruikgemaakt van het geautomatiseerd koppelen van gennaam aan -functie en software voor pathway-analyses. Op dit gebied loopt het RIVM gelijk op met de (internationale) ontwikkelingen. Dit deel van de data-analyse maakt momenteel nog een verdere ontwikkeling door. Hiervoor werkt het RIVM samen met andere instituten, onder andere via het Biomax-platform.
- In de nabije toekomst zullen meer arraydata (publiek) beschikbaar komen. Hierdoor neemt de behoefte toe aan het onderling vergelijken van RIVM-experimenten en het vergelijken en combineren van resultaten met literatuurdata. Voor projecten op het gebied van toxicogenomics, infectieziekten en chronische ziekten zal dit een belangrijke rol zal gaan spelen voor de interpretatie van de resultaten. Daarnaast ontstaat de mogelijkheid voor 'in silico-research' op basis van genexpressiedata uit publieke databanken, allereerst op het gebied van toxicogenomics.
- De komende jaren zullen naast arraydata ook meer andersoortige data, zoals eiwit- en metabolietgegevens beschikbaar komen. Door het integreren van deze multidisciplinaire data, ook wel 'systems biology' genoemd, kan een completer beeld worden verkregen hoe de betrokken biologische processen gereguleerd worden. Voor het RIVM is deze aanpak vooral interessant voor een verfijning van de risicoschatting op stoffengebied.
- De komende jaren zullen nieuwe methoden voor proteomics (het grootschalig bestuderen van eiwitten) en mogelijk metabolomics RIVM-breed worden opgezet. Proteomics zal in de nabije toekomst een belangrijke rol gaan spelen bij bevolkingsonderzoeken en in screeningsprogramma's van micro-organismen. Voor deze technologieën zullen nieuwe bioinformatica-aspecten moeten worden ontwikkeld. Daarnaast bieden de ontwikkelingen op het gebied van systeembiologie mogelijkheden om multidisciplinaire soorten data beter te kunnen combineren ten behoeve van biochemisch en toxicologisch onderzoek.

Literatuur

1. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003; 31(4):e15.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J.Roy.Stat.Soc.B. 1995. 57, 289-300.
3. Ezendam J, Staedtler F, Pennings J, Vandebriel RJ, Pieters R, Harleman JH, Vos JG. Toxicogenomics of subchronic hexachlorobenzene exposure in Brown Norway rats. Environ Health Perspect 2004; 112(7):782-791.
4. Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Nat Biotechnol 2004; 22(10):1249-1252.
5. Pennings JLA, Keltjens JT, Vogels GD. Isolation and characterization of *Methanobacterium thermoautotrophicum* ΔH mutants unable to grow under hydrogen-deprived conditions. J Bacteriol 1998; 180(10):2676-2681.

Bijlagen: protocollen

Op de volgende pagina's zijn de momenteel beschikbare protocollen en handleidingen verzameld. Deze protocollen worden actueel gehouden door de microarray-unit. Aangezien sommige protocollen in de toekomst aangepast, dan wel verbeterd zullen worden, wordt erop gewezen dat de actuele protocollen te vinden zijn op intranet via http://tox/array/.

Bijlage 1: Protocol Image Analysis

- Voor dit deel van de data-analyse wordt gebruik gemaakt van ArrayVision. Momenteel is ArrayVision geïnstalleerd op de computer van Jeroen en de data-analyse-computer. Na het opstarten van ArrayVision kun je de volgende stappen doorlopen.
- Het maken van een protocol. De eerste keer kun je het beste hulp inroepen van Jeroen. Bovendien is voor veelgebruikte types gespotte arrays het protocol meestal standaard aanwezig. Bewaar dit eventueel bij je data. Vaak kun je op basis van zo'n protocol makkelijk een nieuwe versie maken. Het onderstaande dient meer als leidraad voor wijzigingen.
  - Start ArrayVision en open de Protocol editor. Kies via File (linksonder) → select protocol en kies het protocol dat je wilt wijzigen. Doorloop de onderdelen van het protocol als volgt.
  - Analysis: Fluorescent, comparative, 2 arrays, single template mode, 1 extra.
  - Images: Use image filename as channel name aanvinken.
  - Layout: het aantal rijen en kolommen moet je instellen afhankelijk van de layout van je slide, evenals de onderlinge afstand en spotafstand. Roep hier maar hulp in van Jeroen, of laat hem eventueel het hele protocol maken.
  - Blanks: Enable definition of blanks moet uitstaan.
  - Display: display spots, spots outline.
  - Labels: use labeling scheme, overleg met Jeroen over de manier van labelen.
  - Measures: Principal measure = median density, background = median.
  - Background: enable background subtraction aan, corners between spots, defined size 2, individual.
  - References: uit.
  - Segmentation: uit.
  - Quality control: alle getallen op 10% zetten.
  - Alignment: beide getallen op 7 (dit werkt bijna altijd goed).
  - Anchors: uit.
  - Post-analysis: export as tab-delimited text.
  - Vervolgens OK en opslaan.
- De beeldanalyse zelf
  - Start ArrayVision en kies Run the analysis wizard.
  - Selecteer het protocol dat je wilt gebruiken en neem next.
  - Vervolgens moet je de imagefiles laden. Dit zijn Tiff files (*.tif, géén Tif5!). Je moet twee files selecteren, namelijk de Cy3 en de Cy5. Selecteer eerst de Cy5 en daarna de Cy3 file, dan staat in het resultaat eerst Cy3 en dan Cy5 (dat werkt makkelijker). Ga verder met Next.
  - In het alignment-menu kies je Position en klik je het spotje aan dat het meest linksboven staat. Hierna OK.
  - Positioneer de gridblokken over de spotjes.
    - Als je protocol goed in elkaar zit klopt dit al behoorlijk, maar je kunt de blokken rekken en verplaatsen zoals dat in windows gebruikelijk is.
    - Om spotjes goed te kunnen zien helpt het vaak als je de kleurinstellingen wijzigt. In het Visuals-menu (standaard in beeld) kies je linksboven voor (bijvoorbeeld) Ca.vis of Spect2.vis, rechtsboven zit de knop om de kleursterkte te wijzigen. Meestal betekent dit de lijn in de grafiek naar links trekken. De curve op hyperbole zetten werkt vaak het prettigst.
    - Een enkel blok selecteren is ctrl-klik, via ctrl-klik links van of boven een rij blokken selecteer je ze allemaal. Na afloop helemaal linksboven klikken zodat alle blokken geselecteerd worden.
    - Als alles goed ligt kun je Align kiezen en als de computer daarmee klaar is wederom Next. Echter: als je veel spotjes hebt met een laag signaal is het alignen niet aan te raden. Controleer dan liever een keer extra of alle spotjes goed overeenkomen met het grid.
  - Kies nu Perform sampling and post-analysis operations en klik op Finish.
  - Sla nu de data op onder een naam die bij de slides past (1234_Cy3.tif en 1234_Cy5.tif samen in 1234.txt).

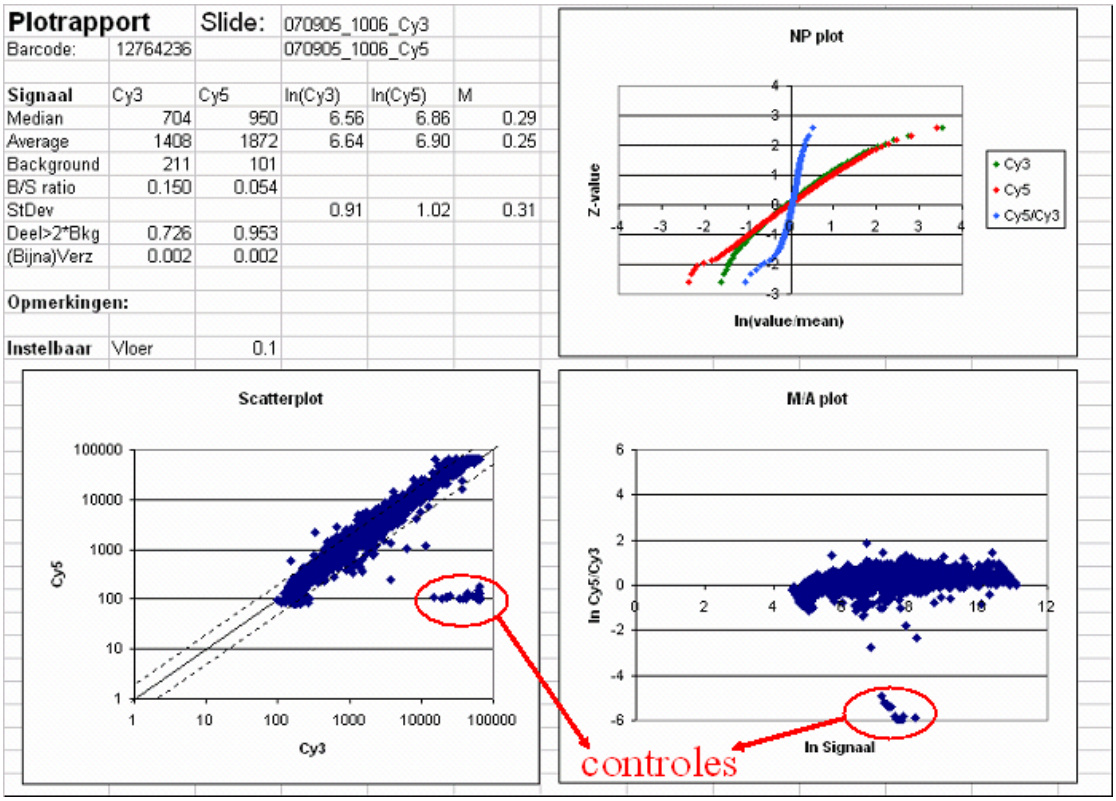
- Na afloop de data verzamelen (alleen indien gewenst):
  - Zet je computer op Engelse settings (VS/UK/Australië/enz) via Start --> Settings --> control panel --> regional settings.
  - Start Excel op en laadt de output files (File --> open --> kijk via all files). Klik twee keer op Next en dan op Finish. Herhaal dit tot je een logische groep slides bij elkaar hebt.
  - Kopieer de volgende gegevens naar een nieuwe file: Spot labels, Cy3 signaal (Median Dens - Levels) en achtergrond (Bkgd), Cy5 signaal en achtergrond. Combineer slides die logischerwijze bij elkaar horen in een gezamenlijk werkblad, maar houdt het aantal redelijk, dwz liever niet meer dan twintig slides per werkblad. Sla de file op als Excel-bestand.
  - Geef de kolommen eventueel een andere (handigere) naam. Let dan goed op wat Cy3 en Cy5 is, controleer of het bij het laden in ArrayVision nergens verwisseld is en zorg dat je ze nu ook niet verwisselt.
  - Zet na afloop de settings eventueel weer terug op Nederlands.
  - Voor geoefende gebruikers is het mogelijk dit proces te automatiseren via R. Zie daarvoor de desbetreffende handleiding. Dit is alleen rendabel voor grotere proeven.
- Veelgestelde vragen en andere handige dingen om te weten:
  - Cy3 = controle/data1 = groen = 532 nm, Cy5 = data(2) = rood = 635 nm.
  - Op de slide is 1 pinafstand (tussen blokken, level 3) = 4,5 mm.
    In ArrayVision is 1 pinafstand = 3 inch = 450 pixels.
  - We verzamelen de info over de achtergrond met het oog op kwaliteitscontrole en het beoordelen van een spotje. We trekken deze echter niet af. Achtergrondaftrek voegt namelijk ruis toe aan je data, geeft afwijkende ratio's voor genen die laag tot expressie komen, en blijkt nadelig bij het opsporen van gereguleerde genen. Denk maar zo: als het aftrekken van je achtergrond een groot effect heeft op je resultaten geeft, dit eigenlijk al aan dat er iets mis was met je spotje.

Bijlage 2: Protocol Kwaliteitscontrole (QC)

- Na de image analysis moet je eerst kijken of de slide naar behoren gelukt is, voordat je verder gaat met de data. Meestal zal Jeroen dit doen, of in ieder geval helpen met het beoordelen. De QC omvat:
- Visuele inspectie van de slide. Dit gebeurt tijdens het scannen en ArrayVision. Hierbij kijk je onder andere naar de achtergrond en de spotmorfologie. Als het goed is heb je mooie ronde spotjes, een gelijkmatige achtergrond en geen sporen van stof, krassen, of luchtbellen. Als je tijdens het scannen en/of ArrayVision iets opgevallen is, maak dan een aantekening voor verderop in de QC.
- Voor de andere onderdelen is in Excel een template-file gemaakt (plotreport.xls). Door je waardes voor je signaal en achtergrond vanuit je ArrayVision output in werkblad Plot1 te plakken (kolom A-G) en kolom H-M door te voeren naar beneden worden er in werkblad Plot2 getallen en grafiekjes berekend. Dat rekenwerk duurt ongeveer een minuut. Let er wel op dat als je datakolommen maar één header-regel hebben je eerst de bovenste rij uit blad Plot1 verwijdert.
- Deze aanpak levert een aantal getallen op:
  - Waarde gemiddelde, mediaan, standaarddeviatie van het signaal, mediaan achtergrond en verhouding signaal/achtergrond. Deze statistiek dient voor het onderling vergelijken van slides.
  - Het aandeel van de spotjes met een signaal > 2 * achtergrond. Dit komt ruwweg overeen met het aantal spotjes met een meetbaar (bruikbaar) signaal. Als dit aantal erg laag is (bijv. < 10%) wijst dat op een slechte labeling.
  - Het aandeel van de spotjes met >95% van de verzadigingswaarde. Dit geeft aan in hoeverre je signaal verzadiging vertoont. Dit percentage moet niet te hoog zijn, liefst <1 %. Een paar verzadigde spotjes is echter nooit helemaal te voorkomen, want sommige genen komen nou eenmaal erg hoog tot expressie.
- Verder komt er een aantal plotjes:
  - De scatterplot en MA-plot voor vorm signaalcurves. Deze geven informatie over de homogeniteit van de verdeling (lengte en breedte puntenwolk). Als het goed is, krijg je een sigaarvormige puntenwolk die in de MA-plot horizontaal ligt. Eventuele afwijkingen hierop kunnen veroorzaakt worden door dye-bias, autofluorescentie Cy3, scannerproblemen of verzadiging. In dat laatste geval kan opnieuw scannen soms helpen.
  - De NP- plot, die iets zegt over de signaalverdeling van Cy3, Cy5 en hun ratio. Een rechte lijn op deze plot geeft aan dat je signaal of de Cy5/Cy3 ratio log-normaal verdeeld is. Dat hoeft echt lang niet altijd zo te zijn, maar de curves voor Cy3 en Cy5 zijn meestal vergelijkbaar en de ratio is meestal wel redelijk normaal verdeeld. Onverklaarbare vreemde afwijkingen hierin of in de ratio kunnen wijzen op een probleem; je kunt dit ook al herkennen aan een verschillende standaarddeviatie voor Cy3 en Cy5. Vaak kan dit opgelost worden door een normalisatie/correctie op de data.
- Daarnaast kun je in werkblad Plot2 nog extra opmerkingen kwijt. Denk hierbij aan de naam (of barcode) van de slide, technische problemen, of opvallende zaken n.a.v. de visuele inspectie.
- Kopieer het QC-rapport (het gedeelte tussen de lijnen) om het te plakken in Powerpoint. Kies hiervoor Edit --> Paste Special --> Device Independent Bitmap. Zo kun je een PowerPoint file maken met alle QC-data bij elkaar om de slides makkelijk onderling te vergelijken.

- Na de QC beslis je of een slide wel of niet verder gebruikt wordt. Jeroen helpt bij het interpreteren van de QC en gezamenlijk beslis je of er slides moeten worden afgekeurd. Het is echter nagenoeg onmogelijk om bij voorbaat criteria te geven waaraan een slide moet voldoen om goed- dan wel afgekeurd te worden. De eisen zijn namelijk mede afhankelijk van het type slide en het gelabelde materiaal. Vergelijk getallen en grafiekjes van een slide daarom met andere slides uit dezelfde serie en eventuele andere vergelijkbare series uit dezelfde proef. Afwijkingen van wat gangbaar is in een serie zijn bijna altijd in negatieve zin. Als je enige ervaring hebt opgedaan met deze stap herken je al snel welke slides problemen zullen geven.
- Gangbare vuistregels voor het afkeuren van slides (maar dit zijn dus geen harde regels!):

- Als de standaarddeviaties van Cy3 en Cy5 meer dan 0,5 verschillen wordt een slide afgekeurd.
- Dat geldt ook als er in de MA-plot een verloop zit van meer dan een factor 10.
- Als het aantal spotjes met een signaal van meer dan twee keer de achtergrond minder dan de helft is van wat gebruikelijk is, is dit vaak ook reden om een slide af te keuren of kritisch te bekijken.
- Vlekken en achtergrond-smeer op de slides kunnen ook reden zijn om een slide af te keuren, al zie je dit niet altijd terug in de getallen.



Plotrapport		Slide:	070905_1006_Cy3		
Barcode:	12764236		070905_1006_Cy5		
Signaal	Cy3	Cy5	ln(Cy3)	ln(Cy5)	M
Median	704	950	6.56	6.86	0.29
Average	1408	1872	6.64	6.90	0.25
Background	211	101			
B/S ratio	0.150	0.054			
StDev			0.91	1.02	0.31
Deel>2*Bkg	0.726	0.953			
(Bijna)Verz	0.002	0.002			
Opmerkingen:					
Instelbaar	Vloer	0.1			

Bijlage 3: Protocol Normalisatie en Statistiek

- Deze analyses lopen vaak gedeeltelijk over in andere delen van de data-analyse, zoals de QC en de datamining. De exacte grens is soms lastig te trekken. Dit gedeelte is dan ook bedoeld voor de wat meer gevorderde gebruiker, en deze stap zal plaatsvinden door of na overleg met Jeroen.
- Er wordt voornamelijk gebruikgemaakt van het programma R, voor meer informatie hierover wordt verwezen naar het Handleiding Grootschalige Arraystatistiek.
- Verzamel de ruwe data in een Tabel. Voor de verdere verwerking wordt uitgegaan van een bepaald formaat invoer. Er is één rij (de eerste) met header-informatie per kolom. Daarna zijn er drie datakolommen met spotinformatie, daarna komen de echte data. Dit omvat per slide de median density voor eerst Cy3 en dan Cy5 (tenzij het om single dye-experimenten gaat). De eerste drie kolommen zijn als volgt:
  - De eerste datakolom bevat een unieke waarde voor iedere spot (bijvoorbeeld het spotnummer uit ArrayVision).
  - De tweede kolom bevat een letter die aangeeft om wat voor type spotje het gaat. A = algemeen, in deze spotjes ben je geïnteresseerd. B = blanco (leeg, spotbuffer, ...). C = controle, denk hierbij aan positieve of negatieve controles zoals luciferase / *Salmonella* / *Arabidopsis* genen of de ‘landing lights’ voor de gridpositionering. Andere letters staan voor andere soorten spotjes die niet tot je feitelijke experiment behoren. Alleen de A-spotjes worden gebruikt voor de analyse.
  - De derde kolom geeft aan wat er in ieder spotje zit, zodat replicaspotjes kunnen worden gemiddeld. Alle spotjes in deze kolom met dezelfde naam worden na normalisatie gemiddeld.
- De data worden genormaliseerd in R, dit bestaat uit een paar stappen. Ten eerste het verwijderen van niet-relevante spotjes (alle spotjes zonder A). Vervolgens een ln-transformatie en quantile normalisatie op alle scans. Indien van toepassing wordt dit gevolgd door correctie van het sample-signaal door het referentie-signaal (op basis van de ratio per spotje, Cy3 --> Cy3/Cy5 * gemCy5). Tot slot worden alle replicaspotjes gepoold. Dit alles verloopt automatisch en het resultaat wordt automatisch opgeslagen. Gedetailleerde uitleg valt te vinden in de Handleiding Grootschalige Arraystatistiek, al is het voor niet-getrainden niet erg gebruikersvriendelijk. Vandaar dat vooral Jeroen er mee zal werken.
- Ook de statistische analyse wordt uitgevoerd in R. Bereken in R eerst per gen een one-way anova tussen alle groepen. Hieruit krijg je voor ieder gen een p-waarde en een maximale FoldRatio. Uit deze data kun je de interessante hits halen volgens bepaalde criteria (bijv. $p<0.001$, $FR>2$). Hierbij krijg je te zien hoeveel genen significant zijn en hoeveel vals-positieve je verwacht. Het is prettig als je lijstje significante genen maar weinig vals-positieve bevat (liefst minder dan 10%). Dit lijstje significante genen wordt verder verfijnd met een minimale FoldRatio om biologisch niet-relevante effecten eruit te halen. Deze FoldRatio ligt vaak op een factor 2, dat is namelijk te bevestigen met Q-PCR. Soms wordt deze op een factor 1,5 gezet, om ook subtiele effecten op te pikken. Veel lager is niet altijd zinnig. Hoe je de p-waarde en FR kiest, hangt af van het experiment.
- Na deze stappen heb je een lijstje met significant gereguleerde genen. Hiermee ga je door naar de volgende stappen.
- De exacte aanpak is afhankelijk van de opzet van je experiment en eventuele praktische omstandigheden. Hoe dan ook: hanteer dezelfde aanpak (voor normalisatie en/of standaardisatie) binnen één experiment. Vergelijken tussen compleet verschillende experimenten doe (en kun) je toch niet. Mocht het overigens nodig zijn, dan kun je immers vanaf de ruwe data altijd een nieuwe analyse uitvoeren. Bij twijfel kun je Jeroen Pennings om raad vragen.

Bijlage 4: Handleiding Grootschalige Arraystatistiek

Dit is een korte uitleg bij het gebruik van R voor de analyse van microarraydata. Hoewel een groot deel van de analyses in Excel plaatsvindt, heeft dit programma als nadeel dat het niet primair een statistisch programma is, en zowel qua mogelijkheden als qua snelheid stuit dit soms op beperkingen. Vandaar dat voor het genomicsproject een aantal nieuwe mogelijkheden in huis is gehaald en getest.
Voor grootschalige arraystatistiek zijn er momenteel binnen het RIVM de volgende mogelijkheden:

- GeneMaths XT van Applied Maths. Speciale software voor microarray-analyse, met veel mogelijkheden (ook met betrekking tot statistiek en normalisatie). Er zijn twee netwerklicenties.
- Splus. Krachtig statistisch programma (en tevens een taal), wordt ook voor andere statististische toepassingen veel gebruikt op het RIVM. Nadeel van dit programma is dat je een licentie nodig hebt en het programma niet gebruikersvriendelijk is.
- R. Dit is een open source programma dat verder nagenoeg hetzelfde werkt als Splus. R is gratis en kan op elke computer geïnstalleerd worden, al is het net als Splus niet erg gebruikersvriendelijk. Deze inleiding richt zich op het gebruik van R voor microarray-analyse in combinatie met de RIVMarray-library.

Zoals genoemd is R een gratis en open source programma en taal voor statistiek in het algemeen, maar deze is de laatste jaren vooral populair geworden onder microarraygebruikers. Het programma zelf kent een aantal standaard aanwezige functies, maar het is mogelijk nieuwe te schrijven en deze code via een library of een package uit te wisselen of verder te verspreiden. Deze zijn meestal ook te gebruiken onder Splus. Binnen de microarraygemeenschap zijn er al veel libraries en packages beschikbaar die speciaal hiervoor zijn geschreven (vooral binnen het BioConductor project), en regelmatig worden er in de literatuur nieuwe beschreven. Van de beschikbare libraries en packages wil ik er twee speciaal noemen. Allereerst het DNAMR package van Amaratunga & Cabrera. Deze bevat veel functies die voor array-analyse erg handig zijn. Ten tweede de RIVMarray-library die speciaal geschreven is voor de microarray-analyses op het RIVM. Deze maakt op zijn beurt overigens weer gebruik van DNAMR en werkt niet zonder.
De genoemde software is te vinden op:

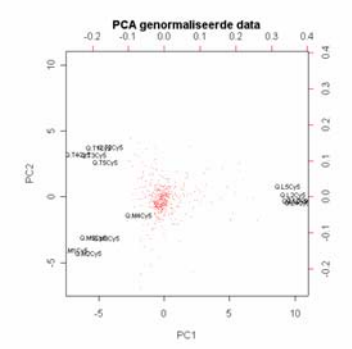
- R home page: http://www.r-project.org/
- BioConductor: http://www.bioconductor.org/
- DNAMR: http://www.rci.rutgers.edu/~cabrera/DNAMR/
- RIVMarray: http://tox/array/

RIVMarray

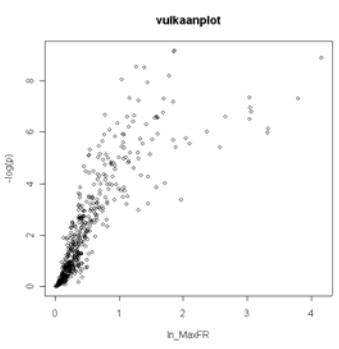
Deze library is geschreven voor de meest voorkomende handelingen bij array-analyse op het RIVM. Voor het gebruik ervan wordt uitgegaan van iemand die enigszins ervaren is met het gebruik van R (of Splus). R voorziet zelf in een prima manual en een redelijk uitgebreide helpfunctie, als je meer over R wilt weten kun je hier vast vinden wat je zoekt. De RIVMarray-library valt te vinden op de tox-array site. Op deze site is tevens een file met voorbeelddata te vinden (rivmarraydata.txt). Deze oefendata bestaan uit vijftien slides met elk vijfhonderd obligo's in meervoud gespot. Er zijn drie groepen samples (L, M, T) met elk vijf biologische replica's. De samples zijn gelabeld met Cy5 en Cy3 bevat een referentiepool.
Hieronder is een voorbeeldsessie gegeven. Om de werking uit te leggen zullen we de tekst verderop even doorlopen.

Voorbeeldsessie 1

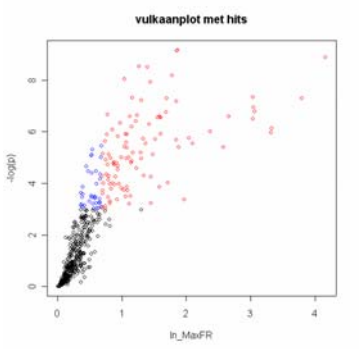
```
> source('rivmarray.r')
DNAMR geladen
RIVMarray geladen
Type uitleg() voor meer hulp
Aan de slag dus maar!

> f.arv.verzamel()
Laden bestanden:   12345001.txt   12345002.txt   12345003.txt   12345004.txt
    12345005.txt   (klaar!)
Verzamelde data opgeslagen als ArrayVision_Verzamel.txt
>
> data1<- f.laaddata('rivmarraydata.txt')
> data2<- f.normaliseer(data1,3)
Begin normalisatie om  Tue Mar 01 15:24:59 2005
Stap 1: Opschonen arraydata: gebeurd
Stap 2: Beschrijving data
aantal spotjes = 1364 , aantal dataspotjes = 1058 , aantal niet-dataspotjes = 306
aantal labelscans = 30
signaalwaardes minimum = 76 , maximum = 65535
sd arrays minimum = 0.8964335 , maximum = 1.372286 , verschil = 0.4758526
NPP-QC minimum = 0.002348214 , maximum = 0.2282498
Resultaten opgeslagen als laatste_QC_Tabel.txt
Stap 3: Log-quantile normaliseren: gebeurd
QNln-waardes minimum = 4.739153 , maximum = 11.05568
Stap 4: Cy3 als referentie gebruikt. Slides zijn
Q.L1Cy5 Q.L2Cy5 Q.L3Cy5 Q.L4Cy5 Q.L5Cy5 Q.M1Cy5 Q.M2Cy5 Q.M3Cy5 Q.M4Cy5 Q.M5Cy5
Q.T1Cy5 Q.T2Cy5 Q.T3Cy5 Q.T4Cy5 Q.T5Cy5
Stap 5: middelen replicaspotjes. Nieuwe aantal =  481
Normaliseren klaar
Resultaten opgeslagen als laatste_genormaliseerde_data.txt
PCA van de monsters staat in andere venster                    FIGUUR 1
Klaar om  Tue Mar 01 15:25:05 2005
> data3<- f.OWA(data2, c(rep(1,5),rep(2,5),rep(3,5)))
Varianties: between.var = 1.778719 , within.var = 0.5617024 (voor poweranalyse)
Gepoolde SD = 0.2003038 , dwz een factor 1.221774 (voor biologische analyse)
Resultaten opgeslagen als laatste_statistieken.txt             FIGUUR 2
> data4<- f.neemPhits(data3, .001, 2)
Selectie hits: 481 entries, 0 vals positief
Je vindt 140 significante hits waarvan 105 met een FR > 2
Resultaten opgeslagen als laatste_hits.txt                     FIGUUR 3
> pairs(data4[,1:3])                                           FIGUUR 4
> data5<-data2[rownames(data4),]
> f.opslaan(data5, 'clusterdata.txt')
> f.tweedee(data5)                                             FIGUUR 5
Kleurschaal = 3.007013
Genvolgorde opgeslagen als laatste_clustering
> savePlot(file='plaatje', type='png')
> q()
```



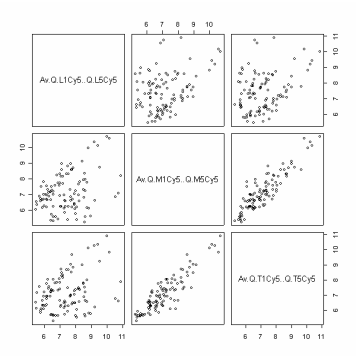
FIGUUR 1



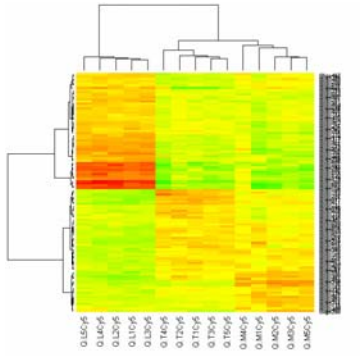
FIGUUR 2



FIGUUR 3



FIGUUR 4



FIGUUR 5

Voorbeeldsessie in R. Ingevoerde commando's zijn vet weergegeven. Figuren die in een ander venster verschijnen zijn als verwijzing geel gekleurd en onderaan weergegeven.

> **source('rivmarray.r')**
Hiermee laad je de RIVMarray-library

> **f.arv.verzamel()**
Deze functie verzamelt alle files die in de directory arv.invoer staan. De functie gaat ervan uit dat dit allemaal ArrayVision files zijn met een layout zoals gebruikelijk is op het RIVM. Dit wil zeggen: in de 1e kolom spotID, in de 2e en 5e kolom spotsignaal zonder achtergrondaftrek voor de twee dyes. Deze waardes worden verzameld in een file ArrayVision_Verzamel.txt, waarbij je zelf nog wel wat moet aanpassen voordat de file opnieuw kan worden ingelezen voor verdere analyse. De kolommen SpotType en GenID moeten nog worden aangepast aan de gridfile van het gebruikte type slide. De eerste regel bevat de naam van de slide én van de scan (dit kan vaak handiger en korter) en de tweede regel zal meestal moeten worden verwijderd (omdat deze een extra header is). Deze bewerkingen worden met opzet niet standaard uitgevoerd om de functie zo algemeen mogelijk te houden als binnen het RIVM reëel is.

> **data1<- f.laaddata('rivmarraydata.txt')**
Hiermee laad je de oefendata. Dit kan natuurlijk ook op andere manieren.

> **data2<- f.normaliseer(data1,3)**
Hiermee normaliseer je de data. Naast de dataset geeft je ook een referentie in die gebruikt wordt voor de normalisatie: 0 = geen referentie (single dye); 1= Ratio; 3 = Cy3 als referentie; 5 = Cy5 als referentie

Voor de normalisatie wordt uitgegaan van een bepaald formaat invoer. Er is één rij (de eerste) met headerinformatie per kolom. De eerste datakolom bevat een unieke waarde voor iedere spot (bijvoorbeeld het spotnummer uit ArrayVision). De tweede kolom bevat een letter die aangeeft om voor type spotje het gaat. De derde kolom geeft aan wat er in ieder spotje zit, zodat replicaspotjes kunnen worden gemiddeld. De verdere kolommen bevatten de echte data. Per slide komt daarbij altijd eerst Cy3 en dan Cy5, tenzij het om single dye-experimenten gaat. Dit formaat is van belang voor de normalisatie, zorg dus dat je data hieraan voldoen! De oefendata kunnen hier als voorbeeld dienen. Verder zijn ontbrekende data niet toegestaan. Volgens de gangbare aanpak treden die ook niet op, maar mochten ze voorkomen dan moet je eerst de 'gaten vullen' via bijvoorbeeld f.impute.knn uit het DNAMR-package (zie aldaar voor uitleg).

De normalisatie bestaat uit een aantal stappen. Als je handig bent, kun je sommige stappen overslaan als ze niet nodig zijn en zo tijd besparen. Doe dit alleen als je ervaren genoeg bent, of laat het aan Jeroen over.

Stap 1: Opschonen arraydata: Hierbij wordt gekeken naar de data in de tweede kolom (het spottype). Met deze kolom kun je informatie over een spotje weergeven, het idee is A = analyse, B = blanco (leeg, spotbuffer), C = controles (landmark-hoekpunten, luciferase, negatieve controles, spike-ins), F = fout gespot, G = Gapdh controles, I = Immunoglobulines e.d., … Alleen spotjes met een letter A (dus meestal gen-oligo's) worden gebruikt in de verdere analyse, de rest wordt weggelaten.

Stap 2: Beschrijving data: Er wordt een aantal waardes genoemd, die meestal voor zich spreken. Een deel hiervan wordt ook uitgevoerd in de file laatste_genormaliseerde_data.txt, zodat je ze nog eens kunt terugkijken. Deze stap dient mede voor de kwaliteitsbeoordeling.
Omdat de normalisatie uitgaat van quantile normalisatie is het van groot belang dat data onderling vergelijkbaar zijn en een soortgelijke verdeling hebben. De term 'sd arrays verschil = 0.4758526' geeft aan hoeveel de maximale en minimale sd van alle scans verschillen. Op basis van ervaring geldt als empirische vuistregel dat deze term niet groter moet zijn dan 0.5. De NPPQC doet iets soortgelijks, deze meet per slide het verschil in sd tussen Cy3 en Cy5 en geeft daarvan de minimale en maximale waarde. Deze waarde geldt telkens binnen eenzelfde slide, dus geeft deze ook informatie over de slidekwaliteit en komt ook kritischer. Hier geldt empirisch als vuistregel dat een verschil tussen de sd van Cy3 en Cy5 van < 0.2 goede slides oplevert, en slides met een sd-verschil > 0.3 a 0.35 af moeten worden gekeurd. Achterliggend idee is dat de verdelingen vergelijkbaar moeten zijn en de correctie in de trend van de MA-plot kleiner moet zijn dan een orde van grootte. Wanneer één van deze waardes groter is dan de genoemde grens wijst dat meestal op een mislukte labeling of een veel te hoge achtergrond, maar meestal op Cy5-afbraak door vocht en/of ozon.

Stap 3: Log-quantile normaliseren: Deze stap vormt de basis van de normalisatie. Het idee komt van Bolstad et al (Bioinformatics, 2003) die dit beschreven voor Affymetrix data. In plaats van per slide de data tijdens het normaliseren te fitten zoals bijv. bij Lowess gebeurt, gebeurt dit nu voor de hele dataset ineens, waarbij in één keer alle scans van alle slides naar elkaar toe worden gefit. Ik heb quantile normalisatie en lowess met elkaar vergeleken voor RIVM-data en de conclusie is dat er hooguit kleine verschillen optreden. Deze verschillen zijn terug te voeren op ex aequo's die ontstaan doordat de ene slide soms wat meer verzadiging heeft dan de andere. Zelfs dan zijn de verschillen klein, en de lijst met hits wordt er niet of nauwelijks door beïnvloed. Deze methode is dus een goed alternatief voor lowess. In de RIVMarray-library is het algoritme enigszins aangepast, zodat het gemiddelde van de logwaardes als basis geldt voor de normalisatie. Dat lijkt wat gevoeliger (geeft meer hits).
Let wel: quantile normalisatie gaat ervan uit dat de algemene dataverdeling tussen slides vergelijkbaar is, voor zowel de monsters als de referentie. Wanneer dit niet geldt door bijvoorbeeld apoptose, metabole stress, heatshock, transcriptieblokkades, overexpressie van een gen, of door een virusinfectie, gaat deze aanname niet op. In dat geval zal het probleem ook met lowess of de meeste andere normalisatie-methodes niet opgelost kunnen worden, aangezien deze er allemaal vanuit gaan dat het globale beeld van de transcriptie gelijk blijft.

Stap 4: Indien van toepassing wordt er genormaliseerd op basis van het gebruikte referentiemonster. Uiteraard is de aanname dat de referentie een dataverdeling heeft die vergelijkbaar is met de monsters, dit wordt in de praktijk het beste gegarandeerd met een referentiepool uit verschillende monsters. De resulterende waarde is niet een gecorrigeerde ratio (lnCy5/Cy3), maar een gecorrigeerde signaalwaarde voor de sample-dye. Op deze manier zijn de waardes ook een maat voor de intensiteit van het spotje. Zo kunnen ze worden gebruikt om te bekijken of een gen wel of niet tot expressie komt bij de voorbereidingen van realtime-PCR.

Stap 5: Middelen replicaspotjes: Statistisch is het mooier (en de ervaring geeft ook duidelijk betere resultaten) als je replicaspotjes per array eerst middelt voor je verder gaat met de analyse. We hebben het hier over daadwerkelijke replicaspotjes, dus niet een 3'- en een 5'-clone of iets dergelijks, maar dezelfde oligo meerdere malen gespot. Deze stap is snelheidsbepalend, vandaar dat eerst wordt gecheckt of er wel sprake is van replica's. Zo niet, dan wordt deze stap overgeslagen.

Nadat het normaliseren klaar is, komt de datum en tijd nogmaals in beeld zodat je kunt zien hoe lang het heeft geduurd en worden de resultaten opgeslagen als laatste_genormaliseerde_data.txt. Deze file bevat alleen de genormaliseerde data zonder extra info-kolommen. Verdere functies maken daar namelijk geen gebruik meer van.
Tevens verschijnt er een PCA van de monsters (Figuur 1)

> **data3<- f.OWA(data2, c(rep(1,5),rep(2,5),rep(3,5)))**
In deze stap voer je een One-Way Anova uit op de data. In dit geval zijn er drie groepen met elk vijf replica's. De uitkomsten bestaan uit een Tabel met het gemiddelde per groep, een F- en p-waarde en de ln van de maximale FoldRatio tussen alle groepen. Ook wordt de q-waarde per gen gegeven, dit is de False Discovery Rate (het gehalte aan vals positieve hits) wanneer de significantedrempel bij de p-waarde van het desbetreffende gen zou worden gekozen.
Deze Tabel wordt opgeslagen als laatste_statistieken.txt. Daarnaast verschijnt er een vulkaanplot (Figuur 2) en wat parameters voor een eventuele poweranalyse en/of biologische interpretatie.
De normale settings geven een F1-statistiek, d.w.z. een genspecifieke p-waarde. Je kunt via de extra parameter Ftype=2 of Ftype=3 een F2- of F3-statistiek berekenen. Een F3-statistiek gaat uit van een gepoolde variantie, d.w.z. de aanname dat alle genen even variabel zijn (deze aanname is overigens onterecht). Een F2-statistiek is een hybride tussen F1 en F3. In de praktijk fungeert F2 als een F1 waarbij de FoldRatio wat zwaarder meetelt.
De functie f.OWA is gebaseerd op de berehandige functie f.F van Amaratunga & Cabrera (DNAMR) met enkele eigen aanvullingen. Let wel: er wordt van uitgegaan dat de replicakolommen per groep al bij elkaar staan!
Een alternatief voor deze functie is de functie f.PMOWA. Daarbij wordt de p-waarde berekend op basis van 1000 permutaties (of een andere waarde naar keuze). Voor de F1-statistiek maakt dit niet veel uit, maar voor de F2 en F3-statistiek geeft dit betrouwbaarder waarden. Let er wel op dat een permutatie-anova nogal traag is, gebruik hem alleen als het meerwaarde heeft.

> **data4<- f.neemPhits(data3, .001, 2)**
Hiermee neem je de relevante hits die (in dit voorbeeld) een p-waarde <0,001 hebben en een maximale FoldRatio >2. Je krijgt tevens te zien wat het aantal significante hits is en het aantal vals-positieven. Als output wordt de lijst uitgevoerd naar laatste_hits.txt, tevens wordt in de vulkaanplot met kleur aangegeven welk deel van de genen alleen significant zijn (blauw) en welke ook aan de FoldRatio voldoen (rood).

> **pairs(data4[,1:3])**
Geeft een scatterplot van alle datakolommen tegen elkaar. In dit geval de drie eerste kolommen van data4, die bevatten de gemiddelde genexpressies per orgaan. Een alternatieve functie is f.pairs, daarbij wordt ook een regressielijn getekend.

> **data5<-data2[rownames(data4),]**
> f.opslaan(data5, 'clusterdata.txt')
> **f.tweedee(data5)**
Eerst selecteer je de genen uit data2 die een hit geven (zo heb je namelijk de waardes van alle samples). Daarna sla je ze op (als tab-delimited text). Tot slot maak je een snelle 2-D clustering met een heatmap. De clustering gebeurt op basis van euclidian distance, met ward linkage voor rijen en average linkage voor kolommen. Deze instellingen werken in de praktijk vaak het beste. De data worden genormaliseerd op het gemiddelde per gen. De schaal loopt van groen (laag) naar rood (hoog), waarbij de extreme waardes het eindpunt van de schaal bepalen. Mochten deze echter te weinig verschillen dan geldt een ondergrens van een ln-factor 2 om opgeblazen ruis te voorkomen. De volgorde van de genen wordt opgeslagen omwille van het leesgemak.
In plaats van zowel de rijen (genen) als kolommen (samples) te clusteren kun je ook alleen de genen clusteren zoals in dit voorbeeld: f.tweedee(data5, kol=FALSE). Voor een eerste grafische cluster-indruk voldoet deze functie ruimschoots, in GeneMaths kun je de rest mooier en handiger uitwerken.

> **savePlot(file='plaatje', type='png')**

Plaatje opslaan (in dit geval als plaatje.png). Andere formaten zijn ook mogelijk.

> **q()**
Afsluiten

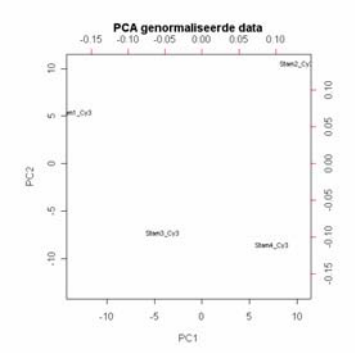
Hiermee heb je de belangrijkste basis wel te pakken. Uiteraard hoef je niet iedere keer met alle data de hele procedure te doorlopen. Het is bijvoorbeeld mogelijk lowess-genormaliseerde ratio's (uit Splus) te importeren en indien van toepassing de replicaspotjes te poolen. Houd daarbij rekening met het juiste formaat voor invoer. Ook kunnen al eerder genormaliseerde data worden ingevoerd voor extra statistische berekeningen of grafische weergave. In deze laatste gevallen zijn extra infokolommen zoals eerder genoemd niet meer nodig.

CGH-arrays

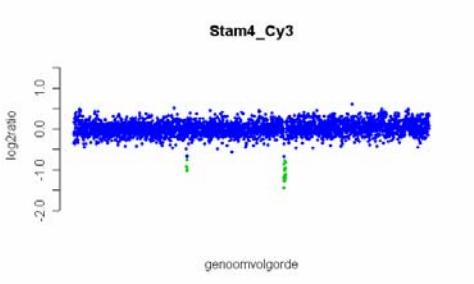
Het meeste werk vindt aan expressie-arrays plaats, maar op het LTR wordt ook gewerkt aan prokaryote CGH-arrays. Daarvoor is ook een aantal functies geschreven. Deze zijn vooral bedoeld voor dit type werk en dus niet voor expressie-arrays of zoogdier-CGH.
Hieronder is een voorbeeldsessie hiervan uitgewerkt.

Voorbeeldsessie 2

```
> source('rivmarray.r')
DNAMR geladen
RIVMarray geladen
Type uitleg() voor meer hulp
Aan de slag dus maar!

> data1<- f.laaddata('kinkhoest.txt')
> data2<- f.CGHnormaliseer(data1)
Begin normalisatie om  Wed Sep 21 11:32:47 2005
Stap 1: Opschonen arraydata: gebeurd
Stap 2: Beschrijving data
aantal spotjes = 13872 , aantal dataspotjes = 11160 , aantal niet-dataspotjes = 2712
aantal labelscans = 6
signaalwaardes minimum = 79.5 , maximum = 65535
sd arrays minimum = 0.8072647 , maximum = 1.057387 , verschil = 0.2501218
NPP-QC minimum = 0.007529059 , maximum = 0.07726262
Resultaten opgeslagen als laatste_QC_Tabel.txt
Stap 3: Log2-quantile normaliseren: gebeurd
QNlog2-waardes minimum = 6.902998 , maximum = 15.99998
Stap 4: Ratio's genomen. Slides zijn
Stam1_Cy3 Stam2_Cy3 Stam3_Cy3 Stam4_Cy3
Stap 5: poolen replicaspotjes (mediaan). Nieuwe aantal =  3582
Resultaten opgeslagen als laatste_genormaliseerde_data.txt
Stap 6: berekenen SD replicaspotjes
Resultaten opgeslagen als laatste_SD_waardes.txt
Normaliseren klaar
PCA van de monsters staat in andere venster                    FIGUUR 6
Klaar om  Wed Sep 21 11:33:15 2005
> data3<- f.CGHhits(data2, -0.7)
Selectie hits: 3582 entries, waarvan 359 geselecteerd
Resultaten opgeslagen als laatste_hits.txt
> f.CGHgenoomplaatje(data2)
CGH genoomplaatjes klaar. Resultaten opgeslagen als CGHplaatje_*.png.
Pas het plot-venster aan voor een ander formaat plaatjes.      FIGUUR 7
> q()
```



FIGUUR 6



FIGUUR 7

> **data1<- f.laaddata('kinkhoest.txt')**
Data laden (in dit geval kinkhoestdata). Hierbij wordt uitgegaan van hetzelfde formaat als voor expressiedata. De eerste datakolom bevat een unieke waarde voor iedere spot (bijvoorbeeld het spotnummer). De tweede kolom bevat een letter die aangeeft om wat voor type spotje het gaat. De derde kolom geeft aan wat er in ieder spotje zit, zodat replicaspotjes kunnen worden gemiddeld. De verdere kolommen bevatten de echte data.

> **data2<- f.CGHnormaliseer(data1)**
De normalisatie vindt plaats vergelijkbaar aan expressiedata, maar met drie kleine verschillen. Ten eerste wordt alles omgeschaald naar een log2-schaal, zoals bij CGH gebruikelijk is. Verder wordt de ratio genomen van de eerste datakolom per slide (het monster) ten opzichte van de tweede datakolom per slide (de controle). Tot slot wordt van replicaspotjes de mediaan en de SD berekend. De mediaan dient voor verdere analyse, de SD is een extra kwaliteitsstap om te zien hoe reproduceerbaar de spotjes zijn. De resultaten worden automatisch opgeslagen.

> **data3<- f.CGHhits(data2, 0.7)**
Hiermee wordt een selectie gemaakt van die genen die in minstens één van de slides een log2-ratio scoren van kleiner dan (in dit voorbeeld) -0,7 of groter dan 0,7. Dit zijn genen waar deleties/amplificaties plaatsvinden en die daarom interessant zijn voor verdere analyse, bijvoorbeeld in GeneMaths. Het resultaat wordt ook weer automatisch opgeslagen.

> **f.CGHgenoomplaatje(data2)**
Geeft een serie plaatjes waarbij de log2-ratio wordt uitgezet tegen de volgorde van de gennaam. Deze volgorde is in principe hetzelfde als de volgorde die uit de normalisatie-stap komt.
Omdat genen zijn genummerd zal dit grofweg overeenkomen met de genoompositie, maar aangezien niet elk gen een oligo heeft en er ook extra oligo's op de array staan, klopt dit niet helemaal. Daarom staan er vooralsnog geen getallen op de x-as.
Elke oligo wordt met een kleurtje weergegeven: blauw als standaard, rood voor amplificatie, groen voor deletie/divergentie. De grens voor deze kleur ligt standaard op 0,7 aan beide kanten, maar dit kun je wijzigen in bijv. 0,8 of een andere waarde, met f.CGHgenoomplaatje(data2, 0.8).
Er wordt een serie plaatjes gemaakt die allemaal dezelfde schaal hebben voor een makkelijke vergelijking. Deze plaatjes worden automatisch opgeslagen in png-formaat. Mocht je de plaatjes groter of kleiner willen, dan kun je het grafische venster van grootte veranderen en de procedure herhalen.

Overzicht alle beschikbare functies:

f.normaliseer(gegevens0, referentiedye)	normalisatie (zie boven) met als subfuncties
f.opschonen(madata)	verwijderen blanco's en controles
f.schrijfQC(madata, referentiedye)	kwaliteitscontrole

f.qnl(gegevens)	quantile normalisatie
f.wisseldye(invoer)	Cy3 en Cy5 kolommen wisselen
f.refnorm(invoer)	referentienormalisatie uitgaande van Cy3
f.rationorm(invoer)	rationormalisatie, dus Cy3/Cy5
f.poolreplica(madata, genenlijst)	poolen replica's
f.OWA(x,g, Ftype=1)	one-way anova (zie boven)
f.PMOWA(x,g, Nperm=1000)	one-way anova met permutaties (zie boven)
f.neemPhits(fdata, Pcut, FRcut)	selectie relevante hits (zie boven)
f.neemFDRhits(fdata, FDRcut, FRcut)	selectie hits op False Discovery Rate en FR
f.tweedee(madata, kol=TRUE)	2-D clustering met heatmap
f.CGHnormaliseer(gegevens0)	prokaryote CGH-normalisatie (zie boven)
f.CGHhits(CGHdata, cutoff)	selectie hits voor prokaryote CGH
f.CGHgenoomplaatje(CGHdata, cutoff=0.7)	prokaryote CGH genoomview
f.laaddata(bestand)	laden data (zie boven)
f.opslaan(gegevens, bestand)	opslaan data
f.kolomsorteer(invoer)	sorteren kolommen op naam
f.arv.verzamel()	Verzamelen ArrayVision-data (zie boven)
f.arv.splits(bestand)	Opsplitsen verzamelde ArrayVision-data
f.analyseer.pubmatrix(pubmx, optie=1)	Textmining analyse op PubMatrix-output. Opties: 0=alleen normaliseren, 1=PCA, 2=clustering, 3=2D-heatmap. Zie ook http://pubmatrix.grc.nia.nih.gov/.
uitleg()	langere uitleg

De functie uitleg() dient tevens als spiekbriefje, want de volgende R-functies komen wel eens van pas maar zijn moeilijk uit het blote hoofd te onthouden.

PCA plaatje van de monsters:
biplot(f.pca(array1), var.axes=FALSE, scale=0, cex=0.7, ylabs=rep(' ',nrow(array1)) , main='PCA data')

Clustering op de data:
plot(hclust(dist(t(array1)), 'ward'), main='titel')

Bekende problemen

Een bekend probleem is het volgende: bij het laden van data kan het zijn dat de data niet door R herkend wordt als arraydata (melding: requires numeric matrix/vector arguments). Dit los je als volgt op:

> **data1<- f.toarray(data1)**
Deze stap is vrijwel altijd noodzakelijk na importeren, behalve als je wilt gaan normaliseren, want in dat geval raak je juist de spotannotatie kwijt! Dat valt wel weer te corrigeren, maar dat is omslachtig, dus begin er niet aan.

Daarnaast is er een aantal voor de hand liggende problemen die je tegen kunt komen.
- Een normalisatie kan niet op minder dan 1 slide worden uitgevoerd
- Een slide moet minimaal één A-spotje bevatten
- Als er een deling door de SD nodig is en de SD is toevallig 0 dan levert dat een fout op

Dit zal echter niet zo vaak problemen geven, want zowel de oorzaak als mogelijke oplossingen liggen voor de hand.

Bijlage 5: Handleiding GeneMaths

Deze handleiding dient als een eerste introductie in het gebruik van GeneMaths versie 2.01. Deze introductie is bewust eenvoudig gehouden want enerzijds is voor dit programma een uitgebreide handleiding beschikbaar en anderzijds vereist het gebruik ervan inzicht in de onderliggende software. Dat laatste kan niet via een korte introductie worden bereikt. Er word daarom uitgegaan van een doorsnee analyse binnen gangbare RIVM-experimenten.

Importeren data

- Begin met je lijstje met hits. Maak in Excel (of in R) een bestand met aan de linkerkant een aantal kolommen met de identifiers en zaken als gennaam of-symbool. Sla het bestand op als tab delimited text (.txt). Doe dit onder Engelse settings, dus met punten in plaats van komma's!
- Start GeneMaths op. Kies onder Scripts --> read_table. Kies het bestand dat je wilt importeren en open het. Stel vervolgens in het menu het volgende in. De maximum en minimumwaarde (-100 tot + 100 werkt vaak goed, een te klein bereik levert fouten op, een te groot bereik werkt minder nauwkeurig), number of gene description fields (het aantal kolommen met gen-informatie), number of array description fields (meestal 1). Use log values is meestal niet nodig omdat de waardes al op een logschaal staan. Nogmaals: denk eraan dat je computer op Engelse instellingen moet staan vanwege het gebruik van punten in plaats van komma's. Klik om OK om te importeren.
- Klik op de menubalk op het icoontje Standardise rows & columns. Klik op row average onder rows. Er staat nu ingesteld (onder rows): perform first, subtract row average, divide by value 1.00, (onder columns) subtract value 0.00, divide by value 1.00. Klik op OK.
  Let op: Als je de data standaardiseert let dan op dat je geen relevante effecten verliest. Als een verschil in het gemiddelde of de spreiding (bijv. bij een tijdreeks) relevant is wil je het opsporen en niet verwijderen. Dit geldt zowel voor standaardisatie als verderop voor de instellingen van een PCA.
- Stel de kleurschaal in (icoontje) op Green/Red I met een schaal van bijvoorbeeld -3 tot 3.
- Sla het bestand op als .cpr bestand. De volgende keer hoef je alleen dit bestand maar te openen om verder te gaan.
- Clustering:
- Wanneer je al een GeneMaths-bestand hebt, kun je deze stap en de vorige overslaan.

Clustering

- Een nuttige eerste stap is even oriënterend kijken naar de complete dataset met behulp van PCA of hiërarchische clustering. Hiermee kom je soms al wat probleemgevallen op het spoor. Typische vragen voor zo'n eerste check zijn:
  - Clusteren replica's samen?
  - Hoe verhouden vergelijkbare monsters zich, bijvoorbeeld ten opzichte van andere groepen? Is er een trend terug te vinden (dosis-respons, tijdreeks)?
  - Zijn er eventuele arrays die als uitbijters beschouwd kunnen worden?
- Voor hierarchische clustering op genen: klik op cluster analysis (rows). Stel dit in op Euclidian, Ward en klik op OK. Dit is de meest gebruikte aanpak, afhankelijk van je proef kan het nodig zijn een andere instelling te kiezen. Zoom in of uit met de vergrootglas-icoontjes. Wanneer je dit wilt kun je clusters samenvatten met je rechtermuisknop (abridge branch).
- Voor hierarchische clustering op monsters: klik op cluster analysis (columns). Stel dit in op Pearson, UPGMA en klik op OK. Dit is de meest gebruikte aanpak, afhankelijk van je proef kan het nodig zijn een andere instelling te kiezen.
- Via print preview kun je plaatjes krijgen die visueel goed genoeg zijn voor PowerPoint.

Principal Component Analysis

- Voor een PCA waarbij je wilt kijken naar de genen: klik op PCA, kies voor de standaard instellingen (subtract average columns, use quantitative values). En klik op OK.
- Voor een PCA waarbij je wilt kijken naar de monsters: kies via het icoontje 'Flip data matrix' voor het draaien van de datamatrix. Klik op PCA, kies voor de standaard instellingen (subtract average columns, use quantitative values). En klik op OK.
- Voor GeneMatsh XT worden de meest gebruikte instellingen PCA op arrays, subtract average arrays.
- Via het copy-icoontje (of alt-PrintScreen) kun je de PCA kopiëren naar PowerPoint.

GeneMaths wordt gebruikt om data en effecten te visualiseren en onder te verdelen in logische groepen. Verder zijn nog een aantal andere programma's in gebruik.

- DAVID/EASE (http://david.abcc.ncifcrf.gov/): voor functionele annotatie en analyse. Aan genen kan de functie gekoppeld worden, verder kan worden gekeken of bepaalde functies vaker voorkomen in een groepje genen (je lijstje hits of een bepaald cluster) dan op de totale array. Dit geeft aanwijzingen of een specifieke pathway of functie gereguleerd wordt. Hiervoor is een korte handleiding beschikbaar.
- GenMapp, voor visualisatie van effecten op pathways. Dit programma is goed in visuele weergave van effecten als je als weet in welke richting je wilt kijken. Het is echter minder geschikt als je (nog) niet weet bij welke pathways er regulatie plaatsvindt.
- PubMatrix (http://pubmatrix.grc.nia.nih.gov/) voor textmining, om te kijken of zoektermen (bijv. namen van genen) samenhangen ten opzichte van andere termen (bijv. organen, effecten, ziektes, stoffen). Deze methode dient vooral om snel in kaart te brengen waar al veel literatuur over is verschenen en of daar een bepaalde samenhang in zit. De interpretatie van de literatuurgegevens blijft wel een kwestie van inzicht en ervaring met het onderwerp.
- Of en hoe deze software gebruikt wordt, is sterk afhankelijk van de opzet van een experiment en de resultaten. Algemene regels voor het gebruik zijn er (nog) niet, al was het maar omdat dit veld nog sterk in ontwikkeling is.

Bijlage 6: Handleiding DAVID/EASE

Deze handleiding dient voor het gebruik van lokaal geïnstalleerde DAVID/EASE-software. Je kunt ook gebruik maken van de website (http://david.abcc.ncifcrf.gov/). Zowel het programma als de website zijn vriendelijk in gebruik en bevatten een uitgebreide helpfunctie met meer informatie.

Algemeen

- Start de lokaal geïnstalleerde versie van EASE op.
- Kies onder '1. Select output' hoe de resultaten moeten worden weergegeven (via Internet Explorer en/of als opgeslagen bestand).
- Selecteer in Excel (of zo) de identifiers van de lijst genen die je wilt analyseren. Plaats de cursur in het veld onder 'Input genes' en klik op Paste.
- Kies onder '2. Input genes' wat voor identifier je gebruikt (bijvoorbeeld Affymetrix probesets, Genbank accessions, LocusLink numbers).

Voor annotatie:

- Klik onder '3. Explore' op 'Select annotation fields' en in het nieuwe venster op 'Add fields'.
- Maak een keuze uit het menu, en voeg ze toe. Gangbare opties zijn: gene name, gene symbol, official gene symbol, alias symbols, chromosomal location, summary, en onder Class de termen GO biological process, GO cellular component, en GO molecular function. Sluit af met 'Done'.
- Klik op 'Annotate genes' en wacht op de resultaten.

Voor pathway-verrijkings analyse:

- Klik onder '3. Explore' op 'Statistical options'. De standaard keuzes zijn onder primary statistic: EASE score, en onder multiplicity correction Bonferroni en Benjamini. Klik op OK om te bevestigen.
- Klik op 'Select categorical systems' en vervolgens 'Add systems'. Gangbare opties zijn: chromosome, GO biological process, GO cellular component, en GO molecular function. Sluit af met 'Done'.
- Klik op 'Find over-represented Gene Categories' en plak (via paste) de lijst met all identifiers van de array in het invoerveld. Meestal zul je al een bestand hiervan hebben, dan kun je dat via 'from a population file' invoeren.
- Klik op 'Run basic analysis' en wacht op de resultaten.
- Bij de uiteindelijke resultaten zijn gangbare criteria dat als False Discovery rate (onder Benjamini) lager is dan 0,1 een pathway significant verrijkt is. Een waarde tussen de 0,1 en 0,2 is niet signicant, maar kan wel op een verrijking wijzen.

Bijlage 7: Handleiding NOAGGG

Deze handleiding dient voor het gebruik van NOAGGG (Numerical Overlap Analysis of Generic Gene Groups ofwel Numerieke Overlap Analyse van Generieke Groepen Genen). Hiermee kun je bekijken of een (sub-)lijstje genen overlap vertoont met andere lijstjes genen. Het principe erachter is dat een lijstje genen dat bij een eerder experiment of in de literatuur is beschreven op dezelfde manier kan worden behandeld als een pathway.

- Open het bestand noaggg.xls. Vanwege vertrouwelijkheid van preliminary en ongepubliceerde data is deze tool niet via intranet te downloaden. Vraag Jeroen als je er gebruik van wilt maken.
- Kopieer vanuit een andere softwaretoepassing de lijst met gen-identifiers en plak deze in kolom A (gele kolom, onder genID). Pas het aantal velden in kolom B en C (groene kolommen) aan aan die van kolom A.
- Kies in het bovenste blauwe vakje, naast annotatie, het type array/identifier dat je hebt. Zie hiervoor de toelichting eronder. Mocht je alleen een lijst met gensymbolen hebben,kies dan voor waarde 1.
- Stel drempelwaarden in voor het minimale aantal genen dat overeen moet komen (standaard 1) en de maximale p-score die je accepTabel acht.
- Aan de rechterkant van het scherm verschijnt nu een overzicht welke experimenten of lijstjes overeenkomen met de genen en criteria die je hebt ingevoerd. Hierbij worden de volgende zaken weergegeven: het aantal genen overlap, de p-score, en een korte omschrijving van het experiment.
- Let op: met deze toepassing kun je alleen vaststellen of er overeenkomst is tussen lijstjes genen. Waar deze overeenkomst op is gebaseerd moet je verder zelf uitzoeken, bijvoorbeeld door te kijken naar welke genen er overeenkomen en hier een nieuwe (pathway-) analyse op uit te voeren. Of de overeenkomst biologisch relevant is, is een kwestie die niet via deze applicatie kan worden beantwoord.
- Ter info: de p-score wordt analoog berekend aan de EASE-score en is enigszins conservatief. Correctie voor meervoudig testen wordt niet toegepast, omdat enerzijds alle lijsten al zijn geselecteerd op relevantie voor RIVM-onderzoek, en anderzijds een mogelijke bias moeilijk kwantitatief valt uit te drukken. De scores moeten dan ook als indicatie worden gezien, niet als harde waarheid.