EINDRAPPORT

GENEXPRESSIEPROFIELEN ALS BIOMERKER IN HUMANE BIOMONITORING:

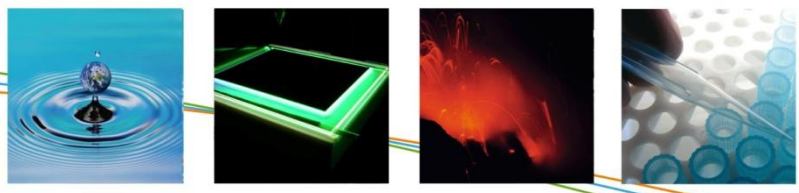
studie van tijdsgebonden variatie in genexpressie en relaties met blootstelling en effecten van milieuvervuiling

(Bestek LNE//OL200700049/8031/M&G)

Coordinator: Prof. G. Schoeters (VITO)
Medewerkers : Dr. P. De Boever
Drs B. Wens
Dr. D. Valkenborg
Universiteit Gent:. Prof. Nik van Larebeke
Drs. Sam De Coster,
Universiteit Maastricht: Prof. Jos Kleinjans
Dr. Danitsja van Leeuwen,
Dr.Danyel Jennen,
Dr. Joost van Delft,
Provinciaal Instituut voor Hygiëne van Antwerpen: Dr. Vera Nelen

Studie uitgevoerd in opdracht van:
Vlaamse overheid, Departement Leefmilieu Natuur en Energie
Afdeling LUCHT, HINDER, RISICOBEHEER, MILIEU EN GEZONDHEID

Februari 2011



VITO NV
Boeretang 200 – 2400 MOL – BELGIE
Tel. + 32 14 33 55 11 – Fax + 32 14 33 55 99
vito@vito.be – www.vito.be

BTW BE-0244.195.916 RPR (Turnhout)
Bank 435-4508191-02 KBC (Brussel)
BE32 4354 5081 9102 (IBAN) KREDBEBB (BIC)

INHOUDSTAFEL

WP3: Exposure-effect relations

WP4: Conclusies en aanbevelingen

MANAGEMENTSSAMENVATTING

In het kader van een preventief milieu- en gezondheidsbeleid is er een sterke behoefte aan de ontwikkeling van nieuwe biomerkers van blootstelling en/of effect van milieuvervuilende stoffen, die reeds in een vroegtijdig stadium een indicatie geven van mogelijke gezondheidseffecten. Het meten van genexpressie in perifeer bloed met behulp van microarray is een aantrekkelijk concept. De onderliggende hypothese is dat circulerend bloed de fysiologische respons weerspiegelt van een organisme en dat deze bloedcellen, en meer specifiek lymfocyten, hun transcriptoom (genexpressieprofiel) aanpassen in functie van de gezondheidstoestand van de gastheer. Het monitoren van genexpressieprofielen wordt bijgevolg aanzien als een veelbelovende aanpak voor de identificatie van gevoelige biomerkers binnen een humaan biomonitoringprogramma. Deze studie beoogt een bijdrage te leveren tot de identificatie van gevoelige merkers van schadelijke effecten die in een vroegtijdig stadium reeds een indicatie geven van gezondheidseffecten. In een eerste werkpakket (WP1) werd een literatuurstudie gemaakt van de mogelijkheden en moeilijkheden bij het gebruik van genexpressie als biomarker in de humane milieu-biomonitoring, en de mogelijke voordelen ten opzichte van traditionele biomerkers. Een belangrijke vereiste om genexpressie te gebruiken in biomonitoring is dat de normale transcriptionele variabiliteit gedocumenteerd is. Hierbij is het belangrijk om de variatie te karakteriseren op populatieniveau, maar ook tijdsgebonden variatie op individueel niveau. Enkel indien dergelijke achtergrondwaarden bekend zijn, kan bij genexpressieprofielen bekomen na blootstelling aan polluenten nagegaan worden of er een significante verhoging of verlaging is van genexpressie. In deze studie werd de stabiliteit van genexpressie in bloed in functie van tijd werd onderzocht op basis van gegevens van een studiepopulatie van gezonde volwassenen. Korte en lange termijn variabiliteit in genexpressie werd beschreven en geanalyseerd in werkpakket 2. De nadruk werd gelegd op seizoensinvloed en invloed van geslacht. In een derde onderzoeksluik (WP3) werd de impact van polluentblootstelling op genexpressie onderzocht. De beleidsbruikbaarheid van genexpressieprofielen in het kader van een humaan biomonitoringmeetnet wordt besproken in WP4.

De studie gaf aan dat in een mileu gezondheidscontext het meten van genexpressie veelbelovend is maar nog niet zo ver ontwikkeld als voor klinische toepassingen. De technologie blijkt robuust te zijn maar er is een voortdurende evolutie op he vlak van dataverwerking en interpretatie. De korte termijn en seizoensvariabiliteit van individuele genen werd gedocumenteerd in een onderzoek bij gezonde jonge volwassenen. Blootstellings- effectrelaties werden geObserveerd waarbij de respons vaak geslachtsspecifiek bleek. De technologie is inzetbaar in een biomonitoringscontext als " early warning signaal " voor complexe blootstelling, maar de interpretatie van de betekenis voor de gezondheid vraagt verdere opvolging.

Literatuurstudie over het gebruik van genexpressieanalyse in een milieu gezondheidscontext

In WP1 werd via een uitgebreide literatuuroverzicht besproken wat de mogelijkheden en moeilijkheden zijn van het gebruik van genexpressie als biomarker in de humane milieu-biomonitoring, en de mogelijke voordelen ten opzichte van traditionele biomerkers. Het idee om genexpressie als biomerker te gebruiken in (humane) biomonitoring stamt voort uit de observatie dat (bepaalde) (milieu)blootstellingen en biologische effecten gepaard gaan met veranderingen in genexpressieniveaus. Ook hebben genexpressiemetingen hun nut reeds bewezen inzake bij de identificatie van werkingsmechanismen van toxische stoffen, bij het onderzoeken (dier)soortspecifieke verschillen, bij onderzoek naar extrapolatie naar lage blootstellingsdosissen (door een hogere gevoeligheid van genexpressie), bij het bepalen van specifieke 'fingerprints' waarbij genexpressiepatronen toelaten het onderscheid maken tussen verschillende blootstellingen of biologische effecten, en in het onderzoek naar multipele (gecombineerde) blootstellingen. Er werd een overzicht gegeven van de verschillende technologieën die worden gebruikt in transcriptomics (genexpressieanalyse), alsook in proteomics (eiwitanalyse) en metabolomics (metabolietanalyse), met inbegrip van hun specifieke kenmerken en beperkingen. Ook de opties voor data-analyse werden besproken. Daarbij werd ingegaan op 'pathway-analyses', die ervoor zorgen dat niet enkel de individuele genen beschouwd worden, maar ook netwerken van genen die samen biologische eenheden vormen. Dit verhoogt de betrouwbaarheid van de resultaten en leunt dichter aan bij de biologische processen die zich afspelen.

De systeembiologische aanpak, die zowel transcriptomics, proteomics als metabolomics includeert, verhoogt verder het potentieel voor biomerkerontwikkeling, niettegenstaande deze aanpak nog zeer duur is en verschillende moeilijkheden kent. We bespreken ook onze keuzes binnen het studie-opzet: de analyse van genexpressie in perifeer bloed door middel van microarrays. Verder werd er een inventaris gemaakt van peer-gereviewde studies die zich richten op de analyse van de genexpressie van het volledige genoom in de context van milieu-, beroepsmatige- en medische blootstellingen, met als doel de ontwikkeling van potentiële biomerkers. Ook initiatieven van internationale milieu- en gezondheidsorganisaties rond het mogelijk gebruik van genexpressiedata voor regulerende doeleinden werden geinventariseerd. Hierbij worden verschillende factoren besproken die een invloed kunnen hebben op de uitkomst van genexpressiemetingen bij het gebuik als biomerker. Daarbij werd duidelijk dat verschillende factoren die de genexpressie (kunnen) beïnvloeden nog amper onderzocht zijn. Voorbeelden zijn de invloed van persoonlijke kenmerken (leeftijd, geslacht, inter- en intra-individuele variatie), levensstijl (voeding, roken,…) en staalname (tijdstip op de dag, seizoen, …). Deze factoren moeten in rekening gebracht worden, en vereisen nog bijkomend onderzoek. Ook geven we een aantal aandachtspunten voor het gebruik van transcriptomics gegevens voor regulerende doeleinden, aangezien verschillende factoren de mogelijkheden om genexpressiedata te gebruiken voor beleidsdoeleinden kunnen bemoeilijken. Hiertoe behoren onder andere de nood tot interdisciplinaire samenwerking en de publieke beschikbaarheid van data (gegevensoverdracht). Verschillende initiatieven zijn reeds opgestart om aan deze noden tegemoet te treden.

Variabiliteit in genexpressie in gezonde jonge volwassenen:

In werkpakket 2 wees een gedetailleerde analyse van PubMed literatuur uit dat onderzoek naar genexpressie in bloed bij de algemene bevolking en tijdsgebonden variatie in de genexpressie metingen eerder beperkt is. Een paar kleine studies (met een studiepopulatie van ongeveer 10 individuen) zijn beschikbaar. De vergelijking tussen deze studies is moeilijk omdat verschillende procedures voor staalname werden gehanteerd. De onderzoekers gebruikten verschillende microarray platforms en er werd vastgesteld dat studies gebruik gemaakt werd van eerste generatie microarrays met een verouderde probe design.

In het kader van verschillende Vlaamse Mileu- en Gezondheidsstudies wordt er aandacht besteed aan de introductie van genexpressie in biomonitoringscampagnes. Met de inschakeling van deze nieuwe meetmethodes hopen de onderzoekers te beschikken over een gevoelige en specifieke methode om een blootstelling aan een complex mengsel van milieupolluenten op te sporen. Verder kunnen de genexpressiemetingen biologisch geïnterpreteerd worden en een verband leggen met de effecten van de blootstelling. Om een idee te krijgen van de impact van milieupolluenten op genexpressie in bloed dient de variabiliteit van genexpressie gekarakteriseerd te worden. Hierbij wordt verondersteld dat er speciale aandacht gegeven moet worden aan de twee parameters: geslacht (mannen versus vrouwen) en tijd (tijdsafhankelijke variatie zoals bijv. een seizoenseffect).

In het kader van de LNE-genexpressie studie werd in WP2 een studie opgezet met 22 gezonde vrijwilligers (evenveel mannen als vrouwen). De leeftijd van de vrijwilligers was tussen 20 en 40 jaar en de enige exclusiecriteria waren roken en eventuele zwangerschap. Deze leeftijdscategorie werd gekozen omdat die enerzijds naar voor geschoven is als belangrijke doelgroep voor biomonitoring in Vlaanderen. Anderzijds werd deze leeftijdscategorie geselecteerd omdat het nemen van bloedstalen bij deze groep vlot kan verlopen. Zes bloedstalen werden gecollecteerd per persoon. Drie stalen werden gecollecteerd in November 2009 en drie stalen in Mei 2010.
De drie stalen werden genomen met een interval van één week. Drie stalen binnen een maand lieten toe om statistische maten zoals gemiddelde expressie, en standaardafwijking te berekenen voor elk gen. Lange-termijn variabiliteit en seizoenseffect konden ingeschat worden door stalen te collecteren over twee verschillende periodes.

Een totaal aantal van 132 bloedstalen werden gecollecteerd met behulp van Tempus tubes (Applied Biosystems). Deze stalen werden verwerkt met Tempus spin kits (Applied Biosystems) en voor elk staal werd totaal RNA van hoge kwaliteit bekomen dat geschikt was voor microarray analyse. Het totaal RNA werd vervolgens gezuiverd van het globine mRNA (niet-informatief mRNA dat een groot gedeelte van het totaal RNA uitmaakt). Een verwijdering van globine mRNA kan helpen om de detectie van genen met zwakke signaalintensiteit te verbeteren. Het effect van de globine verwijdering werd niet getest in dit project. Verschillende rapporten hebben het nut van een globine verwijdering bevestigd voor verschillende microarray platformen (Affymetrix, Agilent, en Illumina).

Vervolgens werd een één-kleur labelling (Cy3) protocol toegepast om absolute signaalintensiteit te meten voor elk gen via microarray technologie. Bovendien laat een één-kleur strategie een maximale statistische flexibiliteit toe. De standaardprotocols van Agilent Technologies werden

gevolgd om de stalen voor te bereiden voor microarray analyse. De stalen werden gehybridiseerd tegenover Agilent 4×44K humane microarrays. Een set van 132 ruwe microarray data werd bekomen. Een eerste kwaliteitscontrole die voorzien is binnen Feature Extraction software van Agilent Technologies gaf aan dat de technische kwaliteit van de data goed is. De technische reproduceerbaarheid was hoog en de variatiecoëfficiënt was lager dan 5%.

Er bestaan verschillende procedures voor het behandelen van ruwe microarray data. Een strikte Agilent kwaliteitsfiltering die gebruikt wordt door de Universiteit Maastricht gaf aan dat ongeveer 8 000 informatieve probes (op een totaal van 41 000 probes) niet bruikbaar waren voor verdere data-analyse omdat de signaalintensiteit te laag was. Deze probes werden eveneens geïdentificeerd via een alternatieve methode binnen werkpakket 2. Er wordt voorgesteld om probes van een lage kwaliteit vooraf te filteren. Deze a priori filtering heeft een positieve impact op latere procedures die gevolgd worden voor correctie voor multipliciteit. In de huidige studie betekende dit dat er voor ongeveer 20% minder genen diende gecorrigeerd te worden.

Vervolgens werden Spearman correlatiecoëfficiënten berekend voor de totale microarray profielen. Zeer hoge Spearman correlaties van >0.94 werden bekomen. Dit wijst enerzijds op kwaliteitsvolle data. Anderzijds is dit ook een indicatie dat verschillen tussen microarray profielen beperkt zijn en dat het effect van de beschouwde parameters (geslacht en tijd (seizoen)) subtiel kunnen zijn. De resultaten gaven ook aan dat de profielen bekomen per persoon sterk vergelijkbaar waren. Dit is indicatief voor het feit dat de expressies in bloed voor een bepaald persoon vrij stabiel zijn. Er kon duidelijk een verschil waargenomen worden tussen de stalen afkomstig van verschillende personen m.a.w. de variatie tussen individuen is groter dan de variatie binnen een individu.

Beschrijvende statistiek werd toegepast op de twee voornaamste factoren (geslacht en seizoen) die bestudeerd werden in werkpakket 2. Er werd vastgesteld dat er een aanzienlijke spreiding is in de signaalintensiteit en de standaarddeviatie. De grootte van de standaarddeviatie wordt bovendien beïnvloed door de signaalintensiteit. Dit aspect kan belangrijk zijn voor een power berekening.
Indien men wil identificeren hoeveel individuen men nodig heeft om een specifiek effect te identificeren in een case-control studie (bijv. hoog-blootgestelde individuen versus laag-blootgestelde individuen), dan zal de relatie tussen standaarddeviatie en signaalintensiteit voor elk gen in rekening moeten gebracht worden.

De variatiecoëfficient varieerde sterk, maar 75% van de probes hadden een variatiecoëfficiënt kleiner dan 0.25 in beide seizoenen. Een kleine hoeveelheid probes (tussen 40 en 60) had een grote variatiecoëfficent in één seizoen en een lage coëfficiënt in het tweede seizoen. De biologische functie van deze genen en hun betrokkenheid in biologische processen en pathways werd opgelijst. Op dit moment is het onduidelijk waarom deze genen zo'n specifiek seizoenseffect vertonen.

De variatiecoëfficiënt werd geëxtraheerd voor twee specifieke genlijsten: i) een panel van huishoudgenen, en ii) een biomerker panel dat voorheen werd geïdentificeerd door de Universiteit Maastricht en de Universiteit Gent. Genen uit deze twee panels vertoonden een groot spreiding in hun variatiecoëfficiënt wanneer de data werden gegroepeerd volgens de factoren geslacht en seizoen. Bovendien was de variatiecoëfficiënt sterk afhankelijk van de probe die werd gebruikt om

het transcript te identificeren. Er wordt aanbevolen om voor kandidaat biomerker genen de signaalintensiteit te controleren, alsook de variatiecoëfficiënt voor één of meerdere probes die gebruikt worden om het transcript te identificeren.

Vervolgens werd binnen WP2 een statistisch model ontwikkeld om de impact van geslacht en seizoen op genexpressie in bloed te evalueren. Een mixed model analyse per gen identificeerde geen belangrijke korte-termijn effecten binnen een seizoen. Voor een totaal aantal van 97 genen werd een trend waargenomen in seizoen 1 (herfst), maar niet in seizoen 2 (lente). Biologische interpretatie van de data identificeerde geen gecoördineerde biologische processen of pathways die werden beïnvloed door de herfst. De waargenomen trend in de herfst kan belangrijk zijn indien precies één van die genen zou gebruikt worden in toekomstige biomonitoringsprogramma's. De trend dient echter verder bevestigd te worden in een grotere populatie alvorens definitieve conclusies kunnen getrokken worden. Omwille van de beperkte trend die werd waargenomen binnen een seizoen, werden de metingen binnen een seizoen beschouwd als herhaalde metingen. Hierdoor kon verder getest worden naar het seizoenseffect op genexpressie. Door gebruik te maken van het concept van herhaalde metingen werd de power van de studie verhoogd.

Er werd een random-effecten model gebruikt om de effecten van de twee factoren seizoen en geslacht te bestuderen. Dit model werd gefit voor elke probe afzonderlijk. Geen enkele interactieterm werd waargenomen en deze parameter werd bijgevolg uit het model verwijderd. Na correctie voor multipliciteit (Benjamini-Hochberg correctie, $p<0.05$) bleven er respectievelijk 110 en 1995 probes over waarvan de signaalintensiteit beïnvloed wordt door geslacht en seizoen. Deze twee probelijsten werden onderworpen aan een pathway analyse. Tal van biologische termen en biologische netwerken bleken significant aangerijkt te zijn voor deze twee lijsten na een Gene Ontology analyse. Het was echter niet duidelijk eenduidig of coherente biologische thema's werden teruggevonden. Met behulp van het commerciële pakket Ingenuity Pathway Analyse werden tal van functies en termen geïdentificeerd die verbonden zijn met infectie en immuniteit. De exacte biologische betekenis is onduidelijk. Pathway analyse is een complex onderzoeksveld omdat er verschillende algoritmes, software en databases zijn. Een gedetailleerde biologische interpretatie en manuele curatie van de bekomen genexpressie data viel buiten het bestek van deze studie.

Een uitgebreide studie werd gemaakt van de variantie via een analyse van de variantie-covariantie matrix. Via verschillende likelihood testen werd nagegaan welke varianties beïnvloed werden door seizoen en/of geslacht. De variantie van ongeveer 8 000 probes was afhankelijk van seizoen of geslacht. Een interindividueel effect op de variantie werd waargenomen voor 25 probes en de functie van de corresponderende genen werd beschreven. Er werd niet verder onderzocht wat de biologische reden was waarom de variantie van deze genen werd beïnvloed.

Tot slot werd, bij wijze van voorbeeld, kandidaat merker genen (20 voor mannen en 20 voor vrouwen) die eerder geïdentificeerd waren door de Universiteit Maastricht en de Universiteit Gent vergeleken met de genlijsten die gegenereerd werden via het random effecten model *i.e.* genlijst voor de factor geslacht en één voor de factor seizoen. Respectievelijk vijf en twee genen van de kandidaat merker genen voor mannen en vrouwen vertoonden een seizoenseffect. De precieze betekenis van deze observatie dient verder onderzocht te worden. Het is mogelijk dat het moment

van staalname nl. periode van het jaar een confounding effect heeft op de signaalintensiteit van deze genen. Bijgevolg dient de impact hiervan op de totale fingerprint nagegaan te worden.

Werkpakket 2 genereerde waardevolle informatie betreffende de whole-genome genexpressie in bloed en de variabiliteit van genexpressie. Dit gebeurde rekening houdend met een mogelijke tijdsgebonden fluctuatie van genexpressie in een kleine populatie van vrijwilligers met een leeftijd tussen 20 en 40 jaar. In andere genexpressie studies wordt er geprobeerd om genexpressie fingerprints te identificeren die gecorreleerd zijn met blootstelling aan complexe mengsels van polluenten (bijv. werkpakket 3 van deze studie). Dergelijke studies trachten genen te identificeren die differentieel tot expressie komen tussen twee groepen nl. mensen met een hoge blootstelling en mensen met een lage blootstelling. Dergelijke studies hebben meestal één meting per individu en zijn niet in staat om variatie binnen het individu (of tijdsgebonden variatie) in te schatten. Het kan zijn dat differentieel tot expressie gebrachte genen met een kleine effect-grootte uiteindelijk vals-positieven blijken te zijn omdat de tijdsgebonden variabiliteit een groter effect heeft dan de behandeling en/of blootstelling. Er wordt ingeschat dat een systematische vergelijking van resultaten van genexpressie metingen met de informatie rond variabiliteit van genexpressie kan helpen om het aantal vals-positieven tot een minimum te beperken. Een van de eerste initiatieven in deze context is om de ruwe en verwerkte data van deze genexpressie studie te stockeren in een database die eenvoudig kan geraadpleegd worden. Verder dienen er criteria ontwikkeld te worden om informatie met betrekking tot variabiliteit van genexpressie te relateren aan data die bekomen worden in het kader van blootstellingsanalyses. Een volgende stap kan bestaan uit het samenvoegen van deze verschillende data door experten in de biostatistiek.

Relatie tussen genexpressie en blootselling en effect van milieupolluenten:

Tijdens het vorige Steunpunt Milieu & Gezondheid is aangetoond dat in het kader van biomonitoring studies onder blootgestelde populaties genexpressie onderzoek tot relevante resultaten aanleiding geeft. Destijds is de expressie van een achttal met kanker verband houdende genen onderzocht in 400 personen uit de Vlaamse populatie van “ouderen”, te weten CYP1B1, SOD2, ATF4, MAPK14, CXCL1, PINK1, DGAT2 en TIGD3. Dit is uitgevoerd met behulp van kwantitatieve PCR, en daarbij de relatie onderzocht met enige biomerkers voor blootstelling en effect.
Het bereikte resultaat was positief, en leverde daarmee een onderbouwing van het nut van genexpressie analyse in bevolkingsstudies. Tegelijkertijd deed het noodzakelijkerwijs beperkt aantal genen die bestudeerd waren, de vraag oproepen, of er belangrijke effecten op genexpressie niveau wellicht nog gemist waren. Daarom wordt in het onderhavige onderzoek bij een subpopulatie (n=100) uit de eerder in het kader van het Steunpunt onderzochte personen (n = 400 voor genexpressie analyse gebaseerd op RT-PCR) de globale genexpressie profielen (alle actieve genen) gemeten door middel van DNA microarray technologie.

De algemene doelstelling van werkpakket (WP3) is de relatie te onderzoeken tussen genexpressieprofielen en biomerkers van blootstelling en effect aan milieuvervuilende stoffen bij blootstellingniveaus die voor de algemene bevolking in Vlaanderen relevant zijn. Hiervoor wordt een beroep gedaan op de gegevens voor blootstellings- en effectmerkers die reeds in het kader van het 1^{e} generatie Steunpunt Milieu & Gezondheid verworven waren. Stof- en dosis-afhankelijke genexpressie veranderingen zijn daarbij geanalyseerd op het niveau van de individuele genen alsmede, teneinde meer zicht te krijgen op de biologische relevantie van deze modificaties, op het niveau van genetische netwerken. Ook zijn de associaties onderzocht met belangrijke effect merkers, met name tumor merkers, micronuclei frequenties en COMET signalen, eveneens verkregen in het kader van het 1^{e} generatie Steunpunt Milieu & Gezondheid. Tenslotte is nagegaan hoe de overlap is tussen de nu verkregen microarray data en de eerder verkregen RT PCR data van het 1^{e} generatie Steunpunt Milieu & Gezondheid.

De onderstaande tabel geeft een overzicht per leeftijdsklasse van het aantal mannen en vrouwen (allen niet-rokers) van wie hoogwaardig RNA, zoals verzameld en geïsoleerd in het kader van de eerdere studie in het Steunpunt, nu geanalyseerd is met behulp van micoarray technologie. Mannen en vrouwen zijn geselecteerd op basis van hun eerder gemeten blootstellingniveaus van meerdere polluenten. Het betreft met name de blootstelling aan: cadmium, lood, PCBs (138+153+180), dioxines (CALUX), PAKs (OH-pyreen) en benzeen (ttMA). De integrale blootstelling is beoordeeld aan de hand van de som van de z-scores van elke polluent. Op basis van deze aanpak worden een groep met hoge vs. een groep met lage blootstelling (gemiddelde van polluentgehaltes) verkregen.

Tabel 1 : Aantal mannen en vrouwen per leeftijdsgroep voor de groep met hoge vs. De groep met lage blootstelling (volgens de z-scores)

		Mannen		Vrouwen		
		H	L	H	L	
Leeftijd	50-55	7	8	8	8	**31**
	55-60	8	8	8	8	**32**
	60-65	9	9	9	10	**37**
		24	**25**	**25**	**26**	

De volgende tabellen geven een overzicht van deze blootstellingmerkers bij geselecteerde mannen en vrouwen (volgende pagina). Via een t-test werd getest of de individuele polluentgehaltes in de hoog blootgestelde groep significant hoger zijn dan de gehaltes in de laag blootgestelde groep (analyse voor mannen en vrouwen afzonderlijk)

Tabel 2: Populatiekenmerken van hoog en laag blootgestelde individuen voor beide geslachten.

Mannen		Hoog blootgestelden			Laag blootgestelden			
		n	Gem.	Std. afw.	n	Gem.	Std. afw.	p-waarde t-test
Cd urine	µg/g crt	24	0.67	0.57	25	0.48	0.23	0.1425
Pb bloed	µg/L	24	56.48	32.42	25	34.38	11.59	0.0024
merker PCBs	ng/g vet	24	474.97	204.18	25	326.74	65.31	0.0012
PCB118	ng/g vet	24	27.21	14.65	25	29.94	12.55	0.4871
p,p'-DDE	ng/g vet	24	588.83	359.97	25	618.89	837.75	0.8720
Hexachlorobenzeen	ng/g vet	24	49.79	22.05	25	46.26	16.86	0.5305
TEQ	pg/g vet	23	37.04	28.61	22	12.91	9.36	0.0005
ttMA (~Benzeen)	mg/g crt	21	0.16	0.16	23	0.08	0.07	0.0447
OH-pyr (~PAHs)	µg/g crt	24	0.40	0.61	25	0.13	0.10	0.0276

Vrouwen		Hoog blootgestelden			Laag blootgestelden			
		n	Gem.	Std. afw.	n	Gem.	Std. afw.	p-value t-test
Cd urine	µg/g crt	25	1.07	0.47	26	0.51	0.16	<0.0001
Pb bloed	µg/L	25	54.63	21.00	26	28.36	13.99	<0.0001
merker PCBs	ng/g vet	25	399.40	156.31	26	324.37	69.67	0.0304
PCB118	ng/g vet	25	34.23	14.76	26	32.04	12.42	0.5676
p,p'-DDE	ng/g vet	25	661.71	537.03	26	740.41	531.07	0.6011
Hexachlorobenzeen	ng/g vet	25	33.97	22.62	25	11.89	9.81	<0.0001
TEQ	pg/g vet	25	76.60	37.21	26	91.69	39.56	0.1673
ttMA (~Benzeen)	mg/g crt	24	0.23	0.19	22	0.08	0.07	0.0006
OH-pyr (~PAHs)	µg/g crt	25	0.40	0.52	26	0.15	0.20	0.0240

De onderstaande tabel geeft een overzicht van het totaal aantal gemodificeerde expressies van genen bij hoog vs laag blootgestelde mannen en vrouwen ('fractie' is het aantal differentiële genen t.o.v. het totale aantal genen op de array).

Tabel 3: Aantal genen met differentiële expressie in t-test tussen hoog en laag blootgestelde individuen, voor de drie datasets.

	UP	DOWN	TOTAAL	fractie
Mannen	320	50	370	1.3%
Vrouwen	902	2016	2918	10.2%
Alle	621	576	1197	4.2%

Het grote verschil in kwantitatieve zin (aantallen genen die differentieel tot expressie komen) tussen mannen en vrouwen wordt meteen duidelijk. Indien wij kiezen voor een striktere evaluatie waarbij alleen effecten in ogenschouw worden genomen die boven het niveau van toevalligheid (5 %) liggen, dan worden alleen statistische significante verschillen tussen hoog- en laagblootgestelde vrouwen aangetroffen. In kwalitatieve zin valt op dat bij vrouwen bij hogere blootstellingniveaus

genexpressies overwegend negatief gereguleerd zijn terwijl bij hoog blootgestelde mannen juist sprake is van genexpressie dereguleringen in de positieve richting.
De volgende tabel toont de resultaten van de correlatieanalyses tussen genexpressie en blootstelling- en effectmerkers, in mannen (n=49), vrouwen (n=51), en voor mannen en vrouwen samen (n=100). Voorts zijn correlaties in negatieve en in positieve richting uitgesplitst.

Tabel 5: Aantallen genen die significant correleren met blootstelling- en effectmerkers.

		Mannen				Vrouwen				Alle				
	Men	n	UP	DOWN	TOTAL	fractie	n	UP	DOWN	TOTAL	fractie	n	UP	DOWN	TOTAL	fractie
Exposures	t-test HvsL	49	320	50	370	1.3%	51	902	2016	2918	10.2%	100	621	576	1197	4.2%
	Cd	49	5053	9735	14788	51.6%	51	958	834	1792	6.3%	100	3399	5733	9132	31.9%
	Pb	49	137	356	493	1.7%	51	831	865	1696	5.9%	100	501	516	1017	3.6%
	PCBs	49	216	77	293	1.0%	51	775	1217	1992	7.0%	100	199	179	378	1.3%
	PCB118	49	118	280	398	1.4%	51	635	1115	1750	6.1%	100	198	778	976	3.4%
	DDE	49	442	230	672	2.3%	51	163	295	458	1.6%	100	187	140	327	1.1%
	HCB	49	954	1170	2124	7.4%	51	821	707	1528	5.3%	100	374	476	850	3.0%
	TEQ	45	710	247	957	3.3%	50	429	670	1099	3.8%	95	998	434	1432	5.0%
	TTMA	44	2778	1027	3805	13.3%	46	317	1355	1672	5.8%	90	880	1086	1966	6.9%
	HPYR	49	605	76	681	2.4%	51	258	551	809	2.8%	100	1150	271	1421	5.0%
Effects	PSA	49	583	476	1059	3.7%	-	-	-	-	-	-	-	-	-	-
	CEA	38	1529	681	2210	7.7%	38	886	532	1418	5.0%	76	1194	419	1613	5.6%
	p53 (SL)	42	261	676	937	3.3%	43	601	733	1334	4.7%	85	507	883	1390	4.9%
	micronuclei	36	241	371	612	2.1%	38	320	321	641	2.2%	74	291	561	852	3.0%
	Comet assay (mediaan)	31	318	125	443	1.5%	36	633	441	1074	3.8%	67				
	HDG	49	3054	5023	8077	28.2%	51	383	178	561	2.0%	100	1024	2001	3025	10.6%

Opnieuw vallen de grote verschillen tussen mannen en vrouwen op, met name wanneer wij opnieuw de striktere evaluatie toepassen waarbij alleen correlaties in ogenschouw worden genomen die boven het niveau van toevalligheid (5 %) liggen. Mannen reageren in vergelijk tot vrouwen zeer sterk op cadmium blootstelling, en ook sterker op blootstelling aan benzeen (c.q. de benzeen-metaboliet t,t'-muconzuur; TTMA). Vrouwen reageren sterker op blootstellingen aan lood, en PCBs. Voorts zien we bij mannen een opvallend hoge correlatie met de urinaire excretie van 8-hydroxy-guanine (HDG), een merker voor het ontstaan en repareren van oxydatieve DNA schade.

Er zijn echter ook overeenkomsten tussen mannen en vrouwen in de mate waarop hun expressieniveaus van bepaalde genen correleren met blootstellingniveaus van bepaalde polluenten en met effect merkers. De onderstaande tabel maakt dit zichtbaar:

Tabel 6: Aantallen genen die zowel bij mannen als vrouwen significant correleren.

	Gelijk teken correlatiecoëfficiënt		Tegengesteld teken correlatiecoëfficiënt		p-waarde binom. Verd
Cd	303	46%	357	**54%**	0.00341
Pb	6	18%	28	**82%**	0.00008
SUMPCB	5	12%	38	**88%**	<0.000001
PCB118	18	62%	11	38%	0.06444
DDE	3	27%	8	73%	0.08057
HCB	36	26%	102	**74%**	<0.000001
TEQ	24	**62%**	15	38%	0.04573
TTMA	80	28%	204	**72%**	<0.000001
hPYR	38	**90%**	4	10%	<0.000001
CEA	53	**54%**	46	46%	0.06259
p53	29	**85%**	5	15%	0.00002
micronuclei	8	**100%**	0	0%	0.00391
Comet assay	6	**60%**	4	40%	0.20508
HDG	31	39%	48	**61%**	0.01450

Met name voor lood, PCBs, HCB (en in mindere mate cadmium en HDG) is de correlatie tussen blootstelling en genexpressie bij beide geslachten significant tegengesteld (teken van de correlatiecoëfficiënt), terwijl voor deze voor PAKs, p53 en micronuclei (en in mindere mate voor TEQ) eerder gelijk zijn aan elkaar (gelijk teken voor de correlatiecoëfficiënt).

De volgende sectie beschrijft de resultaten van de analyse van betrokken genetische netwerken zoals uitgevoerd met behulp van de standaard “pathway finding tool” MetaCore. De onderstaande tabel geeft de aantallen van gemodificeerde netwerken bij mannen en vrouwen weer, in relatie tot de blootstellingniveaus van de gemeten polluenten en effectmerkers:

Tabel 7 Resultaat MetaCore pathway analyse van gen expressie veranderingen correleren met blootstelling – en effect merkers

	# pathways		
parameter	**mannen**	**vrouwen**	**overlap**
cd	27	46	0
hpyr	12	9	0
pb	18	10	0
pcbs	15	4	0
teq	20	7	0
ttma	60	60	7
dde	31	12	0
hcb	22	17	4
pcb118	14	15	0
micronuclei	18	17	0
cea	28	12	0
hdg	30	13	1
comet assay	15	18	0
p53	23	34	3
psa (alleen mannen)	8		
ttest	16	22	2
cleartest	6	8	0

Gele arceringen geven aan dat hier netwerken voorkomen die bij mannen en vrouwen gemeenschappelijk tot expressie komen.

De onderstaande tabel toont dan de genetische netwerken die statistisch significant met name bij hoog blootgestelde mannen voorkomen:

Tabel 8: Gemodificeerde pathways bij hoog blootgestelde mannen

#	Name	pValue	Network objects
1	Development_Regulation of CDK5 in CNS	2.816e-3	3/16
2	Development_MAG-dependent inhibition of neurite outgrowth	3.998e-3	3/18
3	Regulation of metabolism_Bile acids regulation of glucose and lipid metabolism via FXR	6.265e-3	3/21
4	Development_ERK5 in cell proliferation and neuronal survival	7.160e-3	3/22
5	Apoptosis and survival_Role of CDK5 in neuronal death and survival	7.160e-3	3/22
6	Transport_RAB3 regulation pathway	1.109e-2	2/9
7	Development_Neurotrophin family signaling	1.147e-2	3/26
8	Regulation of lipid metabolism_FXR-dependent negative-feedback regulation of bile acids concentration	1.370e-2	2/10
9	Transport_FXR-regulated cholesterol and bile acids cellular transport	1.654e-2	2/11
10	G-protein signaling_RhoB regulation pathway	1.962e-2	2/12
11	Oxidative stress_Angiotensin II-induced production of ROS	2.291e-2	2/13
12	Apoptosis and survival_NGF signaling pathway	2.291e-2	2/13
13	Nitrogen metabolism	2.642e-2	2/14
14	Nitrogen metabolism/ Rodent version	2.642e-2	2/14
15	NGF activation of NF-kB	3.403e-2	2/16
16	Ascorbate metabolism	3.657e-2	1/2

De onderstaande tabel toont vervolgens de genetische netwerken die statistisch significant met name bij hoog blootgestelde vrouwen voorkomen:

Tabel 9: Gemodificeerde pathways bij hoog blootgestelde vrouwen

#	Name	pValue	Network objects
1	N-Glycan biosynthesis p2	3.149e-3	7/18
2	Apoptosis and survival_Regulation of Apoptosis by Mitochondrial Proteins	3.777e-3	9/28
3	Riboflavin metabolism	5.147e-3	4/7
4	Development_Growth hormone signaling via STATs and PLC/IP3	1.069e-2	8/27
5	Apoptosis and survival_Role of CDK5 in neuronal death and survival	1.107e-2	7/22
6	Apoptosis and survival_Granzyme B signaling	1.346e-2	8/28
7	Neurophysiological process_Dopamine D2 receptor signaling in CNS	1.525e-2	4/9
8	Apoptosis and survival_FAS signaling cascades	1.593e-2	10/40
9	Apoptosis and survival_Caspase cascade	1.672e-2	8/29
10	Chemotaxis_Leukocyte chemotaxis	1.892e-2	10/41
11	Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim	2.486e-2	8/31
12	Immune response_CD16 signaling in NK cells	2.584e-2	9/37
13	Signal transduction_AKT signaling	2.810e-2	7/26
14	G-protein signaling_Rac3 regulation pathway	3.291e-2	4/11
15	Transport_FXR-regulated cholesterol and bile acids cellular transport	3.291e-2	4/11
16	Cytoskeleton remodeling_CDC42 in cellular processes	3.333e-2	5/16
17	Apoptosis and survival_Role of IAP-proteins in apoptosis	3.419e-2	7/27
18	Development_Notch Signaling Pathway	3.544e-2	8/33
19	Apoptosis and survival_Endoplasmic reticulum stress response pathway	4.128e-2	9/40
20	Muscle contraction_nNOS Signaling in Skeletal Muscle	4.486e-2	4/12
21	Retinol metabolism	4.729e-2	6/23
22	Apoptosis and survival_Ceramides signaling pathway	4.885e-2	7/29

De in de twee bovenstaande geel gearceerde netwerken worden zowel bij hoog blootgestelde mannen als bij hoog blootgestelde vrouwen gedereguleerd.

Vanwege het verschil tussen mannen en vrouwen in de genomische respons op blootstelling aan met name hormoonverstorende polluenten, is in de MetaCore pathway analyse specifiek gekeken naar hormoon-gerelateerde pathways. Bij de mannen is “Estrogen biosynthesis” gevonden in relatie tot Cd blootstelling. Bij de vrouwen is “Estradiol metabolism” gevonden in relatie tot Cd blootstelling en tot het voorkomen van micronuclei, alsmede “Estrone metabolism” in relatie tot blootstelling aan pcb118.

Tabel 11 toont vervolgens de genetische netwerken die statistisch significant gevonden zijn bij mannen voor de effectmerker hdg (urinaire excretie van 8-OH-guanine):

#	Name	pValue	Network objects
1	Neurophysiological process_Melatonin signaling	2.290e-3	9/11
2	Immune response_CCR5 signaling in macrophages and T lymphocytes	2.639e-3	21/35
3	Transcription_P53 signaling pathway	4.936e-3	19/32
4	Apoptosis and survival_Beta-2 adrenergic receptor anti-apoptotic action	5.351e-3	8/10
5	Delta508-CFTR traffic / Sorting endosome formation in CF	7.025e-3	12/18
6	Transcription_Transcription regulation of aminoacid metabolism	1.227e-2	13/21
7	CFTR folding and maturation (norm and CF)	1.351e-2	8/11
8	Immune response_CXCR4 signaling via second messenger	1.354e-2	11/17
9	Normal wtCFTR traffic / Sorting endosome formation	1.398e-2	9/13
10	Immune response_IFN gamma signaling pathway	1.754e-2	19/35
11	Development_Growth hormone signaling via PI3K/AKT and MAPK cascades	1.754e-2	19/35
12	Propionate metabolism p.2	2.207e-2	12/20
13	Signal transduction_Calcium signaling	2.207e-2	12/20
14	N-Glycan biosynthesis p2	2.378e-2	11/18
15	Transcription_Transcription factor Tubby signaling pathways	2.395e-2	5/6
16	Muscle contraction_ GPCRs in the regulation of smooth muscle tone	2.618e-2	15/27
17	PDGF activation of prostacyclin synthesis	2.739e-2	6/8
18	G-protein signaling_RhoB regulation pathway	2.800e-2	8/12
19	Transcription_CREM signaling in testis	2.800e-2	8/12
20	Translation_Opioid receptors in regulation of translation	2.800e-2	8/12

21	Neurophysiological process _Visual perception	2.800e-2	8/12
22	Development_Ligand-dependent activation of the ESR1/AP-1 pathway	2.836e-2	7/10
23	Transcription_Ligand-dependent activation of the ESR1/SP pathway	3.514e-2	12/21
24	Development_PACAP signaling in neural cells	3.866e-2	11/19
25	Cell cycle_Role of 14-3-3 proteins in cell cycle regulation	4.243e-2	10/17
26	UMP biosynthesis	4.496e-2	3/3
27	Transport_RAN regulation pathway	4.640e-2	9/15
28	DNA damage_Role of Brca1 and Brca2 in DNA repair	4.744e-2	13/24
29	Development_Hedgehog signaling	4.819e-2	16/31
30	Cell cycle_The metaphase checkpoint	4.819e-2	16/31

Tabel 12 toont vervolgens de genetische netwerken die statistisch significant gevonden zijn bij vrouwen voor de effectmerker hdg (urinaire excretie van 8-OH-guanine)::

#	**Name**	**pValue**	**Network objects**
1	Keratan sulfate metabolism p.1	2.417e-3	3/11
2	Signal transduction_Activin A signaling regulation	4.131e-3	4/26
3	Keratan sulfate metabolism p.2	6.183e-3	3/15
4	Immune response_MIF - the neuroendocrine-macrophage connector	6.183e-3	3/15
5	N-Glycan biosynthesis p2	1.048e-2	3/18
6	IMP biosynthesis	2.143e-2	2/9
7	HIV-1 signaling via CCR5 in macrophages and T lymphocytes	2.324e-2	3/24
8	Regulation of lipid metabolism_G-alpha(q) regulation of lipid metabolism	2.634e-2	2/10
9	Immune response_Histamine signaling in dendritic cells	2.877e-2	3/26
10	G-protein signaling_Rac3 regulation pathway	3.165e-2	2/11
11	Development_EGFR signaling via PIP3	3.735e-2	2/12
12	wtCFTR and delta508 traffic / Clathrin coated vesicles formation (norm and CF)	4.341e-2	2/13
13	Cytoskeleton remodeling_Thyroliberin in cytoskeleton remodeling	4.980e-2	2/14

De in de twee bovenstaande tabellen geel gearceerde netwerken worden zowel voor de effectmerker hdg bij mannen als bij vrouwen gedereguleerd.

Tabel 13 toont de genetische netwerken die statistisch significant gevonden zijn bij mannen voor de effectmerker p53:

#	Name	pValue	Network objects
1	Cytoskeleton remodeling_Role of Activin A in cytoskeleton remodeling	7.396e-4	5/16
2	Cytoskeleton remodeling_ESR1 action on cytoskeleton remodeling and cell migration	1.388e-3	4/11
3	Muscle contraction_EDG5-mediated smooth muscle contraction	7.953e-3	4/17
4	Apoptosis and survival_Ceramides signaling pathway	1.202e-2	5/29
5	Chemotaxis_Inhibitory action of lipoxins on IL-8- and Leukotriene B4-induced neutrophil migration	1.444e-2	4/20
6	Cell adhesion_Integrin-mediated cell adhesion and migration	1.444e-2	4/20
7	Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim	1.589e-2	5/31
8	Cytokine production by Th17 cells in CF (Mouse model)	1.589e-2	5/31
9	Inhibitory action of Lipoxins on neutrophil migration	1.718e-2	4/21
10	DNA damage_Mismatch repair	1.823e-2	3/12
11	Development_EGFR signaling via PIP3	1.823e-2	3/12
12	Immune response_CD40 signaling	2.083e-2	6/45
13	Development_NOTCH-induced EMT	2.286e-2	3/13
14	Development_Thrombospondin-1 signaling	2.286e-2	3/13
15	Cytokine production by Th17 cells in CF	2.356e-2	4/23
16	Immune response_IL-6 signaling pathway	2.722e-2	4/24
17	Muscle contraction_ACM regulation of smooth muscle contraction	2.722e-2	4/24
18	Immune response_MIF in innate immunity response	2.722e-2	4/24
19	Bacterial infections in CF airways	2.900e-2	5/36
20	Blood coagulation_GPCRs in platelet aggregation	3.573e-2	5/38
21	Signal transduction_JNK pathway	3.573e-2	5/38
22	G-protein signaling_EDG5 signaling	4.022e-2	3/16

23	Cell cycle_Influence of Ras and Rho proteins on G1/S Transition	4.334e-2	5/40

Tabel 14 toont de genetische netwerken die statistisch significant gevonden zijn bij vrouwen voor de effectmerker p53:

#	Name	pValue	Network objects
1	HIV-1 signaling via CCR5 in macrophages and T lymphocytes	3.075e-4	7/24
2	Immune response_NF-AT signaling and leukocyte interactions	2.246e-3	5/17
3	Muscle contraction_ GPCRs in the regulation of smooth muscle tone	3.876e-3	6/27
4	Neurophysiological process_Thyroliberin in cell hyperpolarization and excitability	5.322e-3	4/13
5	Transcription_Transcription regulation of aminoacid metabolism	6.101e-3	5/21
6	Neurophysiological process_Long-term depression in cerebellum	7.116e-3	4/14
7	Immune response_IL-3 activation and signaling pathway	7.528e-3	5/22
8	Immune response_PGE2 signaling in immune response	7.528e-3	5/22
9	Immune response_Regulation of T cell function by CTLA-4	1.105e-2	5/24
10	Muscle contraction_EDG5-mediated smooth muscle contraction	1.474e-2	4/17
11	Immune response_HTR2A-induced activation of cPLA2	1.474e-2	4/17
12	Heme metabolism	1.555e-2	5/26
13	Immune response_Histamine signaling in dendritic cells	1.555e-2	5/26
14	Immune response_Signaling pathway mediated by IL-6 and IL-1	1.811e-2	4/18
15	Triacylglycerol metabolism p.1	1.821e-2	5/27
16	Development_Growth hormone signaling via STATs and PLC/IP3	1.821e-2	5/27
17	Blood coagulation_GPCRs in platelet	2.120e-2	6/38

	aggregation		
18	Immune response _CCR3 signaling in eosinophils	2.120e-2	6/38
19	Cytoskeleton remodeling_Role of PKA in cytoskeleton reorganisation	2.191e-2	4/19
20	Cardiac Hypertrophy_Ca(2+)-dependent NF-AT signaling in Cardiac Hypertrophy	2.617e-2	4/20
21	Immune response_IFN alpha/beta signaling pathway	2.617e-2	4/20
22	Development_EDG5 and EDG3 in cell proliferation and differentiation	2.617e-2	4/20
23	Immune response_CD28 signaling	2.791e-2	5/30
24	wtCFTR and delta508-CFTR traffic / Generic schema (norm and CF)	2.791e-2	5/30
25	Neurophysiological process_ACM regulation of nerve impulse	3.089e-2	4/21
26	G-protein signaling_RAC1 in cellular process	3.089e-2	4/21
27	Transcription_PPAR Pathway	3.609e-2	4/22
28	Development_GH-RH signaling	3.638e-2	3/13
29	Blood coagulation_Blood coagulation	4.175e-2	4/23
30	Transport_Clathrin-coated vesicle cycle	4.244e-2	7/56
31	Cytoskeleton remodeling_Alpha-1A adrenergic receptor-dependent inhibition of PI3K	4.361e-2	2/6
32	Neurophysiological process_Glutamate regulation of Dopamine D1A receptor signaling	4.437e-2	3/14
33	Cytoskeleton remodeling_Thyroliberin in cytoskeleton remodeling	4.437e-2	3/14
34	Muscle contraction_ACM regulation of smooth muscle contraction	4.789e-2	4/24

De in de bovenstaande tabellen geel gearceerde netwerken worden zowel voor de effectmerker p53 bij mannen als bij vrouwen gedereguleerd.

Aangezien pathway analyse in het algemeen beperkt is omdat slechts een deel van de gecorreleerde of significante genen gebruikt kunnen worden, d.w.z. alleen die genen die voorkomen in de pathways, is er in MetaCore tevebs een GO processes analyse uitgevoerd. Bij deze analyse zijn alle genen betrokken zijn. De onderstaande tabel geeft de aantallen van de significante GO processen bij mannen en vrouwen weer, in relatie tot de blootstellingniveaus van de gemeten polluenten en effectmerkers:

Tabel 15 Resultaten MetaCore GO processes analyse

	# GO processes					
parameter	**mannen**		**vrouwen**		**overlap**	
	P<0.05	**P<0.05 & FDR<0.1**	**P<0.05**	**P<0.05 & FDR<0.1**	**P<0.05**	**P<0.05 & FDR<0.1**
cd	256	54	264	32	57	10
hpyr	162	0	218	12	6	0
pb	152	0	118	1	3	0
pcbs	181	0	186	9	6	0
teq	301	1	332	5	18	0
ttma	355	17	401	72	45	0
dde	284	0	232	1	2	0
hcb	229	0	254	14	14	0
pcb118	159	0	131	1	4	0
micronuclei	221	14	178	0	3	0
cea	165	0	172	0	4	0
hdg	249	50	288	7	23	0
comet assay	174	0	186	12	3	0
p53	275	54	181	0	23	0
psa (alleen mannen)	205	0				
ttest	203	0	222	0	7	0
cleartest	140	0	107	46	1	0

Gele arcering geeft aan dat hier GO processen voorkomen die bij mannen en vrouwen gemeenschappelijk significant P<0.05 zijn en een FDR <0.1 hebben. Deze processen zijn gerelateerd aan RNA metabole processen.

Een overzicht van de meest voorkomende GO processes voor polluenten en effectmerkers staat in de volgende tabel:

Tabel 16 De meest voorkomende GO processes uit de MetaCore analyse

	# GO processes	
parameter	**mannen**	**vrouwen**
cd	RNA metabolic process transcription cell cycle	chromatin assembly RNA metabolic process transcription
hpyr	regulation of immune system	chromatin assembly
pb	protein metabolic process phosphorylation	protein processing
pcbs	immune response	translation macromolecule biosynthetic process
teq	RNA metabolic process	neurological system process
ttma	RNA metabolic process translation cellular response to stress	signaling
dde	translation cell cycle regulation of immune system proces	negative regulation of translation
hcb	homeostatic process	negative regulation of gene expression
pcb118	regulation of immune system	homeostatic process
micronuclei	chromatin assembly	diverse
cea	transport	regulation of immune system process
hdg	RNA metabolic process transcription	translation protein metabolic process
comet assay	diverse	signaling
p53	macromolecule metabolic process	regulation of protein modification process
psa (alleen mannen)	diverse	
ttest	transport	diverse
cleartest	diverse	neurological nervous process signaling

Vergelijking met de RT PCR analyse van 8 genen (van Leeuwen et al. 2008)

Voor de 8 genen waarvan genexpressie reeds eerder met behulp van RT-PCR gemeten is in een eerdere studie in het kader van het Steunpunt Milieu en Gezondheid (MAPK14, SOD2, CYP1B1, PINK1, TIGD3, CXCL1, DGAT2, ATF4) werden bij alle 100 individuen in de microarray analyse een genexpressie-waarde gemeten boven de detectielimiet.
Bij mannen werd een significante correlatie gevonden tussen de PCR-genexpressie-waarden en de microarray-waarden voor MAPK14 ($p<0.0001$), SOD2 ($p<0.0001$), CXCL1 ($p=0.035$), DGAT2 ($p<0.0001$), CYP1B1 ($p=0.008$) en PINK1 ($p<0.0001$), maar niet voor ATF4 ($p=0.5$) en TIGD3 ($p= 0.22$).
Analoog werd bij vrouwen een significante correlatie gevonden voor MAPK14 ($p=0.039$), SOD2 ($p<0.0001$), CXCL1 ($p=0.034$), DGAT2 ($p<0.0001$), PINK ($p<0.0001$), TIGD3 ($p<0.0001$), maar niet voor ATF4 ($p=0.21$) en CYP1B1 ($p=0.20$).

Bij het 1ste Generatie Steunpunt onderzoek zijn 8 genen geselecteerd ten behoeven van een PCR analyse en deze zijn als biomerker toegepast bij 398 Vlaamse volwassenen. De microarray resultaten van de huidige studie bij een subgroep van 100 Vlaamse volwassenen tonen zowel bij mannen als bij vrouwen bij 6 van de 8 genen een significante correlatie. Dit is een aanwijzing voor de technische betrouwbaarheid van de hier gepresenteerde microarray resultaten. Overigens, de verschillen in resultaten voor de twee overige genen kunnen te wijten zijn aan probe-verschillen tussen beide analyses. Dosis-respons relaties tussen blootstelling aan polluenten en genexpressie veranderingen gemeten met behulp van de microarray technologie laten complexe biologische reacties zien. De complexiteit is niet onverwacht aangezien de blootstellingen ook complexe mengsels van milieucontaminanten betreffen.

Het voornaamste resultaat van de onderhavige studie heeft betrekking op de aangetoonde geslachtsverschillen in de genomische respons op blootstelling aan de onderzochte polluenten. Dit is ook waargenomen voor de effect-merkers. De geslachtsverschillen in de genomische reactiepatronen blijken substantieel te zijn. Bij toekomstige biomonitoring studies dient daarom met geslachtsverschillen expliciet rekening gehouden te worden. Van belang is erop te wijzen dat deze reacties optreden in vertegenwoordigers van de algemene Vlaamse bevolking: we spreken dus niet van extreme blootstellingen.

De meest aangetroffen gedereguleerde netwerken bij mannen m.b.t. polluenten en effectmerkers (aangegeven tussen haakjes, waarbij effectmerkers geel gearceerd zijn) hebben betrekking op

- Transcription_P53 signaling pathway (cadmium, hydroxypyrene[1], pp'-DDE[2], hdg[3])
- Transcription_Transcription regulation of aminoacid metabolism (TEQ,[4] pp'DDE, HCB[5] hdg)
- Cell adhesion_ECM remodeling (TEQ, micronuclei, comet assay)
- Immune response_IL-13 signaling via JAK-STAT (HCB, cea[6])

[1] 1-hydroxypyrene is een metaboliet van pyreen dat behoort tot de polyaromatische koolwaterstoffen(hpyr)
[2] *p,p´*-dichlorodiphenyldichloroethylene
[3] 8-hydroxy-guanine is een merker van DNA herstel
[4] TEQ: 2,3,7,8 dioxine toxiciteits equivalenten
[5] Hexachlorobenzene

- Immune response_PGE2 signaling in immune response (hpyr, PCB118[7])
- Immune response_T cell receptor signaling pathway (ttma[8],HCB, comet assay)
- Muscle contraction_ GPCRs in the regulation of smooth muscle tone (Cd, hdg, comet assay)
- Muscle contraction_EDG5-mediated smooth muscle contraction (Cd, cea, p53)
- PDGF activation of prostacyclin synthesis (Cd, TEQ, hdg)
- Transcription_CREM signaling in testis (PCB118, hdg)

De meest aangetroffen gedereguleerde netwerken bij vrouwen m.b.t. polluenten en effectmerkers (aangegeven tussen haakjes, waarbij effectmerkers geel gearceerd zijn) hebben betrekking op

- N-Glycan biosynthesis p2 (cd, pb, hcb, hdg)
- Development_TGF-beta-induction of EMT via ROS (cd, hpyr, ttma, dde)
- Retinol metabolism (pcbs, teq, pcb118)
- Translation _Regulation of EIF4F activity (cd, ttma, dde, micronuclei)
- Apoptosis and survival_Ceramides signaling pathway (cd, ttma)
- Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim (cd)
- Development_Growth hormone signaling via STATs and PLC/IP3 (p53)
- G-protein signaling_Rac3 regulation pathway (cd, hdg)
- HIV-1 signaling via CCR5 in macrophages and T lymphocytes (ttma, hdg, p53)
- Keratan sulfate metabolism p.2 (cd, dde, hdg)
- Neurophysiological process_ACM regulation of nerve impulse (micronuclei, comet assay, p53)
- Neurophysiological process_Glutamate regulation of Dopamine D1A receptor signaling (dde, pcb118, p53)
- Normal and pathological TGF-beta-mediated regulation of cell proliferation (cd, ttma, dde)

[6] Cea:tumormerker carcino embryonaal antigen
[7] PCB118 : polychlorinated biphenyl 118
[8] Ttma: tt muconic acid is een benzeenmetaboliet

Conclusies en aanbevelingen

In werkpakket 4 (WP4) werden een aantal conclusies van WP2 en WP3 gebundeld. Verder werd een evaluatie gemaakt van de genexpressie technologie in het kader van milieu en gezondheid. Details van deze analyse zijn terug te vinden in WP4, maar volgende algemenen conclusies konden naar voor gebracht worden.

De resultaten van de huidige studie toonden aan dat zowel bij de recent gecollecteerde stalen als bij de eerder gecollecteerde stalen, hoog kwalitatief RNA kan bekomen worden uit perifeer bloed bij biomonitoringsstudies van de algemene bevolking. Betrouwbare en reproduceerbare methodes voor behandeling van bloedstalen en voor de bewaarcondities zijn beschikbaar aan de UM and VITO. De natte procedures zijn vergelijkbaar en de ruwe data zijn robuust en uitwisselbaar voor cross lab vergelijking. Er zijn verschillende goede methodes voor datafiltering en statistische analyse van complexe microarray dataset. De inzichten hieromtrent evolueren en in de internationale wetenschappelijke wereld is er geen consensus betreffende een eenduidige strategie. Verschillende methodes kunnen gebruikt worden op voorwaarde dat ze conceptueel aanvaardbaar zijn en gedragen zijn in de wetenschappelijke wereld. Het proces van data processing is een onderdeel van een volledige pijplijn om microarray te gebruiken voor identificatie van genexpressie biomerkers in de context van milieu en gezondheid. Een relevante biologische conclusie en een validatie van de gevonden biomerkers in een onafhankelijke studiepopulatie zijn de finale criteria voor evalutie van de pijplijn voor data-analyse. In een goed en betrouwbaar scenario zou uiteraard goede convergentie moeten bestaan voor de verschillende methodes.

Microarray data analyse en biologische interpretatie met bioinformatica is een complex veld en wordt vaak onderschat bij het opmaken van een projectvoorstel en bijhorend budget. Voor toekomstige projecten is het belangrijk om voldoende tijd en middelen te voorzien opdat de gegenereerde data maximaal kunnen benut worden. Indien onderzoeksprojecten met meerdere labo's worden opgezet is het aan te bevelen dat bij aanvang van het project duidelijke afspraken gemaakt worden over een aantal procedures (zoals formaat van data files, kwaliteitscriteria voor data, gebruik van data of niveau van probe of gen, procedures voor correctie voor multipliciteit, enz.). Dergelijke afspraken zorgen er voor dat data in verschillende werkpakketten vergelijkbaar zijn en optimaal kunnen uitgewisseld worden.

In de context van humane biomonitoring is pathway analyse nuttig om data te aggregeren op een biologisch niveau en die informatie kan gebruikt worden om hypotheses te genereren naar mogelijke gezondheidsaspecten. De experimentele set-up laat niet toe het dynamisch patroon en de interactie van biologische netwerken en bijgevolg de biologische mechanismes ten volle te omvatten. Verandering is genexpressie hoeft niet noodzakelijk vertaald te worden in een fysiologische ontregeling omdat er andere controle mechanismen die zorgen voor een robuust systeem. Complementaire informatie op niveau van epigenetics, proteomics and/of metabolomics niveau zijn belangrijk om de relevantie van veranderingen in genexpressie te bevatten. Bovendien is vanuit fysiologisch standpunt een fenotypische ankering belangrijk. Veranderingen in genexpressie hebben wel een belangrijke signaalfunctie die verdere opvolging vraagt.

De technologie voor genexpressie (microarray, realt-time PCR, etc.) is op dit moment de meest mature (in vergelijking met bijv. proteomics) om kost-efficiënt zo'n whole-genome signaal te genereren. Een eerste aanduiding bestaat hiervoor aangezien er een goede correlatie tussen genexpressie metingen gedaan in het 1ste Steunpunt Milieu- en Gezondheid en de huidige studie. Er wordt dan ook gesteld dat het gebruik van genexpressie in biomonitoring waardevol is.

De methodes voor genexpressie zijn goed onder controle in zowel VITO als Universiteit Maastricht. De gegenereerde analyses zijn robuust en betrouwbaar. De huidige studie heeft de variabiliteit van genexpressie gedocumenteerd in een gezonde volwassene populatie. Hieruit blijkt dat genexpressie een vrij stabiele parameter is. De korte- en lange-termijn variabiliteit in genexpressie is beschreven. De genen die beïnvloed worden door een seizoen of geslachtsafhankelijk zijn werden beschreven. Bovendien werd een inventaris gemaakt van genen met een sterk variabele expressie. Het inventariseren van deze genen is waardevol omdat deze minder geschikt kunnen zijn om later als biomerker gebruikt te worden. Kandidaat merker genen die via andere manieren worden geïdentificeerd, worden best gecontroleerd op hun mogelijk geslacht- of seizoenseffect, alsook hun variabiliteit van expressie. De gegevens worden best gestockeerd in een database die als nulpunt kan fungeren voor latere onderzoeken bij eenzelfde doelgroep.

Werkpakket 3 observeerde een significante verandering in genexpressie die vergecorreleerd is met blootstelling aan polluenten. Een longitudinale studie kan meer informatie geven over de relevantie van de metingen. Enerzijds is er de optie om de metingen te correleren met potentiële morbiditeit. Anderzijds kunnen additionele metingen informatie geven over mogelijke persistente effecten op niveau van genexpressie. Bevestiging van de genexpressieprofielen over verschillende tijdspunten heen, zou de waarde van genexpressie in biomonitoring verder bevestigen. Een kwantificeerbare link leggen met morbiditeit/mortaliteit is momenteel te vroeg. Volgende strategieën zijn daarvoor aangewezen: 1) opname van genexpressiemetingen in prospectieve cohort studies, 2) verband onderzoeken van de relatie tussen genexpressieprofielen en gevaloriseerde effectmerkers of klinische parameters.

Profielen van veranderingen in genexpressie kunnen ingesloten worden in humane biomonitoring in beleidsvoorbereidende context omdat ze een gevoelige merker zijn die op een unieke manier blootstellingspatronen van mengsels weergeeft met een eerste vertaling naar mogelijk biologische impact. De genexpressie metingen hebben een signaalfunctie die verdere opvolging vraagt. De waarde van dergelijke merkers kan slechts tot uiting komen wanneer genexpressiemetingen systematisch parallel worden uitgevoerd met andere biomerker- en effectmetingen zodat de relatieve gevoeligheid en vroege signaalfunctie kan worden bepaald. We kunnen verwachten, dat naar analogie met de klinische geneeskunde, genexpressie in de milieugezondheidscontext zal kunnen worden toegepast als biomerker.

WP1: Literatuurstudie

SAMENVATTING

In dit literatuuroverzicht bespreken we de mogelijkheden en moeilijkheden van het gebruik van genexpressie als biomarker in de humane milieu-biomonitoring, en de mogelijke voordelen ten opzichte van traditionele biomerkers.

Er wordt een overzicht gegeven van de verschillende technologieën die worden gebruikt in transcriptomics, alsook in proteomics en metabolomics, met inbegrip van hun specifieke kenmerken en beperkingen. Ook de opties voor data analyse worden besproken.

In een volgende deel bespreken we onze keuzes binnen het studie-opzet: de analyse van genexpressie in perifeer bloed door middel van microarrays.

Verder wordt ere en inventaris gemaakt van peer-gereviewde studies die zich richten op de analyse van de genexpressie van het volledige genoom in de context van milieu-, beroepsmatige- en medische blootstellingen, met als doel de ontwikkeling van potentiële biomerkers. Ook initiatieven van internationale milieu- en gezondheidsorganisaties rond het gebruik van genexpressiedata voor (potentieel) regulerende doeleinden.

Als laatste bespreken we verschillende factoren die een invloed kunnen hebben op de uitkomst van genexpressiemetingen bij het gebuik als biomerker, zoals leeftijd, geslacht, voedingsgewoonten en levensstijl. Ook geven we een aantal inachtnemingen voor het gebruik van transcriptomics gegevens voor regulerende doeleinden.

ABSTRACT

In this literature overview we discuss the ***opportunities and challenges of using gene expression as a biomarker*** in human environmental biomonitoring, and potential advantages over the use of traditional biomarkers.

An overview of different ***technologies*** used in transcriptomics, as well as the related proteomics and metabolomics technologies, is given, including their specific characteristics and limitations. Several options for data analysis in transcriptomics are discussed.

In a third part of this study, we explain our ***choice of study settings***: the analysis of gene expression in whole blood using microarrays.

Additionally, an inventory is made of peer-reviewed ***studies that involve whole-genome gene expression analysis in human environmental, occupational and medical biomonitoring***, for identifying potential biomarkers. Also, ***efforts of international environment and health organizations*** on the use of gene expression data for (potential) regulatory purposes are listed.

To conclude, we discuss several possible factors that can influence the outcome when using gene expression as a biomarker, such as age, sex, nutrition and lifestyle; and we give some considerations for the use of the eventual ***use of transcriptomics data for regulatory purposes***.

ABBREVIATIONS AND TERMS

Alternative splicing: a process whereby multiple protein isoforms are generated from a single gene. Certain non-coding parts of the pre-mRNA (introns) are removed during splicing, while the remaining (coding) parts (exons) are joined together. This can be done in several alternative ways, by removal or inclusion of certain exons, thus resulting in different proteins.

ANOVA: Analysis of Variance

Bias: a disparity in research or test results due to using improper assessment tools or instruments across groups.

Biomarker: a substance used as an indicator of a biological state (e.g. an exposure, an effect, a susceptibility)

cDNA: single stranded DNA reverse transcribed from an mRNA template

Co-carcinogenesis: The combination of two or more different factors in the production of cancer.

Confounder: a variable in a statistical model that correlates with both the dependent variable and the independent variable.

Covariate: a variable that is possibly predictive of the outcome under study.

Cross hybridization: The annealing of a single-stranded DNA sequence to a probe (e.g. on a microarray) to which it is only partially complementary.

Cy3/Cy5: fluorescent labeling cyanine 3 (green)/cyanine 5 (red)

DNA methylation: a type of chemical modification (methylgroep op DNA – leidt mogelijk tot mutaties?) of DNA that can be inherited and subsequently removed without changing the original DNA sequence (epigenetic modification)

Dynamic range: ratio between the smallest and largest value (e.g. concentrations in a sample, or the range of concentrations that certain a method can detect)

Epigenetics: changes in phenotype or gene expression of a cell caused by mechanisms other than changes in the DNA sequence.

FDR: False Discovery Rate

GO: Gene Ontology

Histone modifications: modifications in proteins that are complexed with DNA in the nucleus.

-mer (60-mer, 25-mer): indicates the length of the oligos on a microarray (number of bases)

Metabolomics: large scale measurement of metabolites

Metabonomics the quantitative measurement of the dynamic response of metabolites to stimuli

mRNA (messenger RNA): RNA, transcribed from DNA, carrying coding information for protein synthesis

Northern blotting: An electroblotting method in which RNA is transferred to a filter and detected by hybridisation to (32)P labelled RNA or DNA.

Pathway: a series of gene or chemical relations in a cell

Phosphorylation: addition of a phosphate group on a molecule. This is an important process for activating or deactivating proteins

Posttranslational modifications: modification of a protein after its translation (e.g. phosphorylation)

Post-transcriptional modification: modification of primary transcript RNA to mature mRNA. This includes mainly processing and splicing.

probe: fragment of DNA or RNA of variable length (usually 100-1000 bases long) spotted on (e.g.) a microarray, which is used to detect in DNA or RNA samples the presence of nucleotide sequences (the DNA target) that are complementary to the sequence in the probe

Promotor: A region of DNA to which RNA polymerase binds before initiating the transcription of DNA into RNA.

Protein isoform: different forms of a protein that may be produced from different genes, or from the same gene by alternative splicing.

Proteomics: large scale measurement of proteins

Realtime-PCR: laboratory technique used to amplify and simultaneously quantify a targeted DNA molecule. Enables both detection and quantification of a specific sequence in a DNA sample.

SAM: Significance analysis of Microarrays, a statistical technique for finding significant genes

in a set of microarray experiments based on the use of repeated permutations of the data to determine if the expression of any genes are significantly related to the response.

TFBS: Transcription Factor Binding Site

Throughput capacity: measure for the number of samples that can be processed in a given time

Transcriptomics: genome wide measurement of mRNA expression

Translation: the process of protein formation by decoding mRNA

INTRODUCTION

As a First part of the study "GENEXPRESSIEPROFIELEN ALS BIOMERKER IN HUMANE BIOMONITORING: STUDIE VAN TIJDSGEBONDEN VARIATIE IN GENEXPRESSIE EN RELATIES MET BLOOTSTELLING EN EFFECTEN VAN MILIEUVERVUILING", a literature framework will be drawn about the use of gene expression analysis in an environment and health context and - more specifically – about the opportunities and challenges related to the development of new biomarkers of exposure or effects of pollutants.

This study has evolved from another initiative from the Flemish government. From 2002-2006 the 1st Flemish Environment and Health Study was set up to assess the impact of the environment on the health of the general Flemish population. As a part of the biomarker research in this study, the expression of a set of 8 genes was measured in 398 adult men and women, and correlated to 'traditional' biomarkers of exposure and effect. In a later phase, 40 RNA samples of this population were used in whole genome analysis to further explore relationships between gene expression and (other) biomarkers of exposure and effect. This study goes even further, and includes whole genome analysis of 100 more RNA samples in the context of biomarker research (part 3 of this study).

Attention will be given to the different technologies used to analyze gene expression and their specific characteristics, as well as related technologies such as proteomics and metabolomics; data analysis, an overview of international efforts on gene expression in biomonitoring; and opportunities and considerations concerning the use of gene expression as a biomarker for regulatory purposes.

1. DEVELOPMENT OF NEW BIOMARKERS: GENE EXPRESSION

1.1. Traditional risk assessment: toxicology and biomarkers

Risk assessment includes both evaluation of toxicology of chemical substances and of (human) exposure to these substances. A number of key problems or challenges in traditional risk assessment, with implications for biomarker development include: mode of action identification of toxicants (Oberemm *et al.* 2005); interpretation of inter- and intraspecies variability and the evaluation of relevance for human exposures (Oberemm *et al.* 2005); dose-response relationships, and especially low dose extrapolations (Oberemm *et al.* 2005); the detection and interpretation of multiple (combined) chemical exposures, which are a reality when dealing with environmental exposures (Aardema & MacGregor 2002).

1.2. Gene expression as a biomarker

The use of gene expression as a biomarker could help overcome some of these challenges. A number of arguments on why gene expression provides promising opportunities for biomarker development are discussed here.

Changes in gene expression as direct or indirect result of *toxicant* ***exposure*** have been increasingly documented over the last years. This has been demonstrated in human and animal cell lines, and in vivo in animal tissue for a wide range of chemicals. Several initiatives have been set up to group these data into central databases. One example is the *Comparative Toxicogenomics Database*[1]. In November 2009, this database contained data from: 5.259 unique chemicals, 16.899 unique genes, 306 unique organisms and 14.672 unique references, together resulting in a total of 199.453 reported gene-chemical associations (including recurrent associations). Numbers of genes reported to be up or down regulated by some example chemicals are illustrated in table 1. These represent the numbers of genes currently reported to be modulated in expression (up or down regulated) as a result of exposure to a certain chemical. These data can be derived from analysis of differential expression after an exposure, as well as from observed dose-response relationships.

[1] The Comparative Toxicogenomics Database includes manually curated data describing cross-species chemical–gene/protein interactions and chemical– and gene–disease relationships to illuminate molecular mechanisms underlying variable susceptibility and environmentally influenced diseases. Data are derived from the published literature. The majority of these studies involve only a relatively limited amount of genes (e.g. by real-time-PCR analysis). (CTD, 2009).

For example, the CTD database currently contains 197 genes that are reported to be differentially expressed as a result of exposure to cadmium (all species included). According to this table, the CTD database includes 197 unique gene-chemical associations for cadmium and a total of 295 gene-chemical associations (thus, including 98 recurrent associations). These associations are derived from 135 publications (CTD 2009). Disturbance of homeostasis by toxicant exposure can thus result in new biomarkers. Modulation of gene expression may lead to certain *pathological* ***effects***. Since most pathological processes are active events under genetic control, gene expression analysis is a powerful tool to monitor these processes (Aardema & MacGregor 2002). Moreover, these gene expression changes will generally appear before a clinical manifestation of toxicity, which even enlarges the opportunities of gene expression as a biomarker (Rockett *et al.* 2004). Several databases list known gene-disease relationships or gene expression profiles of diseases and disorders. Amongst others, the *Comparative Toxicogenomics Database* and *Genelogic Bioexpress* include these (CTD 2009), (Genelogic 2009). Of course, there are also important cellular events that are not directly involving differential gene expression (e.g. phosphorylation, posttranslational modifications, …), but even so it seems that often these changes are still reflected by other transcriptional changes (David *et al.* 2005). The expression of certain genes could thus serve as '*markers of* ***early biological effect'***, when their changes are situated on the timeline between a toxicological insult and a pathological outcome. This offers interesting opportunities for using gene expression as a biomarker.

	Gene-chemical associations		
	#unique	#total	#publi ca-tions
Cd	197	295	135
Pb	44	55	40
PCBs	467	892	202
TCDD	562	994	198
Benzene	353	475	102
PAHs	7044	22094	1941
As	324	244	62

Table 1: Unique and total relationships between gene expression and chemicals reported in the Comparative Toxicogenomics Database for cadmium, lead, PCBs, TCDD, Benzene, PAHs and arsenic (CTD 2009)

Gene expression profiling has been responsible or has helped in the identification of the ***mode of action*** of toxicity of chemicals.

One example of the use of gene expression for mode of action identification is the mechanism of co-carcinogenesis of arsenic. It has been suggested that the inhibition of DNA-repair is responsible for this co-carcinogenesis. Shen et al. (2008) investigated the changes in gene expression using real-time PCR of several key repair proteins involved upstream of the dual incision in the global nucleotide excision repair (NER) pathway: p53, p48, XPC, XPA, and p62 TFIIH. Of these genes - only p53 appeared to be significantly down regulated. These observations suggest that downregulation of p53 gene expression, leading to reduced p53 DNA-binding activity, is at least in part responsible for the co-carcinogenic effects of arsenic.

Interspecies variability is a known problem in risk assessment. Gene expression profiling adds to mechanistic knowledge, which can then lead to a better understanding of the mechanisms behind

Interspecies variability of toxic effects is known for trichloroethene, which causes hepatocellular carcinoma in mice and renal-cell carcinoma in rats. Sano et al. (2009) have shown that these differences are reflected in transcriptomic profiles and biological pathways associated to these profiles. For example inhibition of the TGFbeta pathway and activation of MAPK signalling were specific to mice exposed to TCE.

this variability.

Changes in gene expression associated with toxicity are often more ***sensitive*** and characteristic of the toxic response than currently employed endpoints of pathology. This offers possibilities to monitor defence responses and pre-pathological compensatory responses to cellular damage, which could then result in new biomarkers of sub-pathological cellular changes. Especially in human environmental biomonitoring studies this is an important aspect, because exposure is generally lower than in experimental conditions. This sensitivity also helps for ***low dose extrapolation*** of effects, as it leads to a better understanding of mechanisms involved in the toxicity of exposures at lower doses (Aardema & MacGregor 2002).

Cheng et al. (2003) compared gene expression changes resulting from several levels of exposure to nickel(II). Human lung epithelial cells were exposed to 3 'non-toxic' levels of nickel (50, 100 and 200µM), and 3 toxic levels (400, 800, 1600 µM). Gene expression profiles of non-toxic concentrations were markedly different from those of toxic-concentrations. Among the 113 genes that were differentially expressed by 2-fold or more at the lower concentrations, several genes are potentially cancer related, such as RhoA, dyskertin, IRF1, RAD21 and tumor protein, translationally controlled. Findings like these are obviously important when extrapolating effects from high to low doses.

An interesting opportunity with whole genome monitoring are so called '***fingerprints***' or patterns in gene expression that are typically associated with certain exposures or effects. Identification of these specific fingerprints is being pursued from several scientific disciplines, such as drug discovery, clinical diagnosing, (eco)toxicology and biomonitoring studies, and some promising results have been obtained (Aardema & MacGregor 2002, Rockett *et al.* 2004, Woods *et al.* 2007). In addition, identification of key genes regulating pathways is an opportunity generated by gene expression analysis (Pennie *et al.* 2001).

An illustration of the use of 'fingerprints' is the discrimination of genotoxic and non-genotoxic carcinogens based on their gene expression profiles. van Delft (2005) managed to correctly predict the carcinogenic class in 11 out of 12 cases for the training set (92%), and for all 3 cases for the validation set (100%), using only 20 selected genes.

A common problem in risk assessment and biomonitoring, is ***multiple (combined) chemical exposure***. As environmental exposures are usually 'combined', and interactions between exposures such as synergism, additivity or antagonism can occur, and substantially alter the outcome of an exposure, this is a very relevant issue. Gene expression analysis can help in the comparison of single and multiple exposures in vitro, and - in a later phase - also in vivo, e.g. in biomonitoring because it combines information of different stressors at the effect level (Aardema & MacGregor 2002).

Effects on gene expression of a number of genes related to estrogen and aryl hydrocarbon receptors of combined exposures to nonylphenol (NP) and PCB77 were investigated by Mortensen et al. (2006). Exposure to NP alone resulted in a significant elevation of ERalpha, ERbeta, Vtg and Zrp expression, while combined exposure with PCB77 inhibited NP induced ERs and their target gene expressions. On the other hand, PCB77 exposure lead to a rapid increase of CYP1A1 mRNA and an decrease of AhRalpha and beta expression, while combined with NP, the induction of CYP1A1 was largely reduced and AhRbeta mRNA was restored to baseline levels. Gene expression research helps in the mechanistic research of these exposures, and may prove to be important in the selection of biomarker genes.

Although it may be too early evaluate, the use of gene expression as a biomarker may well be **cost-effective**. This stems from the fact that a huge number of measurements (expression of more than

20.000 genes) can be conducted simultaneously at relatively low cost. Since information on genes active in a wide range of different biological functions and pathways can be revealed, this method could be more cost-effective than the use of traditional biomarkers – requiring separate –labour intensive measurements for each separate biomarker. We also refer to part 6.2, where the use of gene expression in the diagnosis and outcome prediction is considered to be potentially cost-effective, due to the high costs of chemotherapy (Oncotest DX 2009).

At present, the use of gene expression as a biomarker of exposure and (especially) of effect in a human environmental context is still in a phase of (early) development – as is illustrated by the limited amount and scope of the listed studies (see part 5). However, in other areas of research, such as clinical diagnosis and pharmacology, research had advanced to higher levels and has lead to the ***actual development of gene expression biomarkers***, further illustrating the potential of gene expression profiling in other contexts.

*A clear **clinical** advancement in the development of gene expression as a biomarker is the recognition of the **Oncogene DX test** for use as a prognostic and predictive test by the American Society of Clinical Oncology Breast Cancer Tumor Markers Update Committee. This test includes mRNA profiling of 21 genes – RT PCR? (16 cancer related, 5 reference), which is used to form a Recurrence Score for women with early stage node-negative ER+ invasive breast cancer. It is also useful for estimating the magnitude of chemotherapy benefit. Due to the high cost of chemotherapy, these tests can be cost-effective. (Oncotest DX 2009). Another prognostic test for early stage lymph node negative breast cancer - **MammaPrint** – has been approved by the FDA, and uses a gene expression profile of 70 genes to determine the recurrence risk class (10-15% chance of developing metastasis within 10 years vs. 50% chance). These tests are already in use today, and show the potential of gene expression profiles in a practical context (Agendia 2009, Ross 2009).*

*Other examples of actual biomarker development come from **drug testing and development**. For example, Fielden et al (2005, 2007, 2009) have developed gene expression signatures as a biomarker for both hepatic tumor induction and renal tubular toxicity. 64 Nephrotoxic and non-nephrotoxic compounds were used to develop a set of 35 biomarker genes predicting future development of nephral tubular degeneration weeks before appearing histologically. This signature was able to predict the ability of a compound to induce tubular degeneration for 16 out of 21 compounds (76%), which is far better that tradition approaches (Fielden et al. 2005). Similarly, 100 nongenotoxic hepatocarcinogens were used to develop a gene expression biomarker set which was able to detect hepatocarcinogenecity of 47 test chemicals with a sensitivity and specificity of 86% resp. 81%, which is more accurate than actual tests such as liver weight, hepatocellular hypertrophy, hepatic necrosis, serum alanine aminotransferase activity, induction of cytochrome P450 and repression of Tsc-22 or alpha2-macroglobulin mRNA. Moreover, the gene expression profiles provided information on the mode of action of the tested substances*

A potential problem with the use of biomarkers based on gene expression, as with any biomarker, is the need to ***distinguish homeostatic processes or adaptive responses from actual adverse effects*** or processes. As was illustrated above, gene expression markers have been identified for certain effects (e.g. through mode of action identification). However, links to effects are not always clearly made. Therefore, it is important to link gene expression in biomonitoring to other biomarkers (e.g. micronuclei, comet assay,...), as well as other levels of research (mechanistic, mode of action information). Things to consider in the assessment of whether a change is really an adverse effect are: is there a dose-response relationship, what is the normal biological variation, are the effects transient, are the effects isolated or interdependent, is the effect a

Several researchers have investigated the relationship between micronuclei or the comet assay and gene expression profiles. For example, Arsenite-induced DNA damage in mouse cells, detected by the comet assay, and genomic instability detected by the micronucleus test, was accompanied by altered expression of 456 genes, of which 20% had known functions in cell cycle and DNA damage signaling and response, cell growth and/or maintainance (Newman et al. 2008). A similar observation was made by Frotschl. et al. (2005), investigating the effects of chlorpromazine combined with UV radiation on HepG2 cells using gene expression profiling and the comet assay. Komichi et al. (2005) have associated glycochenodeoxycholate exposure in mice cells with 8-hydroxydeoxyguanosine (marker for oxidative DNA damage) with gene expression changes (related to cell proliferation, angiongenesis, invasion and metastasis).

precursor of a known adverse effect (Lewis *et al.* 2002).

Furthermore, distinguishing between biomarkers of exposure and effect is not always easy, both for 'classic' biomarkers as for gene expression as a biomarker. An example is benzene-DNA adducts: these are used as a biomarker of exposure to benzene, but could possibly be used as a biomarker of effect (of genotoxicity of benzene), or even as a biomarker of susceptibility (subjects with a high level of benzene bioactivation). Either way, these adducts are useful in biomonitoring, being a biomarker for exposure or effect, but the use of it will be different: assessment of benzene exposure versus genotoxic effects in the population, or both (Manno *et al.* 2009).

2. OVERVIEW OF GENE EXPRESSION ANALYSIS AND RELATED TECHNOLOGIES

2.1. Omics-technologies

2.1.1. Transcriptomics

a. Introduction

Transcriptomics studies the transcriptional level, the mRNA. The human genome is estimated to consist of 20-25.000 genes, which are transcribed into mRNA before being translated into proteins. Due to alternative splicing of the pre-mRNA (estimated to occur in 80% of the human genes), the number of possible mRNA species is even higher. DNA microarrays and real-time PCR are widely used methods for analyzing gene expression. The possibility to measure gene expression of the whole genome simultaneously provides enormous opportunities for among others, risk assessment, medical diagnosis, pharmacology and biomonitoring (David *et al.* 2005).

b. Technologies

Differential display is a technique relying on reverse transcription, PCR amplification and separation on polyacrylamide gels. Bands are then visualized, for example by autoradiography. Several (reference) samples can then be compared to each other. **Serial analysis of gene expression (SAGE)** is based on the extraction of small "tags" in the mRNA, linking these to each other, amplification in a vector and sequencing of the resulting chain. **Subtractive hybridization** is based on selective amplification of cDNA fragments that are differentially expressed in two samples. It relies on the removal of double strand DNA formed by hybridization of two samples. Genes remaining in the mixed sample are either unique, or strongly up regulated in either of the samples. Genes up or down regulated by 5 or 10fold are reliably detected in the samples (David *et al.* 2005). All of these three techniques require a validation of the identified genes by PCR or Northern Blotting, and have preceded microarray technology (David *et al.* 2005).

DNA microarrays. Gene expression profiling is performed using high density microarrays, consisting of oligonucleotides, or cDNA attached to a solid support (glass or silicon slide). These *probes* can capture specific RNA or DNA sequences in the sample. RNA is isolated from a sample, most often reverse transcribed into cDNA, fluorescently labelled, and hybridised to the microarray. The fluorescent labels are read with a fluorescence scanner (David *et al.* 2005). Since the whole genome is hybridized at once to the microarray, the throughput of this method is much higher than for real-time PCR analysis, in terms of the number of genes that can be measured.
Oligo and cDNA microarrays differ in the type of probe that is used. cDNA is derived from isolated and amplified RNA reverse-transcribed to cDNA from samples, while oligonucleotides are synthesized by coupling or building of single nucleotides. cDNA arrays have the advantage that sequence information is not required, and cross hybridization (hybridization of gene products to the wrong probes) is limited (because of the length of the probes) compared to short oligo (< 50-mer) microarrays. The process of manufacturing the cDNA for the whole genome is, however, labour- and

time-consuming. Oligoarrays are a widely (commercially) available platform, and a generally accepted method for gene expression analysis. Cross hybridization is a potential problem, but the inclusion of several oligos representing the same gene and/or the use of longer oligos (>50 nucleotides) has made it possible to eliminate this problem (Barrett & Kawasaki 2003).
We can further distinguish *two-colour and one-colour microarrays***.** With two colour microarrays, two samples are hybridized together on one slide, each labelled with a specific fluorescent label (Cyanine-3 and Cyanine-5). This could be used to compare two treatments or conditions, or to compare a sample to a common reference sample. The Cy3/Cy5 ratio is then calculated based on the fluorescence measurements. With one-colour microarrays, only a single sample is hybridized on each array. Hybridizations are fully independent. Measured fluorescence represent absolute values of gene expression. Thus, two colour microarrays yield a ratio of fluorescence for each gene between one sample and another (another independent sample, or a reference sample), while single colour arrays yield absolute values of gene expression for each gene. As is discussed in 3.2., these values (ratio's or absolute values) will be normalised and transformed, before further statistical analysis is done (Barrett & Kawasaki 2003).

There are several microarray-platforms available. The three most used platforms will be discussed. *Agilent* offers one- and two colour microarrays, based on 60-mer oligonucleotide probes, manufactured by photolithographic synthesis[2] on a glass slide. The *Affymetrix*-platform relies on 25-mer oliginucleotides manufactured by photolithographic synthesis on a silica substrate, but includes – next to a perfect match oligo – also a 'mismatch oligo', differing only in one nucleotide. This mismatch oligo should not hybridize, and is used as a control for mismatch hybridization. *Illumina* uses a different approach, as oligonucleotides are manufactured separately and spotted on microbeads that are randomly attached to the microarray. This requires an extra probe decoding step for each probe using a molecular address, since probes are not placed on a predefined spot on the array. Illumina arrays include several (technical) replicates per probe (Barnes *et al.* 2005).

The intensity of fluorescence of spots on a microarray slide is quantified, by selectively exciting fluorophores with a laser and measuring the fluorescence with a filter (optics) photomultiplier system, yielding the relative abundance of nucleic acid sequences in the target.
Costs per (Agilent 44k) microarray analyzed are currently ca. €250 plus an additional €100 for materials and fluorescent labels (info:GRAT june 2009). Additionally, 8x15k arrays (slides with 8 times 15k probes) are available which include a specific selection of genes, reducing costs per sample to 50% of that of whole genome arrays.

Quantitative real-time PCR. The amount of a specific mRNA reverse transcribed to cDNA in a sample is amplified using the polymerase chain reaction, and the accumulation of the PCR product is measured in real-time during the course of this reaction, using a fluorescent label that binds the DNA. The amount of cDNA (transcribed from mRNA) in the original sample is calculated by comparing the measured curve derived from the sample, to the standard curve(s) of a known amount of DNA. Throughput is limited to the number of genes that are measured, as opposed to (whole-genome) microarrays, whereas sensitivity is higher for RT-PCR as opposed to microarrays (David *et al.* 2005).

[2] light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array

An interesting improvement of real-time PCR is the *TaqMan Custom Array*, where 8 samples can be simultaneously analysed on expression of 384 genes (pre-spotted plates) (TaqMan 2009). The Biomark 96.96 Dynamic Array (Fluidigm), is a similar technology, allowing real-time PCR analysis of 96 samples on 96 genes simultaneously (Fluidigm 2009). These developments may be one step closer toward a higher throughput real-time PCR technique.

Costs per 96-well PCR analysis (43 samples in duplo for 1 gene) are currently ca. €75 plus an additional €3.5 per sample for cDNA synthesis (useable for all genes measured) (info: GRAT, june 2009).

Next-generation sequencing is another promising approach to gene expression analysis. This includes (for transcriptomic purposes), the complete sequencing of all mRNA copies present in a sample (after being converted to cDNA using reverse transcriptase) (Shendure & Ji 2008).

Digital mRNA profiling is yet another new approach to gene expression analysis. The Nanostrings nCounter system relies on the mixing of sample mRNA with reporter and capture probes that hybridize. Excess probes are then washed away, and the purified ternary complexes are bound to the imaging surface and immobilized. Finally, reporter probes, representing individual copies of mRNA are decoded and counted. Sensitivity is reported to be close to that of PCR based methods. In a recent report over 500 genes could be multiplexed in a single reaction (Fortina & Surrey 2008).

c. Considerations

A question that has been raised is concerning the comparability of the different microarray platforms. The Microarray Quality Control project was initiated to address this issue and evaluate reproducibility, specificity, sensitivity and accuracy. Identical RNA samples were analyzed using several different platforms, including the Affymetrix, Agilent and Illumina platforms. This study showed intra-platform consistency across laboratories, as well as high interplatform concordance in terms of genes identified as differentially expressed. Furthermore, Affymetrix, Agilent, and Illumina platforms displayed high correlation values of 0.90 or higher with TaqMan (RT-PCR) assays based on comparisons of expression values of 450–550 genes. This indicates both qualitative (genes identified as differentially expressed) and quantitative (expression ratios) concordance and reproducibility across platforms and technologies. Data quality is concluded to be essentially equivalent across several different platforms, and it is suggested that platform type should not necessarily be a determining factor in decisions regarding experimental microarray design (Patterson *et al.* 2006, Shi *et al.* 2006).

An obvious consideration of transcriptomic analysis is that several functions are determined post-transcriptional, at the translational and posttranslational level. Thus, it has to be kept in mind that, however monitoring the whole genome, not all changes induced for example by exposure to environmental pollutants will be detected on the level of gene expression. However, it appears that often, these changes can in turn be reflected by other transcriptional changes (David *et al.* 2005).

Alternative splicing of mRNA – is another important factor to consider when measuring gene expression, since one gene can give rise to several different new mRNA species (estimated to occur in 80% of the human genes). By inclusion or omission of certain parts of the pre-mRNA before mRNA

formation, this results in several possible isoforms of the protein. For known splice variants, this can be dealt with by using multiple probes per gene, representing different splice variants. Current whole genome arrays include several known splice variants for a number of genes (David *et al.* 2005).

2.1.2. Proteomics

a. Introduction

Proteomics studies the translational expression level, the proteins. Serum is estimated to contain about 20.000 proteins, with a concentration range of 10 orders of magnitude. Proteomics provides further insight into systems biology due to a number of factors: post-translational changes are not detected by transcriptomics, mRNA degradation may alter the amount of mRNA actually translated into protein, and phosphorylation state of proteins affects their activity (Anderson & Anderson 2002).

b. Technologies

Two dimensional gel electrophoresis (2DGE) is traditionally used to separate, visualise and quantify proteins, usually followed by mass spectrometry for protein identification. These procedures are limited to small amounts of samples, high amount of available material and medium to highly abundant proteins. Identifiable amount of proteins in a sample: several hundred (Kussmann *et al.* 2008).

Differential imaging electrophoresis (DIGE) uses sample and control proteins, each labelled with a specific dye, and co-separated, visualised and quantified on one gel. This technique makes identification of the proteins easier, but still has limited throughput capacity and dynamic range (Kussmann *et al.* 2008).

Multi-dimensional protein identification technology (MudPIT) is based on direct analysis of protein mixtures. Separation is performed on the level of the peptide (using multi dimensional liquid chromatography), instead of the protein (as in 2DGE). Separated peptides are then fragmented by tandem MS and proteins identified based on the resulting spectra. Identifiable amount of proteins in a sample >1000 (Motoyama & Yates, III 2008).

Protein-chip: proteomic profiling on chips is another new approach. Microarrays have been fabricated with capacities ranging from a few to several thousands of proteins. Antibodies or antigens on chips are mostly used to capture the proteins in solution, but also nucleic acids, receptors, enzymes, and proteins have been used. Fluorescence of the (labeled) bound proteins is used for detection and quantification (Collings & Vaidya 2008, Wingren & Borrebaeck 2004)

Protein quantification can be achieved through staining (see DIGE). Another option is the incorporation of stable isotopes in peptides and/or proteins in living cells. Chemical tagging can be used in ex vivo tissues of fluids. Mass spectrometry is not quantitative, as the intensity of the signal is dependent on the nature of the protein, but due to its high reproducibility, it has the potential to enable so-called label-free proteomics (Kussmann *et al.* 2008).

Because of the huge dynamic range of proteins in mixtures, protein and peptide separation are usually insufficient, requiring selective **depletion and enrichment**. Depletion of highly-abundant proteins can be performed using antibody columns. Similarly, enrichment of the low abundance proteins can be performed, e.g. using chemical scavengers with affinity to phospho- or glycoproteins (Kussmann *et al.* 2008).

c. Considerations

The major challenge in proteomics is the large dynamic range of proteins (Oberemm *et al.* 2005). Because of the complexity of protein samples, extensive pre-separation, depletion and enrichment steps are required, before proceeding to qualification and quantification of the proteins (Kussmann *et al.* 2008). At the moment, all strategies (still) have one or more limitations (throughput, quantification, data processing and interpretation), and several of the used techniques are still in a phase of development or optimization.

2.1.3. Metabolomics

a. Introduction

Metabolomics measures endogenous metabolites in tissues en body fluids. These metabolites have a relatively low molecular weight (<1000Da), in contrast to the much larger proteins and nucleotides. Included are amino acids, oligopeptides, sugars, bile acids, simple fatty acids, and intermediates of many pathways (e.g. tricarboxylic acid cycle, glycolysis, …). Metabolic profiling gives insight into the physiology of the cell.

b. Technologies

Nuclear magnetic resonance is a technique that scans for a proton spectrum of a sample. These spectra are based on the resonance of C-H bonds, distinguishing between more stable (e.g. aromatic rings) and more flexible bonds (e.g. aliphatic methyl groups). Biological samples typically generate several 1000 peaks. Detection is quantitative, with a detection limit of ~10^{-5}M (Clarke & Haselden 2008).

Mass spectrometry (MS). MS is usually preceded by liquid or gas chromatography separation. Separated molecules are then ionized and sprayed in the MS. The detection limit is situated around 10^{-12}M, resulting in a detection of smaller molecules such as sulphate, amine and carboxyl groups. Biological samples generate up to 15000 peaks (Clarke & Haselden 2008).

c. Considerations

A large number of molecules can be detected, but only a small fraction (5-10%) consists of known components. The limited database of known metabolites is a current restriction of the technique, compared with the further developed transcriptomic and proteomics databases (Clarke & Haselden 2008).

2.2. Data analysis in toxicogenomics

2.2.1. Design en statistics

Scanning the microarrays yields microarray images that have to be processed in order to obtain relative gene expression data – per gene and per sample. These resulting data are then normalized. Further possible pre-processing includes transformation and data filtering (Allison *et al.* 2006). Subsequently, statistical analysis of ***differential expression*** can be carried out by among others t-tests, SAM (significance analysis of microarrays), ANOVA, or correlation analysis.

There are several possibilities for experimental design in microarray experiments, the choice of which will depend on the approach and goals of the experiments. These designs are mainly important for two-colour arrays, where relative values are measured, as opposed to single-colour arrays where absolute values are measured, and no 'comparison' is needed. A main subdivision can be made into direct and indirect designs. ***Direct designs*** involve only the samples of interest, whereas ***indirect designs*** use a reference sample to compare all samples of interest to. An example of direct design is the *loop design*, where every sample is compared to a 'next' sample, and where the last sample is compared to the first one. Although the loop design results in a lower variance (for a same amount of microarrays), it is usually only suited for relatively small numbers of samples, as the loop grows with the sample number and one bad sample can disrupt the whole readout. Also, statistical analysis gets more complex with higher sample numbers. An example of an indirect design, is the *common reference design*, where all samples are compared to a common reference sample. This reference can be a pool of all (or a part of the) samples in the experiment, or 'universal' reference sample used for all experiments in a laboratory. This design is preferred for larger numbers of samples, because all samples are compared independently to the same sample, which simplifies statistical analysis, and limits the consequences of one bad sample for the experiment. Furthermore, several variations or combinations are possible, for use in time-course experiments or multifactorial experiments where several conditions are compared to each other (Kerr & Churchill 2001, Yang & Speed 2002).

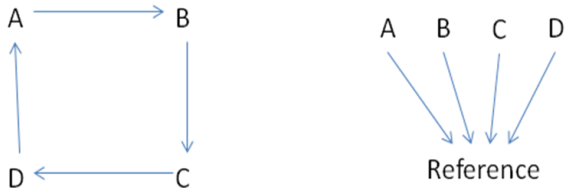


Figure 1: loop design(left) and common reference design (right)

Microarray studies can generally be subdivided into three categories: class comparison, class prediction, and class discovery. ***Class comparison*** is the comparison of results from differently *predefined groups*. In environmental biomonitoring, this could be a comparison of a group of high exposed vs. a group of low exposed individuals in a population. Statistical analysis in class comparison can be t-tests, which yield a p-value and a fold change value (measure for the actual difference between the groups). ***Class predictions*** involve *predefined* classes as well, but emphasis lies on the ability to develop a gene (or protein) based multivariate function that reliably predicts the class membership of a new sample based on gene (or protein) expression levels. ***Class discovery*** is unique in that it does not involve predefined groups. Clustering is used to identify classes in the data,

or multivariate evaluation is used to find common patterns of gene (or protein) expression across different samples, to identify genes (or proteins) with related functions, or to group samples with common properties. Obviously these three categories, each require a different approach in study setup (Gundert-Remy 1999, Gundert-Remy *et al.* 2005). Most environmental biomonitoring studies involving gene expression analysis are class comparisons, where a difference in (environmental) exposure is linked to differences in gene expression.

Microarray data analysis usually involves *large numbers of statistical tests* (at least one test per analysed gene). The relatively high cost of microarray chips usually also limits the number of arrays used (per group) and/or the number of replicates. As a result, there is an increased possibility of erroneous inference. Usually, a statistical test involves a null-hypothesis, that states that there is no association between two given variables (e.g. a gene and another variable, or the mean expression of a gene in two groups). The lower the resulting p-value, the lower the chance that this null-hypothesis is true. When such tests are repeatedly performed (i.e. for each gene one or more times), the total chance of false positives rises (null-hypothesis falsely rejected). However, when the threshold (p-value) is lowered, there is a decrease in false positives, but also an increase of false negatives. Therefore, carefulness is required when choosing a p-value threshold.

The most commonly used multiple-testing procedure for microarray data is the **false discovery rate (FDR)**, and can be interpreted as the expected proportion of significant findings that are indeed false positives. P-values can be used to *estimate* or *control* the FDR and related error rates. FDR must be set to a certain value, which neither allows too many false positives in the results, as well as prevents too many false negatives (i.e. sensitive enough). A number of statistical methods are available to calculate the FDR (Gusnanto *et al.* 2007, Pounds 2006).

2.2.2. Gene Ontology and pathway analysis

Gene expression microarrays analyses typically yield large lists significantly differentially expressed genes. However, these genes don't function independently, but actually form a complex biological unity.

A first way to deal with this fact is by using **gene ontology (GO)**, an expert-curated database assigning genes to various functional categories. These categories are split into three subdivisions: biological processes, molecular functions and cellular components. Currently ca. 17500 genes are annotated to one or more categories, and the database is regularly updated. Several tools use these GO-annotations to verify if in a list of genes – significantly up- or downregulated – there is an over- of underrepresentation of GO-terms compared to a background list of genes (for all non-significant genes in an experiment, or the rest of the genome). These over/underrepresentation could tell us something about biological, molecular mechanisms that are hidden behind these genes that are differentially expressed. Examples of software using these GO-terms are Babelomics FatiGO and DAVID (Werner 2008).

Yet a level closer to system biology are **pathway analysis**. This approach focuses on physical and functional interactions of genes. These pathways can be viewed as chains of events. Although there are steps in these pathways not controlled by mRNA, but by other (e.g. protein-based) events, there

is enough transcriptional feedback regulation to allow identification of pathways trough mRNA level changes Most pathway analysis tools are based on databases derived from literature, and are therefore constantly updated in accordance with new insights into pathways (Werner 2008).

Biological processes usually rely on an interaction of several pathways, which acts like a network. Research around the **network approach** of gene expression data is ongoing, and involves regulatory sequences, such as transcription factor binding sites (TFBS) of promoters (Veerla & Hoglund 2006).

2.2.3. Interpretation and integration of data: the systems biology approach

Systems biology is an integration of gene, protein and metabolite information. This stems from the fact that – of course – these levels do not function independently. A few studies on the integration of these different levels have been published, and interesting results have been obtained. For example, Fan *et al.* (2006) investigated differential gene expression, proteome and metabolic networks induced in A549 lung cancer cells by selenite. Metabolic profiles were used to support or clarify transcriptomic data. Amongst others, they found changes in the mRNA expression profile superimposed on glycolysis, pentose phosphate pathway, citric acid cycle glutaminolysis and fatty acid metabolism. In these pathways, downregulation of 5 glycolytic genes was linked to reduced synthesis of lactate or glycolytic flux in selenite treated cells; similarly, reduced expression of mitochondrial pyruvate dehydrogenase, a-ketoglutarate dehydrogenase, isocitrate dehydrogenase, and malate dehydrogenase in selenite treated cells were linked to an attenuated citric acid cycle, which –together with a reduced glycolysis - lead to the reduced synthesis of glutamic acid, fumarate and malate from glucose. This example illustrates the possibilities and opportunities of a multi-level approach.

To date it is impossible to select the better of these ‘levels’ with regard to delivering biomarkers. Apart from technical and practical limitations that are seen in proteomics and metabolomics (see before), up to now – if possible - a combination of the levels seems the most interesting approach when looking at the whole of systems biology. Future research will further point out strengths and weaknesses of each level for biomarker development.

3. CHOICE OF STUDY SETTINGS

3.1. Gene expression as a biomarker

In an ideal situation we would want to monitor using a combination of different omic-technologies in accordance with the path leading from exposure to effects, beginning with transcriptomics (gene expression), trough proteomics (protein formation) and finally metabolomics (endogenous metabolites). However, compared to whole genome expression analysis, whole proteome and whole metabolome analysis still suffer far more from technical and practical issues. The huge dynamic ranges and the incompleteness of proteomic and metabolomic databases are two major disadvantages, which have been overcome to a larger extent for transcriptomics. Additionally, microarray-based transcriptome analysis was the first 'mature' genome-wide profiling technology, and as a result is also the most used technology, which offers more opportunities to compare results with those derived from other projects (Kussmann *et al.* 2008).

3.2. Microarrays to measure gene expression

As is discussed in §3.1.1., the two main technologies used to measure gene expression each have their own strengths. Roughly, real-time PCR is more useful when measuring the expression of a limited set of genes in a larger population, whereas microarrays are more useful when measuring the whole genome in a more limited population. When trying to identify biomarkers of exposure and effect in a population, whole genome microarrays obviously offer a more complete picture of what is biologically going on at the level of gene expression. The reduced sensitivity of microarray measurements of gene expression compared to real-time PCR is a drawback, but is currently inevitable, when monitoring the whole genome. However, this may be only a limited problem, because gene expression changes significantly picked up by microarrays might well be more interesting than changes that are too small to measure by microarray. Validation of microarray data by RT-PCR can be used to verify microarray results.

3.3. Choice of medium: peripheral blood

Of course, in biomonitoring, there is a need of an **easily accessible** tissue, since biopsy of inaccessible tissues is only feasible when there is a strong medical reason to do so. Peripheral blood is **easily accessible** and is already routinely used for biomonitoring, or medical diagnosis. One example is the screening for prostate specific antigen (PSA), to diagnose prostate cancer. Also, only a few hundred microliters of blood is sufficient to conduct gene expression analysis (Rockett *et al.* 2004).

Peripheral blood contains **living, nucleated cells** (with extractable RNA), which is another condition that is required when measuring gene expression (Rockett *et al.* 2004).

A number of studies indicate that **gene expression in peripheral blood cells partly reflects gene expression in other tissues**, giving good opportunities of using peripheral blood as a surrogate tissue for biomonitoring and diagnosis (Liew *et al.* 2006, Rockett *et al.* 2004, Sullivan *et al.* 2006). In a comparative study of gene expression in different tissues, Liew *et al.* (2006) found an overlap of 80%

or more (16.000 genes) between the list of genes expressed in 9 different tissues[3] and the list of genes expressed in peripheral blood cells. These amounts can clearly not be covered by housekeeping genes alone. An ideal surrogate tissue should express many genes, in particular genes responsive to exposures and effects. The fact that most genes come to expression adds to the proof that blood cells can be a successful surrogate tissue (Liew *et al.* 2006). Inevitably, the expression of some genes is cell-type dependent. These should be identified and considered when working with gene expression as a biomarker.

An important extra criterion is that peripheral **blood passes through many different parts of the body,** and through different organs, and is thus exposed to the physiological environment as the target cells (Sullivan *et al.* 2006).

Based on the above assessment, blood is currently the most obvious and practical surrogate tissue for human gene expression biomonitoring. Several studies confirm the usefulness of peripheral blood for exposure and/or effect assessment in biomonitoring, as is illustrated in section 5 (Amundson *et al.* 2004, Rockett *et al.* 2004).

In addtion, we mention that experiments with saliva as a medium for gene expression analysis have also been conducted. Several pioneer studies have indicated a potential for the use of saliva transcriptomics as an early biomarker for oral or even systemic diseases (Hu *et al.* 2006, Li *et al.* 2004b, Li *et al.* 2004a). However, several problems exist: saliva usually contains low concentrations of informative molecules, and it naturally contains bacteria and food debris, all of which complicate gene expression analysis. Therefore, one study focussed on the use of cell-free saliva, but – not surprisingly – rather low amounts of genes were detected. Over time, saliva could turn out to be a non-invasive medium for transcriptomic analysis in biomonitoring. However, the use of saliva for these purposes is still in an early stage of development, compared to blood transcriptomics, and needs further study (Hu *et al.* 2006, Li *et al.* 2004b).

[3] Tissues were derived from the brain, colon, heart, kidney, liver, lung, prostate, spleen and stomach.

4. PROJECTS INVOLVING THE USE AND DEVELOPMENT OF GENE EXPRESSION AS A BIOMARKER

4.1. Introduction

The purpose of this section, was to make an inventory of projects that explore the possibilities of gene expression analysis for human biomonitoring and biomarker development. First, the scope of the literature research is outlined. Second, an overview of studies (past and ongoing) is given, including initiatives around the regulatory use of gene expression. Third, an overall discussion, with attention to general conclusions and evolution within these studies is presented.

4.2. Methods

Since gene expression research extends into a wide scope of scientific branches and includes a large number of possible study-settings (in vivo/in vitro, human/animal, ...), this inventory of relevant studies is limited to (mostly) ***whole-genome*** studies of ***human*** populations ***environmentally, occupationally or medically exposed*** to chemical exposures.

Peer reviewed publications complying with these criteria were retrieved using - amongst others - following search terms: 'microarray', 'gene expression', 'whole genome', 'human', 'environmental exposure', 'occupational exposure', 'cadmium', 'lead', 'polychlorinated biphenyls', 'polyaromatic hydrocarbons', 'tobacco smoke', 'metal', 'arsenic', 'benzene', 'ionizing radiation', ...

4.3. Results: overview of past and ongoing studies involving gene expression as a biomarker

4.3.1. Peer reviewed publications

A list of peer reviewed publications (and ongoing projects awaiting publication) are listed in annex 1. An overview of characteristics, results and conclusions of these publications is given in table 2. These publications are grouped by the nature of the exposure under study: environmental, occupational and medical exposures.

TABLE 2

Exposure	Population	Set-up	Potential confounding factors taken into account	Medium	Platform	Statistical analysis	Valida-tion	Results	Most significant genes/ proposed biomarkers	GO terms/pathways enriched	Conclusions
Environmental exposures											
Cadmium Dakeshita *et al.* 2009	20 exposed vs. 20 non-exposed (all female)	CC	BMI, present illness, medical history	peripheral blood	80-mer oligo microarray (1867 probes)	spearman rank correlation, multiple regression, ingenuity pathway analysis	real-time PCR	137UP 80DOWN (p<0.05)	CASP9, SLC3A2, GPX3, ITGAL, TNFRSF1B, BCL2A1, COX7B	cell death, cellular growth, proliferation and development, gene expression, cell morphology, protein ubiquitination, glucocorticoid receptor signalling, death receptor signalling, JAK-STAT cascade	Oxidative stress and apoptosis related genes correletate with exposure to Cd
Arsenic Andrew *et al.* 2008	11 (9 male, 2 female) high exposed, 10 (8 male, 2 female) low exposed	CC	smoking	peripheral blood	Affymetrix GeneChip U133 Plus 2.0 oligonucleotide microarray (47,000 probes)	2 class comparison (SAM statistical package), GO-analysis	real-time PCR	259 DIFF EXP (SAM p<0.05)	HSPA9B, CD69, MALT1 (defence); IL2RB, CHST2, NFATC3, ALOX3, PTX3 (immune response)	defence response, immune response, cell growth, signal transduction, apoptosis, regulation of cell cycle, JAK-STAT cascade, T-cell receptor signalling, GTP-ase activity, type-1 diabetes mellitus	Consistent with known health effects of As, genes associated with cell growth, apoptosis, cell cycle, t-cell receptor signalling and diabetes correlate with As-exposure
Arsenic Fry *et al.* 2007	23 high vs. 9 low exposed (all pregnant women)	CC	food consumption, drinking water, health history, birth and pregnancy information	cord blood	Affymetrix GeneChip HGU133 Plus 2.0 oligonucleotide microarray (54.675 probes)	t-test, correlation analysis, linear regression, Ingenuity pathway analysis		404UP 43DOWN	Networks found around NF-kB/IL1, STAT1LHIF-1A, JUN/FOS/IL8	immune response; inflammatory response; response to stress, other organism, pest/pathogen/parasite, wounding, biotic stimulus, external stimulus; cytokine activity; cell death	As could act as an inflammatry stimulus that activates NFkB signalling cascade
	Training set 6 high vs. 6 low exposed, validation set 17 high and 6 low exposed	CP							Biomarker set: CXCL1, DUSP1, EGR-1, IER2, JUNB, MIRN21, OSM, PTGS2, RNF149, SFRS5 and SOC3	stress response, cell cycle regulation	11 Genes were identified that could be a biomarker geneset for As exposure

Combined environmental exposure van Leeuwen *et al.* 2008	398 Flemish adults from 9 regions (aged 50-65, 207 male/191 female	CC	age, sex, smoking status, food consumption, alcohol and tobacco use, medical history, medication use, exposure to solvents, pesticides, …	peripheral blood	real-time PCR analysis of 8 genes (CYP1B1, MAPK14, SOD2, CXCL1, ATF4, TIGD3, DGAT2, PINK1)	ANOVA, pearson correlation		DIFF EXP between regions was observed, as well as many correlations between biomarkers of exposure/effect and expression of the 8 genes			Correlations between gene expression of selected genes, and biomarkers of exposure and effect are found, indicating the potential of using gene exrpession profiling as a biomonitoring tool
Combined environmental exposure De Coster *et al.* in preparation	40 Flemish adults (aged 50-56) – 20 male/20female		See "combined environmental exposre van Leeuwen *et al.* 2008"	peripheral blood	Agilent Whole Human Genome 4x44k oligonucleotide microarray	Correlation analysis, regression analysis, Metacore Pathway analysis		Genes selected for each exposure regression p<0.01, Metacore p<0.05 Male/Female: Cd (6/6), Pb(8/4), PCBs(124/5), dioxin(0/7), ttmA(~benzene:15/9), OH-pyr(~PAHs:10/5)			Correlations between gene expression and biomarkers of exposure are found. Using regression and pathway analysis, genes were selected for PCR analysis in the new Flemish Environment and Health Study (planned 2009-2010). Sexes were analyzed separately, because intersex differences were found while exploring the dataset.
Air pollution van Leeuwen *et al.* 2006	24 children from a mining area (Teplice) and 23 children from a rural area (Prachatice). Teplice: 7 boys and 5 girls age 5-7, 6 boys and 6 girls age 7-11. Prachatice: 6 boys and 6 girls age 6-7, 5 boys and 6 girls age 7-11.	CC		peripheral blood	Agilent Human 22k oligonucleotide microarray (22.000 probes)	t-test, PCA, EASE gene functionality analysis	real-time PCR	1001 genes UP, 726 DOWN (p<0.05)	CXCL1, PINK1, DGAT2 and TIGD3	nucleosome assembly, chromatin assembly/disassembly, microtubule based movement, M-phase associated processes, muscle development, immune response and vitamin metabolism	Relatively small differences in exposure generate large numbers of differentially expressed genes, which makes generating a discriminative profile feasable

Diesel Exhaust Peretz *et al.* 2007	5 healty non smoking men (before vs. after exposure)	CC		peripheral blood	Affymetrix Genechip Human Genome U133 Plus 2.0 (47.000 probes)	t-test, EASE gene functionzlity analysis		188 UP and 176 DOWN 6h after exposure ($p<0.05$), 307 UP and 561 DOWN 22h after exposure ($p<0.05$), 17 UP and 41 DOWNregulated genes in both timepoints ($p<0.05$)		Oxidative stress, inflammation, Leukocytes activation, Cell Adhesion, Cell migration, Vascular homeostasis	Diesel exhaust exposure seems associated with a gene expression signature consisting of genes implicated with
Environmental/ Lifestyle exposures											
Tobacco smoke Lampe *et al.* 2004	smokers (26 male, 16 female), non-smokers (22 male, 21 female)	CC, CP	food consumption, smoking, exercise, demographics	peripheral blood	Hu25k oligonucleotide microarrays (25.000 probes)	Pearson correlation, Montecarlo simulation of reporter genes		861 DIFF EXP	36 genes selected as predictor genes		Smoking is associated with a biologically relevant mRNA expression pattern
Tobacco smoke van Leeuwen *et al.* 2007	9 smoking discordant monozygotic twin pairs (4 male 5 female)	CC		peripheral blood	Phase-1 Human tox 600 cDNA microarrays - toxicologically relevant mechanisms (600 probes)	confidence analysis, wilcoxon test, signal-to-noise ratio	real-time PCR	34/76/44 DIFF EXP (confidence analysis/wilcoxon ranks/singal to noise)	Genes signifficant in all tests: ATF4, MCL1, MAPK14, SERPINA1 PTEN, PXN, SOD2 . Genes correlating (Spearman) significantly with DNA adducts: BAK1, CSF1R, IL10, APP, AXIN1, OGG1, PCK2		Five genes are proposed as a biomarker of effect induced by smoking: SOD2, MAPK14, ATF4, CYP1B1, SERPINB2. The first four of these genes were included in PCR analysis conducted for the Flemish Environment and Health Study (see van Leeuwen *et al.* 2008)

Occupational exposures											
Metal welding fumes Wang *et al.* 2005, Wang *et al.* 2008	15 exposed vs. 7 unexposed (all male)	CC	smoking, medical history, occupational history	whole blood	Affymetrix Human Genome U133A Genechips (39.000 probes)	paired t-test, GO-analysis		533 DIFF EXP pre- vs. postexposure, (86 DIFF EXP in controll group)	IL8, IL1A, CXCR4, RALBP1, SCYE1 (related to immune response)	proinflammatory and immune response, oxidative stress, phosphate metabolism, cell proliferation, apoptosis	Systemic response to metal welding fumes is suggested by the significance of several relevant pathways
Benzene Forrest *et al.* 2005, McHale *et al.* 2009, Smith *et al.* 2005	8 exposed vs. 8 non-exposed (both 4 male/4female)	CC	environmental exposure to solvents and pesticides, past and current alcohol and tobacco use, food consumption, medical history, medication use, familiy history	peripheral blood	Affymetrix Human U133 GeneChip (48.000 probes), Illumina HumanRef-8 Bead-Chips (24.500 probes)	paired t-test, Ingenuity pathway analysis	real-time PCR	Affymetrix: 2692 DIFF EXP ($p<0.05$), Illumina: 1828 DIFF EXP ($p<0.05$)	JUN, ZNF331, CXCL16 and PF4	Apoptosis, Immune response, defence response, stress response, inflammatory response, chromatin assembly	Several biomarker gene exposuree are proposed. Links to effects are made through gene ontology analysis.
Mediacal exposures											
Ionizing radiation Amundson *et al.* 2000, Amundson *et al.* 2004)	1 non hodgkin lymphoma patient (before vs. after radiation)			peripheral blood	6485 element cDNA microarrays	Ease analysis	real-time PCR	82 DIFF EXPR ($p<0.05$ in two or more measurements during the course of the therapy)	CDKN1A, GADD45A, DDB2	heat shock proteins, immune response, inflammatory response	Development of gene expression biomarkers for radiation exposure seems promising

Ionizing radiation Meadows *et al.* 2008	18 healthy controls, 36 pre-irradiated patients, 34 post-irradiated patients, 36 pre-chemotherapy patients, 32 post-chemotherapy patients (mean 47.9 years)	CC, CP	peripheral blood	Operon's human genome oligo set v4.0 (35.035 probes)	2 way mixed model ANOVA	training set used, resulted in the correct identification of 90% of the irradiated patients, and 81% of the chemotherapy patients	Genes that distinguished radiation status:XPC, GTP3A, PCNA, CDKN1A, PPM1D, ACTA2, TIMM8B, MOAP1, DDB2, C19orf2, HNRPDL, BBC3, BAX. Genes that distinguished chemotherapy status: FKBP5, SAP30, SOCS1, CRAMP1L, UVRAG, ASGR1, BLVRA, RAI17, TRAF3, LILRB1, BID, HMOX1, TIEG, NOTCH2, ZFP36L1, IFI30, WARS, CPVL, SCO2	Peripheral blood gene expression profiles can be identified in mice and humans which are specific, accurate over time, and not confounded by inter-individual differences.

Table 2: Overview of characteristics, methods and conclusions of peer reviewed studies investigating the relationship between gene expression and environmental, occupational or medical exposures in human populations. For details, see annex 1. CC=class comparison, CD=class discovery, CP=class prediction.

4.3.2. International environment and health organisations employing gene expression analysis for (potential) regulatory purposes

The **U.S. Environmental Protection Agency** (EPA) has adopted an Interim Policy on the use of genomics data, and plans to use these data in four areas: 1° prioritization of contaminants and contaminated sites to help focus resources on greater hazards or risks, 2° monitoring: chemical, physical analysis of air, water, soil and sediment, 3° toxicity testing, tissue testing, ecological community testing, 3° reporting provisions, 4° risk assessment: mode of action identification, human relevance testing, mixtures testing, … (Dix *et al.* 2006). Currently, genomics data are integrated in the decision making process, but are, on their own, considered insufficient as a basis for decisions (EPA 2004). Several workgroups have been set up to facilitate the transition of genomics data to decision-making processes, with focuses on data submission (toward consistent data submissions), quality assurance (interlaboratory and procedural comparability), data analysis (consistent data analysis approach), data management (storage of collected data), training (prepare EPA for the use of genomics data) (Gallagher 2005).

The **National Institute of Environmental Health Institute** (NIEHS) is involved in the ***Genes, Environment and Health Initiative (GEI)***, that aims to accelerate understanding of genetic and environmental contributions to health and disease. This includes a genetic component that will focus on genetic differences between people with an illness, and healthy individuals. The other component will focus on 1° developing environmental sensors for measuring exposures to toxins, dietary intake, physical activity, psychosocial stress and addictive substances, 2° identifying biomarkers in the human body that indicate activation of disease, and 3° integrating sensor and biomarker technologies.
Another project involving the NIEHS is the ***Epigenomics Initiative***, which will investigate the role of environmental exposures in triggering the epigenetic changes responsible for silencing and overexpressing genes. The objectives are 1° To establish an international committee on standard practices and platforms, develop new antibody reagents and create a reference epigenome database, 2° to develop epigenomic mapping data and infrastructure to facilitate research in human health and disease, 3° to evaluate epigenetic mechanisms in aging, development, environmental exposure (physical, chemical, behavioural, social environments) and modifiers of stress, 4° to develop new technology for single cell analysis and remote imaging of epigenetic activity in cells, tissue and whole animals (NIEHS 2007).

TOXGEN Environment and Health is an initiative from the **European Institute for Health and Consumer protection (IHCP)** with the objective to identify and assess novel methodologies to approach the complexity of environmental health science, including studies on the use of toxicogenomics for the characterization of health effects from exposure to chemicals and chemical mixtures. Attention is given to individual susceptibility factors, multiple chemical exposures, endocrine disruption and blood gene expression in biomonitoring, more specific in studies identifying sensitive individuals and populations (IHCP 2005).

The **Organisation for Economic co-operation and Development (OECD)** has started activities to explore and evaluate regulatory applications of Toxicogenomics and molecular screening assays. This research is focusing on risk assessment of chemicals, developing new biomarkers for exposure and toxicity, and making inventory of omics tools available to several countries. Results of a survey were published in 2008 (See table). Risk evaluation of chemicals and biomarker research is ungoing (OECD 2009).

	2004	2007
Countries	5	8
Total studies	42	62
Transcriptomics	33	32
Proteomics	7	19
Metabolomics	2	11
Publications	10	32

Table 3: Surveys taken by the OECD in 2004 and 2007 about the use of -omics tools by countries. For each year, numbers of countries responding, studies and publications are given.

The **Registration, Evaluation, Authorisation and Restriction of Chemical substances (REACH)** programme of the European Union is also including toxicogenomics, among others in the research around "Classification of carcinogens by toxicogenomics", which should allow for a better classification than traditional methods (Degen & Hengstler 2008).

All of these regulatory driven programs are still ongoing, and at present (november 2009), no concrete results are available.

4.4. Discussion: overall observations regarding these projects

Most studies focus on two groups of individuals, usually a low/non-exposed vs. a (highly-)exposed group, in which gene expression is submitted to class comparison analysis. A few studies (also) include class prediction, involving the development of gene sets to distinguish classes (exposed vs. non-exposed). Some studies also evaluate dose-response relations of exposures and gene expression, using correlation or regression analysis (see table 2).

The ***number*** of studies regarding the effects of environmental exposures on human gene expression in vivo is still very limited.

In addition to the low number of studies, the heterogeneity of exposure conditions (e.g. toxicant studied) in these studies is high. For each of the exposures that have been included, only a few studies are available, which means that ***validation*** of the obtained effects is not yet possible.
Usually little information is available on ***other toxicants to which the subjects have been (environmentally) exposed*****,** apart from the exposure of interest. However, this is especially important in a context of exposure to complex mixtures, like environmental exposures are.

Inclusion of ***effects*** is mostly limited to using gene ontology and pathway analysis, which evaluates genes in their biological context, rather that each gene separately. This way connections between (effect related) pathways and exposures are made. Apart from this, the comparison of gene

expression with ***biomarkers of effect*** in the context of environmental biomonitoring is scarce in these studies. van Leeuwen *et al.* (2007) included DNA aducts in their study of tobacco smoke, and van Leeuwen *et al.* (2008) included several tumor markers and biomarkers of genotoxicity in their study of combined exposures (both conducted for the Flemish Environment and Health Study).

Furthermore, there is rather little attention for several potentially ***important influences*** such as *sex, age, nutrition, time and season of sampling*, ... Since all of these factor can influence the outcome of gene expression profiles, it is important to include these parameters whenever possible. This is discussed in part 5.1.

Exposure	Number of studies	References
Cadmium	2	Dakeshita *et al.* 2009, van Leeuwen *et al.* 2008
Arsenic	2	Andrew *et al.* 2008, Fry *et al.* 2007
Air pollution (general)	1	Van Leeuwen *et al.* 2006
Diesel exhaust	1	Peretz *et al.* 2007
Tobacco smoke	2	Lampe *et al.* 2004, van Leeuwen te al. 2007
Metal welding fumes	1	Wang *et al.* 2005,2008
Benzene	1	Forrest *et al.* 2005, McHale *et al.* 2009, Smith *et al.* 2005, van Leeuwen *et al.* 2008
Ionizing radiation	2	Meadows *et al.* 2008, Amundson *et al.* 2000,2004
Lead, HCB, p,p'-DDE, PCBs, dioxins, PAHs	1	van Leeuwen *et al.* 2008

Table 4: overview of the number of studies, including specific exposures

The ***number of arrays used*** in these studies is however – probably due to the lowering costs – increasing in the more recent studies, which also increases the confidence on their outcomes. Most recent studies also include ***gene ontology*** or ***pathway analysis*** such as Ingenuity or Metacore, and thus includes system biology as opposed to information on the expression of individual genes only. This further increases confidence in the obtained results, as the expression of genes in biological units is compared rather than just analyzing single genes.

In accordance with these observations, the ***present study*** includes - in *part 2 - an investigation of inter- and intra-individual differences* (including sex related differences) in gene expression. In *part 3, differences in gene expression are related to measured internal exposure* of cadmium, lead, PCBs, dioxin, hexachlorobenzene, p,p'-DDE, benzene, and PAHs, *as well as to the effect biomarkers* prostate specific antigen, carcinoembryonic antigen, p53 protein (tumor markers) in serum, the micronucleus test (chromosomal damage), the comet assay (DNA damage) and 8-hydroxydeoxyguanosine (oxidative DNA damage). An extensive database is available with background information on all subjects as well. In total 140 samples of Flemish adults (aged 50-60) will be analyzed: 40 samples in an ongoing study (samples already analyzed) plus 100 additional samples in the present study). Gene expression data will then be related to these biomarkers of exposure and effect. In contrast with most other studies we will both use a large number of samples and a wide range of exposure and effect biomarkers. This range of exposure biomarkers is especially useful in environmental studies such as these, because the population is usually exposed to (low levels of) a number of chemicals.

5. CONSIDERATIONS FOR GENE EXPRESSION USE AS A BIOMARKER

5.1. Aspects requiring further attention and study

Despite the general use of gene expression analysis in a range of scientific disciplines, some important topics have received only limited attention, and should be investigate to further found gene expression studies and biomarker research. Especially in biomonitoring, where it's impossible to control for all covariates, attention is required for these possible causes of confounding.

Personal characteristics

Radich *et al.* (2004) have investigated ***inter-individual variation*** of gene expression in peripheral blood lymphocytes in 17 individuals (9 male, 8 female). The objective was to determine whether stable differences in gene expression between individuals could be identified. Samples were taken on several time points, at least one week apart over a total course of 7 weeks plus one additional follow up around 6 months after initial sampling. In addition, 17 individuals were used for validation (reference pool). 72 genes were identified that differ significantly in expression between individuals, including genes involved in immunologic/inflammatory pathways, genes of the major histocompatibility complex (MHC), and interferon induced genes (CIG5, IFIT1, IFIT4, MX1, USP18). This set of genes allowed to perfectly identify individuals in samples obtained 6 months later. Whitney *et al.* (2003) found significant interindividual differences in 75 persons notably in multi-histocompatibility complex genes, BRCA1 (highly polymorphic), PRNP, and in 6 interferon regulated genes (OAS3, MTAP44, INADL, MX1, GS3686, IFIT1). Eady *et al.* (2005) found inter-individual variation of genes of the major histocompatibility complex, interferon regulated genes (FER1L3, Ly6E, G1P2/IFI15/ISG15, OAS3, IFI44, IFI44L , STAT1), histone genes. Interestingly, interferon-related genes, and MHC class II genes seem to overlap between these three studies. Genes that show considerable inter-individual variation could be of interest when selecting biomarker genes, if they can be linked to specific exposures of effects. However, because such information was not available in this study and sample size was rather low, further investigation of the indentified genes would be required. Obviously, if this variability cannot be traced back to exposures or effects, these genes are unfavourable when selecting biomarker genes.

As to ***intra-individual variation*** contrasting conclusions have been drawn. Eady *et al.* (2005) found an average within subject variation less than 20% for 80% of the genes (6777 out of 8489 genes) analysed, and concluded that within-individual variation is low. Highest intra-individual variation was found in immunoglobulin genes. Whitney *et al.* (2003) did observe intra-individual variation in gene expression, and several highly variable genes, including MHC class II-genes, but stated that this was not the dominant source of variation among samples. Eady *et al.* contribute somewhat differing conclusions concerning intra-individual variation between the two studies to a difference in study design. Further investigation of intra-individual variation is necessary.

Also, the influence of ***sex*** on gene expression was investigated by Whitney *et al.* (2003). Sample population included 40 males and 35 females, all healthy, with an average age of 36.5 years (+/- 14.8 years). A total of 47 genes were found to be differentially expressed (gene expression change > 43%

and False Discovery Rate of less than 6.5 accepted), 11 genes were overexpressed in males, 36 genes were overexpressed in females. Several genes appearing to be differentially expressed between males and females, were located on the X or Y chromosomes, but not exclusively. These genes were associated with amongst others immunity, cell cycle control, proliferation and apoptosis. Eady *et al.* (2005) identified 51 genes differentially expressed in both sexes in a linear regression model including age, sex and BMI. Interestingly, there was a substantial overlap between the sex-specific gene lists identified by the studies of Whitney and Eady. In addition, preliminary results from a whole genome gene expression analysis on a subset of 40 samples of the Flanders Environment and Health Study, where the influence of several environmental pollutants was investigated, indicate a substantial difference in gene expression response to environmental contaminants such as PCB's and lead (unpublished results). This could possibly be associated with influences of these chemicals on endocrine function. However preliminary, this indicated the importance of a well balanced study population, and demands attention to possible differences in gene expression patterns for both sexes.

Whitney *et al.* (2003) have also looked to the influence of ***age*** on gene expression, and report a negative correlation between Ig gene expression and age. The effect of aging on whole genome gene expression is poorly investigated and understood. Additional research is needed.

Lifestyle characteristics

Nutritional factors can influence the outcome of gene expression results. For example, high protein and high carbohydrate breakfasts resulted in 317 respectively 919 differentially expressed genes in a study of van Erk *et al.* (2006). Similarly, 24 and 48h of fasting resulted in 1200 respectively 1386 differentially expressed genes (Bouwens *et al.* 2007). This points to the importance of accounting for nutritional factors when biomonitoring using gene expression.

Because the similarities to environmental pollutants, the influence of ***smoking*** on gene expression was already discussed in chapter 5. Since the effect of smoking is clearly affecting gene expression, smokers should be excluded, or be well balanced in the study design.

Sampling characteristics

There is some evidence that ***time of sampling*** has an influence on gene expression, and that some genes are characterised by a diurnal change in expression. These processes could be associated with circadian cycles of cell growth, or nutrient availability, but are still poorly understood (Whitney *et al.* 2003). Additionally, the season of sampling could affect measured gene expression, bus this has yet to be investigated.

Since microarray gene expression analysis is becoming cheaper, it is to be expected that the number of conducted studies and their sample sizes will continue to increase, and therefore knowledge of these covariates too will expand.

5.2. Considerations for regulatory use of omics-data

Focus of research involving (whole genome) gene expression analysis has mostly been on risk assessment of chemicals, over biomonitoring, especially for humans (Boverhof & Zacharewski 2006). Therefore, the use of toxicogenomics for biomarker development and biomonitoring, is still in an early stage of development (Steinberg *et al.* 2008). As studies accumulate confidence will possibly raise in the obtained biomonitoring results, which should eventually lead to gene expression biomarkers – similarly to what clinical and pharmaceutical studies have proven.

An important issue is ***interdisciplinary cooperation*** and ***the public availability*** of gene expression (microarray) data. As is discussed before, gene expression analysis is used in a wide range of scientific disciplines: medical sciences, pharmacology, human and ecotoxicology, amongst others. Findings from these studies can have inter-disciplinary implications, so transfer of knowledge is essential, also for governments and regulatory bodies, who want to use an integrated approach. In practice, such inter-disciplinary exchange is no obvious. However, efforts are being made to (partially) overcome these problems. One is the development of microarray data standards such as MIAME (minimum information about a microarray experiment) that defines the minimum quantity and quality of information that is required to interpret and verify results (Steinberg *et al.* 2008). Also, several online-databases have been in use for online microarray data submission, to make these data publicly available. Examples are CEBS (Chemical Effects in Biological Systems – NIEHS - http://cebs.niehs.nih.gov), GEO (Gene Expression Omnibus – NCBI - www.ncbi.nlm.nih.gov/geo), ArrayExpress (EBI - http://www.ebi.ac.uk/microarray-as/ae/) and CTD (Comparative Toxicogenomics Database - http://ctd.mdibl.org). Associations between gene expression and exposures, and between gene expression and disease can be retrieved in these databases, and compared to own results.

The development of these new technologies, and the use of gene expression in biomonitoring and as a biomarker generates opportunities and challenges for governments and regulatory bodies. Currently, as is discussed in §4.2, governments are preparing themselves for the use of toxicogenomics data through several organizations (EPA, NIEHS, IHCP, …). EPA had stated that these data are currently useful in the decision making process, but are alone considered insufficient as a basis for decisions, and have to be integrated with more traditional data-sources. However, preparations are being made that may lead to the use of toxicogenomics data in decision-making (Boverhof & Zacharewski 2006).

CONCLUSIONS: OPPORTUNITIES AND CHALLENGES IN THE USE OF GENE EXPRESSION IN ENVIRONMENT AND HEALTH BIOMONITORING

Throughout the different chapters, we have come across opportunities that exist and challenges that have to be overcome or have to be taken into account when using gene expression as a biomarker.

The idea of using gene expression as a biomarker (in environmental biomonitoring) stems from the observation that (certain) exposures and effects are reflected by changes in gene expression (§1.2). Furthermore, gene expression analysis has proven useful in the identification mode of action of toxicants, in dealing with interspecies variability, low dose extrapolation (due to the often higher sensitivity), development of specific fingerprints to distinguish between exposures or effects, and in research of multiple exposures (§1.2).

The use of gene ontology and pathway analysis has taken gene expression analysis from analysis of single genes to analysis of networks of genes, increasing confidence in the output. The system biology-, or integrated approach (involving transcriptomics, proteomics and metabolomics) further increases opportunities for biomarker development, however expensive and still suffering from several difficulties (§2.2).

However, there are still challenges for the actual use of gene expression as a biomarkers (e.g. for regulatory purposes).

An important aspect is the distinction between homeostatic processes or adaptation, and actual adverse effects. Therefore, at this moment, it is still necessary to integrate other markers of effect into biomarker research involving gene expression. Similarly, it's not always clear whether a change in gene expression is a biomarker of exposure or a biomarker of effect (or both). In any case, this change in gene expression may be useful, but different in purpose (§1.2).

Several important factors potentially influencing gene expression outcome are still poorly investigated. These include personal characteristics (inter- and intra-individual variation, sex, age), lifestyle (nutrition, smoking) and sampling characteristic (time of day, season, …). These factors have to be taken into account and require additional research (§5.1).

Factors that influence the possibility (and timing) to use gene expression information in regulatory context include interdisciplinary cooperation and the public availability of data (knowledge transfer). Several initiatives have been developed to accommodate these needs (§5.3).

BIBLIOGRAPHY

Aardema, M. J. & MacGregor, J. T. (2002). Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. Mutat. Res. 499, 13-25.

Amundson, S. A., Do, K. T., Shahab, S., Bittner, M., Meltzer, P., Trent, J., & Fornace, A. J., Jr. (2000). Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. Radiat. Res. 154, 342-346.

Amundson, S. A., Grace, M. B., McLeland, C. B., Epperly, M. W., Yeager, A., Zhan, Q., Greenberger, J. S., & Fornace, A. J., Jr. (2004). Human in vivo radiation-induced biomarkers: gene expression changes in radiotherapy patients. Cancer Res. 64, 6368-6371.

Anderson, N. L. & Anderson, N. G. (2002). The human plasma proteome: history, character, and diagnostic prospects. Mol. Cell Proteomics. 1, 845-867.

Andrew, A. S., Jewell, D. A., Mason, R. A., Whitfield, M. L., Moore, J. H., & Karagas, M. R. (2008). Drinking-water arsenic exposure modulates gene expression in human lymphocytes from a U.S. population. Environ. Health Perspect. 116, 524-531.

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., & Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res. 33, 5914-5923.

Barrett, J. C. & Kawasaki, E. S. (2003). Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov. Today 8, 134-141.

Bouwens, M., Afman, L. A., & Muller, M. (2007). Fasting induces changes in peripheral blood mononuclear cell gene expression profiles related to increases in fatty acid beta-oxidation: functional role of peroxisome proliferator activated receptor alpha in human peripheral blood mononuclear cells. Am. J. Clin. Nutr. 86, 1515-1523.

Boverhof, D. R. & Zacharewski, T. R. (2006). Toxicogenomics in risk assessment: applications and needs. Toxicol. Sci. 89, 352-360.

Cheng, R. Y., Zhao, A., Alvord, W. G., Powell, D. A., Bare, R. M., Masuda, A., Takahashi, T., Anderson, L. M., & Kasprzak, K. S. (2003). Gene expression dose-response changes in microarrays after exposure of human peripheral lung epithelial cells to nickel(II). Toxicol. Appl. Pharmacol. 191, 22-39.

Clarke, C. J. & Haselden, J. N. (2008). Metabolic profiling as a tool for understanding mechanisms of toxicity. Toxicol. Pathol. 36, 140-147.

Collings, F. B. & Vaidya, V. S. (2008). Novel technologies for the discovery and quantitation of biomarkers of toxicity. Toxicology 245, 167-174.

CTD (2009). Comparative Toxicogenomics Database http://ctd.mdibl.org.

Dakeshita, S., Kawai, T., Uemura, H., Hiyoshi, M., Oguma, E., Horiguchi, H., Kayama, F., Aoshima, K., Shirahama, S., Rokutan, K., & Arisawa, K. (2009). Gene expression signatures in peripheral blood cells from Japanese women exposed to environmental cadmium. Toxicology 257, 25-32.

David, D. C., Hoerndli, F., & Gotz, J. (2005). Functional Genomics meets neurodegenerative disorders Part I: transcriptomic and proteomic technology. Prog. Neurobiol. 76, 153-168.

Degen, G. H. & Hengstler, J. G. (2008). Developments in industrial and occupational toxicology: REACH, toxicogenomics, mycotoxins, lead, asbestos, boron, bitumen, deletions polymorphisms and SNP interactions: meeting report of the 16th EUROTOX training and discussion session. Arch. Toxicol. 82, 483-487.

Dix, D. J., Gallagher, K., Benson, W. H., Groskinsky, B. L., McClintock, J. T., Dearfield, K. L., & Farland, W. H. (2006). A framework for the use of genomics data at the EPA. Nat. Biotechnol. 24, 1108-1111.

Eady, J. J., Wortley, G. M., Wormstone, Y. M., Hughes, J. C., Astley, S. B., Foxall, R. J., Doleman, J. F., & Elliott, R. M. (2005). Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. Physiol Genomics 22, 402-411.

EPA (2004). Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA. http://ihcp. jrc. ec. europa. eu/docs/IHCP_annual_report/ihcpar05. pdf.

Fan, T. W., Higashi, R. M., & Lane, A. N. (2006). Integrating metabolomics and transcriptomics for probing SE anticancer mechanisms. Drug Metab Rev. 38, 707-732.

Forrest, M. S., Lan, Q., Hubbard, A. E., Zhang, L., Vermeulen, R., Zhao, X., Li, G., Wu, Y. Y., Shen, M., Yin, S., Chanock, S. J., Rothman, N., & Smith, M. T. (2005). Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. Environ. Health Perspect. 113, 801-807.

Fortina, P. & Surrey, S. (2008). Digital mRNA profiling. Nat. Biotechnol. 26, 293-294.

Frotschl, R., Weickardt, S., Staszewski, S., Kaufmann, G., & Kasper, P. (2005). Effects of chlorpromazine with and without UV irradiation on gene expression of HepG2 cells. Mutat. Res. 575, 47-60.

Fry, R. C., Navasumrit, P., Valiathan, C., Svensson, J. P., Hogan, B. J., Luo, M., Bhattacharya, S., Kandjanapa, K., Soontararuks, S., Nookabkaew, S., Mahidol, C., Ruchirawat, M., & Samson, L. D. (2007). Activation of inflammation/NF-kappaB signaling in infants born to arsenic-exposed mothers. PLoS. Genet. 3, e207.

Gallagher, K. (2005). Potential Implications of Genomics for Regulatory and Risk Assessment Activities at EPA. http://dels. nas. edu/emergingissues/docs/gallagher. pdf.

Genelogic (2009). Genelogic http://www.genelogic.com.

Gundert-Remy, U. (1999). Strong bones in later life: luxury or necessity? The problem of reimbursement. Bull. World Health Organ 77, 434-435.

Gundert-Remy, U., Dahl, S. G., Boobis, A., Kremers, P., Kopp-Schneider, A., Oberemm, A., Renwick, A., & Pelkonen, O. (2005). Molecular approaches to the identification of biomarkers of exposure and effect--report of an expert meeting organized by COST Action B15. November 28, 2003. Toxicol. Lett. 156, 227-240.

Gusnanto, A., Calza, S., & Pawitan, Y. (2007). Identification of differentially expressed genes and false discovery rate in microarray studies. Curr. Opin. Lipidol. 18, 187-193.

Hu, S., Li, Y., Wang, J., Xie, Y., Tjon, K., Wolinsky, L., Loo, R. R., Loo, J. A., & Wong, D. T. (2006). Human saliva proteome and transcriptome. J. Dent. Res. 85, 1129-1133.

IHCP (2005). IHCP TOXGEN Environment and Health. http://ihcp. jrc. ec. europa. eu/docs/IHCP_annual_report/ihcpar05. pdf.

Kerr, M. K. & Churchill, G. A. (2001). Experimental design for gene expression microarrays. Biostatistics. 2, 183-201.

Komichi, D., Tazuma, S., Nishioka, T., Hyogo, H., & Chayama, K. (2005). Glycochenodeoxycholate plays a carcinogenic role in immortalized mouse cholangiocytes via oxidative DNA damage. Free Radic. Biol. Med. 39, 1418-1427.

Kussmann, M., Rezzi, S., & Daniel, H. (2008). Profiling techniques in nutrition and health research. Curr. Opin. Biotechnol. 19, 83-99.

Lampe, J. W., Stepaniants, S. B., Mao, M., Radich, J. P., Dai, H., Linsley, P. S., Friend, S. H., & Potter, J. D. (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol. Biomarkers Prev. 13, 445-453.

Lewis, R. W., Billington, R., Debryune, E., Gamer, A., Lang, B., & Carpanini, F. (2002). Recognition of adverse and nonadverse effects in toxicity studies. Toxicol. Pathol. 30, 66-74.

Li, Y., St John, M. A., Zhou, X., Kim, Y., Sinha, U., Jordan, R. C., Eisele, D., Abemayor, E., Elashoff, D., Park, N. H., & Wong, D. T. (2004a). Salivary transcriptome diagnostics for oral cancer detection. Clin. Cancer Res. 10, 8442-8450.

Li, Y., Zhou, X., St, J. M., & Wong, D. T. (2004b). RNA profiling of cell-free saliva using microarray technology. J. Dent. Res. 83, 199-203.

Liew, C. C., Ma, J., Tang, H. C., Zheng, R., & Dempsey, A. A. (2006). The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J. Lab Clin. Med. 147, 126-132.

Manno, M., Viau, C., Cocker, J., Colosio, C., Lowry, L., Mutti, A., Nordberg, M., & Wang, S. (2009). Biomonitoring for occupational health risk assessment (BOHRA). Toxicol. Lett.

McHale, C. M., Zhang, L., Lan, Q., Li, G., Hubbard, A. E., Forrest, M. S., Vermeulen, R., Chen, J., Shen, M., Rappaport, S. M., Yin, S., Smith, M. T., & Rothman, N. (2009). Changes in the peripheral blood transcriptome associated with occupational benzene exposure identified by cross-comparison on two microarray platforms. Genomics 93, 343-349.

Meadows, S. K., Dressman, H. K., Muramoto, G. G., Himburg, H., Salter, A., Wei, Z., Ginsburg, G. S., Chao, N. J., Nevins, J. R., & Chute, J. P. (2008). Gene expression signatures of radiation response are specific, durable and accurate in mice and humans. PLoS. One. 3, e1912.

Mortensen, A. S., Tolfsen, C. C., & Arukwe, A. (2006). Gene expression patterns in estrogen (nonylphenol) and aryl hydrocarbon receptor agonists (PCB-77) interaction using rainbow trout (Oncorhynchus Mykiss) primary hepatocyte culture. J. Toxicol. Environ. Health A 69, 1-19.

Motoyama, A. & Yates, J. R., III (2008). Multidimensional LC separations in shotgun proteomics. Anal. Chem. 80, 7187-7193.

Newman, J. P., Banerjee, B., Fang, W., Poonepalli, A., Balakrishnan, L., Low, G. K., Bhattacharjee, R. N., Akira, S., Jayapal, M., Melendez, A. J., Baskar, R., Lee, H. W., & Hande, M. P. (2008). Short dysfunctional telomeres impair the repair of arsenite-induced oxidative damage in mouse cells. J. Cell Physiol 214, 796-809.

NIEHS (2007). http://www. niehs. nih. gov/news/newsletter/2007/october/trans. cfm.

Oberemm, A., Onyon, L., & Gundert-Remy, U. (2005). How can toxicogenomics inform risk assessment? Toxicol. Appl. Pharmacol. 207, 592-598.

OECD (2009). OECD Activities to Explore and Evaluate Regulatory Application of Toxicogenomics and Molecular Screening Assays. http://www. oecd. org/document/29/0,3343,en_2649_34377_34704669_1_1_1_1,00. html.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T. M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., Walker, S. J., Zhang, L., Hurban, P., de, L. F., Fuscoe, J. C., Tong, W., Shi, L., & Wolfinger, R. D. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat. Biotechnol. 24, 1140-1150.

Pennie, W. D., Woodyatt, N. J., Aldridge, T. C., & Orphanides, G. (2001). Application of genomics to the definition of the molecular basis for toxicity. Toxicol. Lett. 120, 353-358.

Peretz, A., Peck, E. C., Bammler, T. K., Beyer, R. P., Sullivan, J. H., Trenga, C. A., Srinouanprachnah, S., Farin, F. M., & Kaufman, J. D. (2007). Diesel exhaust inhalation and assessment of peripheral blood mononuclear cell gene transcription effects: an exploratory study of healthy human volunteers. Inhal. Toxicol. 19, 1107-1119.

Pounds, S. B. (2006). Estimation and control of multiple testing error rates for microarray studies. Brief. Bioinform. 7, 25-36.

Radich, J. P., Mao, M., Stepaniants, S., Biery, M., Castle, J., Ward, T., Schimmack, G., Kobayashi, S., Carleton, M., Lampe, J., & Linsley, P. S. (2004). Individual-specific variation of gene expression in peripheral blood leukocytes. Genomics 83, 980-988.

Rockett, J. C., Burczynski, M. E., Fornace, A. J., Herrmann, P. C., Krawetz, S. A., & Dix, D. J. (2004). Surrogate tissue analysis: monitoring toxicant exposure and health status of

inaccessible tissues through the analysis of accessible tissues and cells. Toxicol. Appl. Pharmacol. 194, 189-199.

Sano, Y., Nakashima, H., Yoshioka, N., Etho, N., Nomiyama, T., Nishiwaki, Y., Takebayashi, T., & Oame, K. (2009). Trichloroethylene liver toxicity in mouse and rat: microarray analysis reveals species differences in gene expression. Arch. Toxicol. 83, 835-849.

Shen, S., Lee, J., Weinfeld, M., & Le, X. C. (2008). Attenuation of DNA damage-induced p53 expression by arsenic: a possible mechanism for arsenic co-carcinogenesis. Mol. Carcinog. 47, 508-518.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135-1145.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de, L. F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T. M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X. H., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q. Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., & Slikker, W., Jr. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 24, 1151-1161.

Smith, M. T., Vermeulen, R., Li, G., Zhang, L., Lan, Q., Hubbard, A. E., Forrest, M. S., McHale, C., Zhao, X., Gunn, L., Shen, M., Rappaport, S. M., Yin, S., Chanock, S., & Rothman, N. (2005). Use of 'Omic' technologies to study humans exposed to benzene. Chem. Biol. Interact. 153-154, 123-127.

Steinberg, C. E., Sturzenbaum, S. R., & Menzel, R. (2008). Genes and environment - striking the fine balance between sophisticated biomonitoring and true functional environmental genomics. Sci. Total Environ. 400, 142-161.

Sullivan, P. F., Fan, C., & Perou, C. M. (2006). Evaluating the comparability of gene expression in blood and brain. Am. J. Med. Genet. B Neuropsychiatr. Genet. 141B, 261-268.

van Delft, J. H., van, A. E., van Breda, S. G., Herwijnen, M. H., Staal, Y. C., & Kleinjans, J. C. (2005). Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling. Mutat. Res. 575, 17-33.

van Erk, M. J., Blom, W. A., van, O. B., & Hendriks, H. F. (2006). High-protein and high-carbohydrate breakfasts differentially change the transcriptome of human blood cells. Am. J. Clin. Nutr. 84, 1233-1241.

van Leeuwen, D. M., Gottschalk, R. W., Schoeters, G., van Larebeke, N. A., Nelen, V., Baeyens, W. F., Kleinjans, J. C., & van Delft, J. H. (2008). Transcriptome analysis in peripheral blood of humans exposed to environmental carcinogens: a promising new biomarker in environmental health studies. Environ. Health Perspect. 116, 1519-1525.

van Leeuwen, D. M., van Herwijnen, M. H., Pedersen, M., Knudsen, L. E., Kirsch-Volders, M., Sram, R. J., Staal, Y. C., Bajak, E., van Delft, J. H., & Kleinjans, J. C. (2006). Genome-wide differential gene expression in children exposed to air pollution in the Czech Republic. Mutat. Res. 600, 12-22.

van Leeuwen, D. M., van, A. E., Gottschalk, R. W., Vlietinck, R., Gielen, M., van Herwijnen, M. H., Maas, L. M., Kleinjans, J. C., & van Delft, J. H. (2007). Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. Carcinogenesis 28, 691-697.

Veerla, S. & Hoglund, M. (2006). Analysis of promoter regions of co-expressed genes identified by microarray analysis. BMC. Bioinformatics. 7, 384.

Wang, Z., Neuberg, D., Su, L., Kim, J. Y., Chen, J. C., & Christiani, D. C. (2008). Prospective study of metal fume-induced responses of global gene expression profiling in whole blood. Inhal. Toxicol. 20, 1233-1244.

Wang, Z., Neuburg, D., Li, C., Su, L., Kim, J. Y., Chen, J. C., & Christiani, D. C. (2005). Global gene expression profiling in whole-blood samples from individuals exposed to metal fumes. Environ. Health Perspect. 113, 233-241.

Werner, T. (2008). Bioinformatics applications for pathway analysis of microarray data. Curr. Opin. Biotechnol. 19, 50-54.

Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J. C., Relman, D. A., & Brown, P. O. (2003). Individuality and variation in gene expression patterns in human blood. Proc. Natl. Acad. Sci. U. S. A 100, 1896-1901.

Wingren, C. & Borrebaeck, C. A. (2004). High-throughput proteomics using antibody microarrays. Expert. Rev. Proteomics. 1, 355-364.

Woods, C. G., Heuvel, J. P., & Rusyn, I. (2007). Genomic profiling in nuclear receptor-mediated toxicity. Toxicol. Pathol. 35, 474-494.

Yang, Y. H. & Speed, T. (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. 3, 579-588.

ANNEX1:

LIST OF PEER REVIEWED GENE EXPRESSION STUDIES LISTED IN TABLE 2

A. Environmental exposures

Cadmium (Dakeshita *et al.* 2009)

Dakeshita S, Kawai T, Uemura H, Hiyoshi M, Oguma E, Horiguchi H, Kayama F, Aoshima K, Shirahama S, Rokutan K, Arisawa K. ***Gene expression signatures in peripheral blood cells from Japanese women exposed to environmental cadmium.*** *Toxicology. 2009 Mar 4;257(1-2):25-32.*

Materials and methods:

- Population: 20 females in a Cd-polluted area, 20 females in a non-Cd polluted area, individually matched for age
- Medium: peripheral blood
- Internal exposure measurement: blood cadmium (exposed: mean 3.55µg/L – range: 1.9-6.5µg/L, non-exposed: mean 1.5µg/L – range: 0.6-2.2µg/L), urinary cadmium (exposed: mean 8.25µg/g crt – range: 3.4-22.2µg/g crt, non-exposed: mean 4.35µg/g crt – range 1.9-26.5 µg/g crt)
- Confounders: BMI, present illness, medical history
- 80-mer oligoDNA microarray with 1867 genes (stress response and drug metabolism related)
- Statistical analysis: spearman rank correlation, multiple regression, Ingenuity pathway analysis
- Validation: RT-PCR on selected genes (>30% up or downregulated, correlation coefficients,

Results:

- 137 genes UP, 80 DOWN-regulated in the high exposed group ($p<0.05$)
- Genes related to apoptosis, cell cycle regulation, oncogenes, stress response, immune system, transporters, metabolism
- Significantly enriched annotations: cell death; cellular growth, proliferation and development; gene expression; cell morphology; protein ubiquitination; glucocorticoid receptor signalling, death receptor signalling, JAK-STAT cascade
- Most significantly differentially expressed genes related to stress response (CASP9, SLC3A2, GPX3, ITGAL, TNFRSF1B, BCL2A1, COX7B

Conclusions:

- Several genes were identified as correlating with low levels of Cd-exposure, predominantly oxidative stress and apoptosis related.
- Whole genome gene expression analysis of more heavily exposed individuals is currently underway.

Arsenic (Andrew *et al.* 2008, Fry *et al.* 2007)

Andrew AS, Jewell DA, Mason RA, Whitfield ML, Moore JH, Karagas MR. ***Drinking-water arsenic exposure modulates gene expression in human lymphocytes from a U.S. population.*** *Environ Health Perspect. 2008 Apr;116(4):524-31.*

Materials and methods:

- Population: 11 (9 male, 2 female, mean 66 years) high exposed (>5µg As /l urine), 10 (8 male, 2 female, mean 67 years) low exposed individuals
- Medium: peripheral blood lymphocytes
- Internal exposure measurement: urinary arsenic, drinking water arsenic: high-exposed mean 32µg/L (range 10.4-74.7µg/L) low exposed mean 0.7µg/L (range 0.007-5.3µg/L)
- Confounders: smoking (high exposed: 27% smokers, low exposed:10%)
- Affymetrix GeneChip U133 Plus 2.0 oligonucleotide microarray (47,000 probes)
- Statistical analysis: 2 class comparison (SAM statistical package), GO-analysis
- Validation: RT-PCR of selected genes

Results:

- 259 genes differentially expressed between high and low exposed groups (SAM test, $p<0.05$)
- Significantly enriched annotations: defence response, immune response, cell growth, signal transduction, apoptosis, regulation of cell cycle, JAK-STAT cascade, T-cell receptor signalling, GTP-ase activity, type-1 diabetes mellitus.
- Most significantly differentially expressed genes related to defence and immune response: HSPA9B, CD69, MALT1 resp. IL2RB CHST2, NFATC3, ALOX3, PTX3
- Genes analysed by RT-PCR correlating to urinary As: IL2RB, PRF1 and KIR3DL1 (UP), HLA-DRB1 (DOWN)

Conclusions:

- Several genes were identified as correlating with exposure to arsenic, predominantly defence and immune response associated, as well as involved in cell growth, apoptosis, cell cycle regulation, T-cell receptor signalling and diabetes, which is consistent with documented health effects of arsenic.
- Genes and pathways identified are candidates for biomonitoring of individuals with history of arsenic exposure

Fry RC, Navasumrit P, Valiathan C, Svensson JP, Hogan BJ, Luo M, Bhattacharya S, Kandjanapa K, Soontararuks S, Nookabkaew S, Mahidol C, Ruchirawat M, Samson LD. ***Activation of inflammation/NF-kappaB signalling in infants born to arsenic-exposed mothers.*** *PLoS Genet. 2007 Nov;3(11):e207.*

Materials and methods:

- Population: 23 pregnant women from the Ron Pibul District (high As-exposed) and 9 women from the Bangkok Area (Thailand) – age, education, and socioeconomically matched
- Medium: cord blood
- Internal exposure measurement: toenail arsenic concentration, high exposed >= 0.5µg/g (range 0.1-0.38), low exposed <0.5µg/g (range 0.5-68.63µg/g)
- Confounders: food consumption, drinking water, health history, birth and pregnancy information
- Affymetrix GeneChip HGU133 Plus 2.0 oligonucleotide microarray (54.675 probes)
- Statistical analysis: t-test, correlation analysis, linear regression, Ingenuity pathway analysis

- Validation: /

Results:

- Training/validation sample sets
  - Three training-sets of samples were selected based on the following criteria: 1° random selection, 2° 6 highest and 6 lowest exposed individuals, 3° a combination of 1° and 2°. These training sets led to a fingerprint of respectively 170, 38 and 11 genes. In all three cases these training sets lead to an accurate prediction of 80% of the validation samples (high or low exposed group).
  - The 11 genes identified as the strongest biomarker set for arsenic exposure (least number of genes needed to predict 80% of the samples) are: CXCL1, DUSP1, EGR-1, IER2, JUNB, MIRN21, OSM, PTGS2, RNF149, SFRS5 and SOC3.
  - Significantly enriched annotations (11 biomarker genes): stress response, cell cycle regulation
- Differential expression high vs. low exposed groups
  - 404 genes UP, 43 genes DOWN-regulated in the high exposed group ($p<0.05$)
  - Significantly enriched annotations: immune response; inflammatory response; response to stress, other organism, pest/pathogen/parasite, wounding, biotic stimulus, external stimulus; cytokine activity; cell death.
  - Network analysis identified three networks: 1° network around NF-kB/IL1-B (apoptosis/inflammation related), 2° network around STAT1/HIF-1a (cytokine response related), 3° network around JUN/FOS/IL8 (stress response related)
  - Most significantly differentially expressed genes:

Conclusions:

- A set of 11 genes was identified that could be a biomarker geneset for prenatal arsenic-exposure.
- Prenatal exposure to arsenic could act as an inflammatory stimulus that activates the NF-kB signalling cascade.
- Differential gene expression observed in this study could possibly in part be explained by other environmental exposures (like other heavy metals), however no information on other exposures is available

Air pollution (van Leeuwen *et al.* 2006)

van Leeuwen DM, van Herwijnen MH, Pedersen M, Knudsen LE, Kirsch-Volders M, Sram RJ, Staal YC, Bajak E, van Delft JH, Kleinjans JC. ***Genome-wide differential gene expression in children exposed to air pollution in the Czech Republic.*** *Mutat Res. 2006 Aug 30;600(1-2):12-22.*

Materials and methods:

- Population: Czech 24 children from a mining area (Teplice) and 23 children from a rural area (Prachatice). Teplice: 7 boys and 5 girls age 5-7, 6 boys and 6 girls age 7-11. Prachatice: 6 boys and 6 girls age 6-7, 5 boys and 6 girls age 7-11.
- Medium: peripheral blood cells

- Exposure measurement: regional air pollution (SO2, NOx, NO2, NO, CO, PM2.5, PM10, PAHs, c-PAHs)

- Confounders:
- Agilent Human 22k oligonucleotide microarray (22.000 probes)
- Statistical analysis: t-test, PCA, EASE gene functionality analysis
- Validation: real-time PCR

Results:

- 1001 genes UP, 726 DOWN-regulated in the high-exposed group ($p<0.05$)
- Significantly enriched annotations: nucleosome assembly, chromatin assembly/disassembly, microtubule based movement, M-phase associated processes, muscle development, immune response and vitamin metabolism
- Genes correlating with micronuclei frequencies: 747 positively, 705 negatively
- Significantly enriched annotations (genes correlation with micronuclei): mRNA metabolism, cell communication, RNA processing
- Real-time PCR validation was performed on 8 genes: 3 genes were not abundant enough to be measured, results of CXCL1, PINK1, DGAT2 and TIGD3 were confirmed by real-time PCR.

Conclusions:

- Due to the large amounts of genes that responded to rather low differences in exposure between the two populations, it appears feasible to generate discriminative profiles.
- Transcriptomic analysis is found to be a promising tool for monitoring adverse health effects due to environmental exposure.

Diesel exhaust (Peretz *et al.* 2007)

Peretz A, Peck EC, Bammler TK, Beyer RP, Sullivan JH, Trenga CA, Srinouanprachnah S, Farin FM, Kaufman JD. ***Diesel exhaust inhalation and assessment of peripheral blood mononuclear cell gene transcription effects: an exploratory study of healthy human volunteers.*** *Inhal Toxicol. 2007 Nov;19(14):1107-19.*

Materials and methods:

- Population: 5 healthy, non-smoking men
- Medium: peripheral blood mononuclear cells
- Exposure measurement: 2h exposure sessions to PM2.5, NO2, NO and CO
- Microarray: Affymetrix Genechip Human Genome U133 Plus 2.0 (47.000 probes)
- Validation: /

Results:

- 188 UP and 176 DOWNregulated genes 6h after exposure ($p<0.05$), 307 UP and 561 DOWNregulated genes 22h after exposure ($p<0.05$), 17 UP and 41 DOWNregulated genes in both timepoints ($p<0.05$)

- Significantly enriched annotations: Oxidative stress, inflammation, Leukocytes activation, Cell Adhesion, Cell migration, Vascular homeostasis

Conclusions:

- It is suggested that diesel exhaust exposure is associated with a characteristic gene expression signature, and genes part of this signature have been implicated in oxidative stress and inflammatory processes.
- Small sample size limits conclusions. Future studies will be carried out with a larger sample size

Tobacco smoke (Lampe *et al.* 2004, van Leeuwen *et al.* 2007)

Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, Friend SH, Potter JD. ***Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke.*** *Cancer Epidemiol Biomarkers Prev. 2004 Mar;13(3):445-53.*

Materials and methods:

- Population: smokers (26 men, 16 women), non-smokers (22 men, 21 women),
- Medium: peripheral blood leukocytes
- Internal exposure measurement: plasma cotinine
- Plasma cotinine levels smokers: men mean 217ng/mL (SD:106ng/mL) women mean 241ng/mL (SD: 85ng/mL)
- Hu25k oligonucleotide microarrays (25.000 probes)
- Statistical analysis: Pearson correlation
- Validation: /

Results:

- 861 genes differentially expressed
- 36 genes correlating with plasma cotinine were selected with lowest type1+type2 error rates (25 directly correlated, 9 inversely correlated).

Conclusions:

- Active exposure to tobacco smoke is associated with a biologically relevant mRNA expression pattern
- Knowledge on other exposures and variables and its influence on gene expression could help in studying exposure-effect relations in human population.

van Leeuwen DM, van Agen E, Gottschalk RW, Vlietinck R, Gielen M, van Herwijnen MH, Maas LM, Kleinjans JC, van Delft JH. ***Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs.*** *Carcinogenesis. 2007 28 (3):691-697*

Materials and methods:

- Population: 9 smoking-discordant monozygotic twin pairs (4 male and 5 female)
- Medium: peripheral blood mononuclear cells
- Internal exposure measurement: plasma cotinine

- Plasma cotinine smokers mean 247.6 ng/mL (SD: 54.7), non-smokers mean 5.6 ng/mL (SD: 0,2)
- Effect markers: DNA adducts
- Phase-1 Human tox 600 cDNA microarrays (Phase-1 Molecular Toxicology, Santa Fe, NM, USA) toxicologically relevant mechanisms (600 probes)
- Statistical analysis: confidence analysis, wilcoxon-test, signal-to-noise ration (smokers vs. non-smokers)
- Validation: real-time PCR

Results:

- Differentially expressed genes: 34 (confidence analysis), 76 (Wilcoxon signed ranks test), 44 (signal to noise ratio test). Genes significant in each test: ATF4, MCL1, MAPK14, SERPINA1 PTEN, PXN, SOD2
- Genes correlating (Spearman) significantly with DNA adducts: BAK1, CSF1R, IL10, APP, AXIN1, OGG1, PCK2
- Results confirmed by real-time PCR for SOD2, MAPK14, ATF4. Newly discovered genes by real-time PCR: CYP1B1, SERPINB2

Conclusions:

- Five genes are proposed to be suitable biomarker of effect induced by smoking: SOD2, MAPK14, ATF4, CYP1B1, SERPINB2. These genes (apart from SERPINB2) have further been used in the Flemish Environment and Health Study, in which they were linked to other markers for exposure and effect (see van Leeuwen *et al.* 2007).

Combined environmental Exposure: cadmium, lead, PCBs, DDE, HCB, dioxins, benzene, PAHs (van Leeuwen *et al.* 2008, De Coster *et al.* in preparation)

van Leeuwen DM, Gottschalk RW, Schoeters G, van Larebeke NA, Nelen V, Baeyens WF, Kleinjans JC, van Delft JH. ***Transcriptome analysis in peripheral blood of humans exposed to environmental carcinogens: a promising new biomarker in environmental health studies****. Environ Health Perspect. 2008 Nov;116(11):1519-25.*

Materials and methods

- Population: 398 adults (age 50-65years, 207male/191female)
- Medium: peripheral blood mononuclear cells
- Internal exposure measurement: cadmium (Cd), lead (Pb), polychlorinated biphenyls (PCBs), dioxins (teq), p,p'-DDE (DDE), hexachlorobenzene (HCB), benzene (ttMA), PAHs (OH-pyrene); effect biomarkers: Prostate specific antigen (PSA), carcinoembryonic antigen (CEA, p53 in serum, micronucleustest (MN), comet assay (comet), 8-hydroxydeoxyguanosine (8HDG).
- Real-time PCR analysis of 8 genes: CYP1B1, MAPK14, SOD2, CXCL1, ATF4, PINK1, DGAT2, TIGD3
- Statistical analysis: correlation analysis
- Confounders: age, sex, smoking status, food consumption, alcohol and tobacco use, medical history, medication use, exposure to solvents, pesticides, …

Results:

- Variation in gene expression was found between several Flemisch areas. Most significant were Olen (most upregulated gene exporession) and Gent/Fruit area (most down-regulated).

- Significant correlations were found for:
    - CYP1B1: ttma (all,F), CEA(F), HCB(M)
    - ATF4:OH-pyr(all)
    - MAPK14: PCBs(all,M), HCB(all,M), DDE(all,M), Cd blood (all), Cd urine (all,M)
    - SOD2: PCBs(all,M), DDE (all,M), Cd urine (all,M), MN (F), CEA (F)
    - CXCL1: teq (all,M), DDE (F), CEA(M)
    - DGAT2: teq(all), p53(all), Pb (F), comet (F), teq(M), PCBs (M), Cd urine (M)
    - TIGD3: PCBs(all), OH-pyr (all),
    - PINK1: PCBs(all,M), DDE (all,M), HCB (M)

All=whole population, M=males, F=females

Conclusions:

- Correlations between gene expression of selected genes, and biomarkers of exposure and effect are found, indicating the potential of using gene exrpession profiling as a biomonitoring tool

De Coster et al. in preparation

Materials and methods:

- From the previous study (see van Leeuwen *et al.* 2008), 40 samples (20male/20female) were analysed on
- Platform: Agilent whole human genome 4x44k oligonucleotide microarrays.
- Statistical analysis: Regression analysis, Metacore Pathway analysis
-

Results (preliminary)

- Sexes were analyzed separately, because intersex differences were found while exploring the dataset.
- Based on regression ($p<0.01$) and pathway analysis results ($p<0.05$ in Metacore), a selection of genes was made (Male/Female): Cd (6/6), Pb(8/4), PCBs(124/5), dioxin(0/7), ttmA(~benzene:15/9), OH-pyr(~PAHs:10/5)

Polychlorinated biphenyls (Sisir *et al.* 2007)

Sisir K. Dutta , Somiranjan Ghosh , Eric P. Hoffman , Tomas Trnovec , Lubica Palkovicova , Dean Sonneborn , Irva Hertz-Picciotto. ***Early Disease Biomarkers of PCB-exposed Human Population.*** *Grant Number: 1UO1ES016127-01 (2007)*

Materials and methods: In vitro gene expression patterns in peripheral blood mononuclear cells exposed to PCBs/OH-PCBs have been observed previously. Validation of gene expression biomarkers

from in vitro study in a population study is planned. Two Slovakian districts have been chosen to identify gene expression biomarkers of environmental exposure to PCBs in the population: Michalovce (high PCB contamination due to industrial waste in local rivers) and Svidnik ('background' levels of PCBs). Results are not yet available.

Genotoxic and immunotoxic chemicals (Newgeneris 2009)

NewGeneris-project: Newborns and Genotoxic exposure risks. www.newgeneris.org

NewGeneris is an Integrated Project conducted within the European Union's 6th Framework Program, priority area Food Quality and Safety. Its objective is to investigate the role of prenatal and early-life exposure to genotoxic chemicals present in food and the environment in the development of childhood cancer and immune disorders. Several publications are about to be submitted (results are not yet available) (NewGeneris 2009).

B. Occupational exposures

Metal welding-fumes (Wang *et al.* 2005, Wang *et al.* 2008)

Wang Z, Neuberg D, Su L, Kim JY, Chen JC, Christiani DC. ***Prospective study of metal fume-induced responses of global gene expression profiling in whole blood.*** *Inhal Toxicol. 2008 Nov;20(14):1233-44.*

Wang Z, Neuburg D, Li C, Su L, Kim JY, Chen JC, Christiani DC. ***Global gene expression profiling in whole-blood samples from individuals exposed to metal fumes.*** *Environ Health Perspect. 2005 Feb;113(2):233-41.*

Materials and methods:

- Population: 15 exposed males (in a welding school), 7 unexposed males
- Medium: whole blood
- Exposure measurement: air concentration of metal fumes exposed group: median 2.44 mg pm2.5 /m³ (range 1.3-3.42 mg pm2.5 /m³), non-exposed group: 0.04 mg pm2.5 /m³ (range 0.02-0.17 mg pm2.5 /m³)
- Confounders: smoking
- Affymetrix Human Genome U133A Genechips (39.000 probes)
- Statistical analysis: paired t-test, GO-analysis
- Validation: /

Results:

- 533 genes differentially expressed post exposure compared to pre-exposure ($p<0.05$) (control group: 86 genes differentially expressed in two timepoints)
- Significantly enriched annotations (pre vs. post exposure): proinflammatory and immune response, oxidative stress, phosphate metabolism, cell proliferation, apoptosis
- 35 genes were identified from significant pathways (pre vs. post exposure), among others 5 related to early inflammatory response: IL8, IL1A, CXCR4, RALBP1, SCYE1).

Conclusions:

- Environmental exposure to metal fumes in healthy individuals induced observable changes in gene expression profiles
- Significance of proinflammatory pathways, immune response, oxidative stress, phosphate metabolism, apoptosis, cell proliferation suggests systemic response in peripheral blood in response to environmental particle exposure.
- Smoking is an important confounder

Benzene (Forrest *et al.* 2005, McHale *et al.* 2009, Smith *et al.* 2005)

McHale CM, Zhang L, Lan Q, Li G, Hubbard AE, Forrest MS, Vermeulen R, Chen J, Shen M, Rappaport SM, Yin S, Smith MT, Rothman N. ***Changes in the peripheral blood transcriptome associated with occupational benzene exposure identified by cross-comparison on two microarray platforms.*** *Genomics. 2009 Apr;93(4):343-9.*

Smith MT, Vermeulen R, Li G, Zhang L, Lan Q, Hubbard AE, Forrest MS, McHale C, Zhao X, Gunn L, Shen M, Rappaport SM, Yin S, Chanock S, Rothman N. ***Use of 'Omic' technologies to study humans exposed to benzene.*** *Chem Biol Interact. 2005 May 30;153-154:123-7.*

Forrest MS, Lan Q, Hubbard AE, Zhang L, Vermeulen R, Zhao X, Li G, Wu YY, Shen M, Yin S, Chanock SJ, Rothman N, Smith MT. ***Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers.*** *Environ Health Perspect. 2005 Jun;113(6):801-7.*

Materials and methods:

- Population: 8 high (professionally) exposed (4male, 4 female, mean age 33.5 years) and 8 unexposed (4 male, 4 female, mean age 35.4 years), sex and age-matched.
- Medium: peripheral blood
- Internal exposure measurement: urine benzene levels (exposed mean 778.7 µg/L – SD 1433.02, non-exposed mean 0.36 µg/L – SD 0.51) , air exposure to benzene monitoring
- Affymetrix Human U133 GeneChip (48.000 probes), Illumina HumanRef-8 Bead-Chips (24.500 probes)
- Statistical analysis: paired t-test, Ingenuity pathway analysis
- Validation: real-time PCR in a larger dataset of 28 individuals

Results:

- Affymetrix: 2692 genes differentially expressed ($p<0.05$), 180 UP/ 65 DOWN with FC>1.5
- Illumina: 1828 genes differentially expressed ($p<0.05$), 83 UP/ 88 DOWN with FC>1.5
- 346 genes cross-validated
- Significantly enriched annotations: Apoptosis, Immune response, defence response, stress response, inflammatory response, chromatin assembly
- Significantly enriched canonical pathways: lipid metabolism related
- Genes identified as candidate biomarkers resulting from McHale *et al.* 2009 and Forrest *et al.* 2005: JUN, ZNF331, CXCL16 and PF4

Conclusions:

- Several genes were identified that could be potential biomarkers for benzene exposure

C. Medical exposures

Ionizing Radiation (Amundson *et al.* 2000, Amundson *et al.* 2004, Meadows *et al.* 2008)

Amundson SA, Do KT, Shahab S, Bittner M, Meltzer P, Trent J, Fornace AJ Jr. ***Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation.*** *Radiat Res. 2000 Sep;154(3):342-6.*

Amundson SA, Grace MB, McLeland CB, Epperly MW, Yeager A, Zhan Q, Greenberger JS, Fornace AJ Jr. ***Human in vivo radiation-induced biomarkers: gene expression changes in radiotherapy patients.*** *Cancer Res. 2004 Sep 15;64(18):6368-71.*

Materials and methods:

- Population: one non-Hodgkin lymphoma patient has gene expression changes followed up during radiotherapy (accululated dose: 9Gy), 6 male and 2 female radiotherapy cancer patients exposed to X-rays (1.5Gy)
- Medium: peripheral white blood cells
- Exposure measurement: doses of radiation received during treatment 6485 element cDNA microarrays
- Validation: quantitative RT-PCR

Results:

- Differential expression after treatment in one non-hodgkin patient: 82 genes significantly regulated in two or more measurements during the course of therapy
- Significantly enriched annotations: heat shock proteins, immune response, inflammatory response
- Genes validated by QRT-PCR: CDKN1A, GADD45A, DDB2
- Validated genes were compared by RT-PCR between the 7 other patients

Conclusions:

- Development of gene expression biomarkers for radiation exposure seems promising
- Due to individual differences in response to radiation, a set of biomarker genes would seem more informative than single genes.

Meadows SK, Dressman HK, Muramoto GG, Himburg H, Salter A, Wei Z, Ginsburg GS, Chao NJ, Nevins JR, Chute JP. ***Gene expression signatures of radiation response are specific, durable and accurate in mice and humans.*** *PLoS One. 2008 Apr 2;3(4):e1912*

Materials and methods:

- Population: 18 healthy controls, 36 pre-irradiated patients, 34 post-irradiated patients, 36 pre-chemotherapy patients, 32 post-chemotherapy patients (mean 47.9 years)
- Medium: peripheral blood
- Exposure: irradiated dose (50cGy, 200cGy, 1000cGy)
- Operon's human genome oligo set v4.0 (35.035 probes)
- Statistical analysis: 2 way mixed model ANOVA

- Validation: /

Results:

- A training set used, resulted in the correct identification of 90% of the irradiated patients, and 81% of the chemotherapy patients
- Genes that distinguished radiation status:XPC, GTP3A, PCNA, CDKN1A, PPM1D, ACTA2, TIMM8B, MOAP1, DDB2, C19orf2, HNRPDL, BBC3, BAX
- Genes that distinguished chemotherapy status: FKBP5, SAP30, SOCS1, CRAMP1L, UVRAG, ASGR1, BLVRA, RAI17, TRAF3, LILRB1, BID, HMOX1, TIEG, NOTCH2, ZFP36L1, IFI30, WARS, CPVL, SCO2

Conclusions:

- Peripheral blood gene expression profiles can be identified in mice and humans which are specific, accurate over time, and not confounded by inter-individual differences.

WP2: Normal blood gene expression variability

1. INTRODUCTION

Toxicogenomics is a scientific field that studies how the genome is involved in responses to environmental stressors and toxicants. In its broadest sense toxicogenomics holds promise for identifying biomarkers of exposure with improved sensitivity and selectivity, and in easily accessible biological fluids and tissues (e.g. blood, urine, buccal epithelial cells). The ability to measure hundreds of thousands of genes, proteins, or metabolites from a single sample has been commonly referred to by the suffix 'omics'. The four general categories of omics technologies include genomics, transcriptomics, proteomics and metabolomics. Transcriptomics is the full complement of activated genes that encode mRNA or transcripts in a tissue or a sample. This analysis using microarray technology is now widely used because of its technological maturity, reproducibility and affordability. Gene expression changes constitute of the first responses of the individual when homeostatic conditions are perturbed. It is envisioned that gene expression profiling will enable to identify genes and pathways that are activated and/or suppressed during cell injury and recovery as a result of the challenge of the individual with environmental stressors and toxicants. Hence, the methodology may hold promise to identify biological changes before any clinical manifestation of toxicity. The purpose is to identify sentinel genes or gene clusters that are responsive to exposure to chemical toxicants or environmental pollutants. Historically, single biomarkers have been used for toxicity assessment. The single biomarkers are limited in their ability to inform on mechanism, but variability within a single marker can also reduce the likelihood of classifying in a particular state. The ideal approach would be to identify a profile of biomarkers that contain sufficient information to i) positively identify the occurrence of toxicity and pathology against a background of biological and environmental variability and ii) identify the mode of action by which the toxicant or pollutant is acting. The power of transcriptomics (and in extension all other omics technologies) stems from the information within the profile. Furthermore, annotated genomes and bioinformatics tools for statistical analysis and biological interpretation of omics data are providing unprecedented opportunities to elucidate mechanisms of toxicity, identify susceptible populations and novel biomarkers and support risk assessment.

Gene expression signatures have been proved useful for classifying disease subtypes (Alizadeh et al. 2000;Khan et al. 2001), predict clinical outcomes (van 't Veer et al. 2002), identify diagnostic markers (Patino et al. 2005) or stratifying patient groups (van de Vijver et al. 2002) and to understand infectious disease processes. Much scientific and industrial effort has been undertaken to develop gene signatures in many clinical areas. One of the success stories is the MammaPrint or the first fully commercialized microarray-based multigene assay designed to individualize treatment for patients with breast cancer. The 70 genes that comprise the MammaPrint assay are focused primarily on proliferation with additional genes associated with invasion, metastasis, stromal integrity and angiogenesis (Kim and Paik 2010). It is envisioned that more gene signatures will become available

for clinicians to diagnose patients and to stratify them in order to select the most appropriate treatment.

Transcriptomics as part of the toxicogenomics analysis suite may also be useful for human (prospective) biomonitoring. There are a number of case studies showing that environment exposure to chemicals modifies the human transcriptome. The number of studies at present are quite limited and include populations exposed to ionizing radiation (Amundson et al. 2000), benzene (Forrest et al. 2005), dioxin, arsenic, metal fumes (Wang et al. 2005), and complex environmental exposures such as cigarette smoke (Lampe et al. 2004), diesel exhaust (Peretz et al. 2007) and air pollution (Leeuwen et al. 2006). Each of these studies identified potential biomarkers of exposure and/or early effect. The genes altered by these exposures represent a diversity of mechanisms including systemic effects on inflammation, which may underlie the development of associated diseases. The studies suggest that transcriptomics analysis are valuable for studying the biological response of the general population in the exposome-responsome continuum. In this respect, human disease is thought to arise when the normal physiological conditions of an individual, determined by the unique genetic background (genome), is perturbed by the exposome, a term describing all exposures from conception onwards. Such perturbations are then assessed by measuring the different components of the responsome using toxicogenomics technologies. Taking into account i) the complex nature of the exposome and its many interacting compounds (lifestyle, socio-economic status, etc.) and ii) the subclinical manifestation of the effects (or the large time gap between exposure and the actual manifestation of the disease), one needs to invest more effort in evaluating the long term value of transcriptomics (gene signatures). A correct exposure and/or phenotypic anchoring of the gene expression data needs to be done. Furthermore, influence of the dynamic nature of gene expression, within- and between-subject variability, lifestyle factors needs to be well documented. Lifestyle characteristics can also influence the outcome of gene expression studies. For example, nutritional factors like high protein or carbohydrate diet or a period of fasting induces differential expression of certain genes (Bouwens et al. 2007;van Erk et al. 2006). Also smoking is an important factor affecting gene expression in peripheral blood cells (Lampe et al. 2004;van Leeuwen et al. 2005).
One prerequisite to study gene expression in the general population is the availability of an accessible matrix, which is different in clinical context where researchers have access to biopsies, tumour samples etc. Blood is routinely used in human biomonitoring for measuring pollutants and metabolites and other well established markers such as hormones, cytokines. It is therefore the most obvious and practical surrogate tissue for studying gene expression in the human population. Peripheral blood is receiving a lot of interest for biomarker detection because of its critical role in immune response, metabolism and communication with cells (Mohr and Liew 2007). Studies indicate that gene expression in peripheral blood cells partly reflects gene expression in other tissues, making blood a surrogate tissue for studying effects of exposure in the human body (Liew et al. 2006;Rockett et al. 2004;Sullivan et al. 2006). Therefore, peripheral blood became the tissue of choice for developing molecular diagnostics (Baird 2006;Borovecki et al. 2005;Valk et al. 2004).

One of the objectives of our study was to analyze the **gene expression variation in the peripheral blood of a healthy population**. The underlying idea is to start documenting the population variability in gene expression. This can be considered as a **benchmark initiative** for future evaluation of gene expression changes in biomonitoring studies that are thought to result from exposure to environmental pollutants. Gene expression values that are correlated with exposure measurements and that are within the *normal* population variability might reflect true biological changes and these genes may hold potential as candidate biomarker.

A couple of studies that looked at gene expression variation in the peripheral blood of normal individuals are available (Eady et al. 2005;Karlovich et al. 2009;Palmer et al. 2006;Radich et al. 2004;Whitney et al. 2003). The between-subject variation of gene expression was studied by Radich et al. (Radich 2004), Whitney et al. (Whitney 2003) and Eady et al. (Eady 2005). The results of these studies are divers, but they show some overlapping genes (mainly interferon-related and MHC class II genes) with high variability between the individuals studied. Next to intra-individual variation, also inter-individual variation was studied. Contrasting conclusions have been drawn from these studies. Eady et al. concluded from their study that the within-individual variation is low (less than 20% for 80% of the genes). Whitney et al. observed intra-specific variation in gene expression and several highly variable genes. However, they stated that this was not the dominant source of variation among samples. Both studies also investigated the influence of sex and age on gene expression. These two factors appeared to induce some degree of differential gene expression. Interestingly, there was a substantial overlap between the sex-specific gene lists identified by the studies of Whitney et al. and Eady et al. Inter-individual sample variation was also found to be associated with the time of day the sample was taken (Radich et al. 2004;Whitney et al. 2003). Another study, performed by Meaburn at al. (Meaburn et al. 2009) investigated the reliability of gene expression profiles over 4 hours and the stability over 10 months. For this study, 5 twin pairs were recruited within a longitudinal study of behavioural development in twins. The authors identified probe sets that reliably detected individual differences across 4 hours and 73.7% of these probes was also stable over 10 months. The probe sets that reliably detect individual differences from a single peripheral blood collection and stably detect individual differences over 10 months are promising targets for research on the causes and correlates of individual differences in gene expression. Transcripts whose expression is not stable over time are potentially interesting, especially as an index of environmental effects (Meaburn et al. 2009). The main conclusion from these studies is that stable gene expression profiles can be obtained over a short time period, but there are significant differences between the participants in the study.

The studies mentioned above are valuable scientific documents with sound scientific conclusions. Nevertheless, we were convinced that additional research was warranted because those studies have investigated only short term variability. Samples have been collected over a period of maximum 5 weeks, with the exception of the study of Karlovich and coworkers who (Karlovich et al. 2009) who sampled over a period of 6 months (baseline, day 14, day 28, day 90 and day 180). In one part of this study, they focused on 11 inflammation-related genes and found that each gene was stably expressed, independent of age and gender. In another part of this study, microarray technology was used to study gene expression. Expression levels appeared to be constant over 1 month, but after three months, a small percentage of genes appeared to vary. A small proportion of genes was found to be differentially regulated according to gender, none by age. The difference between short term

and long term variation such as seasonal effect have never been documented. The studies have been performed with different analytical techniques (both sample collection and processing, as well as the microarray analysis) and different statistical approaches. This makes comparison between the studies almost impossible. Moreover, the studies of Whitney et al. and Eady et al have been performed with one of the first types of microarray platforms that were available. Since, then scientific community has experienced a maturation of the microarray technology. The reproducibility and reliability of the results has improved tremendously over the last decade. Also, those studies have outdated and a limited number of probes on the array. Within the context of the Environment and Heath research commissioned by the Flemish government there is considerable effort in developing gene expression signatures that can be useful for monitoring exposure to pollutants. This scientific effort has been centred around the University of Maastricht. They are in the processing of revealing interesting gene signatures using the latest Agilent microarray technology (see also WP3 of this study).

The aim of our investigation in WP2 of this study is to document gene expression population variability using the latest Agilent microarray technology (4X44K human arrays). This will allow integration of this new data set into existing microarray data aiming at biomarker identification. With WP2 we want to document short term and long term (season) variability. We also looked at the impact of gender on gene expression levels and potential interaction terms between time and gender.

More specifically, we investigated blood gene expression changes in a group of 22 non-smoking healthy volunteers (age between 20 and 40), with an equal distribution between male and female donors. Each volunteer provided a series of 6 blood samples, collected in Tempus™ blood collection tubes. Both short- and long-term variability was studied. Short-term variability was studied by 3 sampling days spread over 3 weeks (one per week) and long-term variability was investigated in view of possible seasonal influence on gene expression. Questionnaires were included in this study to characterize the health status, eating habits and possible sources of exposure (e.g. passive smoking) of the participants. The experimental setup was setup such that 2 main factors (gender and season) could be tested. The age of the participants was chosen to reflect on of the three ages groups that are also under scrutiny in the Environment and Health project of the Flemish government. We selected this age group for our study because it was the most convenient group from practical and ethical point of view to sample on several occasions. Moreover, the participants could be easily recruited within our company. A minimum of three samples within a season was chosen in order calculate basic descriptive statistics. The experimental setup was selected to maximize the interpretability of the study taking into account the budgetary restrictions. The WP2 study is the first pilot study to document gene expression variability in the general population over a longer period of time. The information gained from this study is valuable for evaluating the feasibility of such studies, future power calculations for gene expression studies. Also, robustness of candidate biomarker genes can be evaluated against a background of population expression variability.

2. MATERIALS AND METHODS

2.1 Selection of participants

Blood samples ware obtained from 22 healthy, non-smoking individuals. All volunteers were employees of the Flemish Institute of Technological Research (VITO) with an age ranging from 20-40 years. The average age (and body mass index) are 32.9 (24.6) and 31.4 (22.1) for male and female participants, respectively. For assessing inter- and intra individual short-term variation, as well as seasonal variation, individuals were asked to give blood samples once a week (2 x 3 ml) during three weeks in autumn of 2009 (week 46, 47 en 48) and again three samples in the spring of 2010 (week 17, 18 en 19). Blood was collected between 8 am and noon at the medical department of the Research Centre for Nuclear Energy (SCK) in Mol. A volume of 3 mL of whole blood was collected by phlebotomy in Tempus Blood RNA tubes in which 6 ml of stabilization reagent was present. After collection, the contents of the tubes was mixed by inverting the tubes. Samples were frozen within 1 hour at -20°C or kept at 4°C for maximum 24h before freezing them. Together with each blood sampling, subjects were asked to fill in a questionnaire (Annex 1) about health status, eating habits and exposure to environmental pollutants during the last 24 hours. This study was approved by the ethics committee of the Antwerp University (UA A09 21). The approval is added as Annex 2. All subjects gave written informed consent to participate in the study.

2.2 Selection of method

When using blood for gene expression profiling, some important issues need to be considered. First, the technology used to perform the analyses is a factor that needs to be taken into account when studying gene expression. The method used to collect blood, isolate and purify the RNA greatly influences the results of the study (Asare et al. 2008). In this study it is demonstrated that peripheral blood RNA isolation methods can critically impact differential gene expression results, particularly in a setting where fold-change differences are typically small and there is inherent variability within biological cohorts. Moreover, different gene expression measurement technologies deliver slightly different results (Arikawa et al. 2008).

Blood is a complex tissue, containing a variety of cell types (erythrocytes, granulocytes, lymphocytes, monocytes, natural killer cells and platelets). Each of these cell types has a unique expression signature and the relative proportions contribute to the general gene expression profile (Whitney 2003). The relative proportions of the cell types can change rapidly with states of health and disease leading to variability that may confound the interpretation of gene expression differences between control and disease groups. In some studies, blood is fractionated prior to RNA extraction. This process allows studying homogenous cell populations. Nevertheless, there are many advantages of profiling gene expression from whole blood rather than from subpopulations. Cell fractionation requires additional processing steps and it is accompanied by some degree of cell activation that may alter gene expression patterns and thus may be considered as processing artefacts. Moreover, isolation of RNA from whole blood can be performed with easy and rapid procedures, which results in a reduced cost. Two products for whole blood RNA collection are commercially available that enable minimal blood handling procedures thus minimizing the risk of inducing changes in gene expression through blood handling or processing. In both systems (PAXgene™ and Tempus™) the

blood is immediately lysed when collected into the blood collection tube and RNA is stabilized using proprietary reagents.

The relative high proportion of globin mRNA present in total RNA extracted from whole blood can reduce the efficacy of the microarray assay (Feezor et al. 2004;Liu et al. 2006). The impact of globin RNA transcripts on gene expression results can be reduced by using methods that remove or block globin RNA during the microarray assay. Globin reduction results in a consistent and significant increase in the quality of microarray data. Detection sensitivity is improved most dramatically for low abundance genes, especially when using Affymetrix array technology (Field et al. 2007;Tian et al. 2009). Recently, Wright et al. have shown that absolute concentrations of globin and differences in transcript concentrations within a sample set are factors that cause globin interference in the case of Agilent microarray technology.

Extrapolation of microarray results from one study to another needs to take into account many different aspects of sample processing and data processing and those need to be communicated at length in order to make sure that there are technical issues that cause differences between microarray results. The approach for sample collection was the following in WP2: blood collection and preservation in Tempus tubes (Ambion). Extraction of total RNA using the recommended Tempus spin kit.

2.3 RNA extraction

Total RNA was extracted using the Tempus Spin RNA Isolation kit (Applied Biosystems) according to the manufacturer's instructions. First, samples were thawed at room temperature. Total volume was adjusted to 12 ml with phosphate buffered saline provided with the kit. Samples were vortexed vigorously and centrifuged at 3 000 g for 30 minutes at 4°C. After centrifugation, the supernatans was poured off. The pellet, containing the RNA, is resuspended by adding Resuspension Solution. This mixture was brought onto a filter and after a few washing steps, the RNA was eluted in 90 µl of elution buffer. RNA concentration was determined using a NanoDrop Spectrophotometer (NanoDrop Technologies, Wilmington, DE). In a next step, RNA was precipitated with ammonium acetate and ethanol and GlycoBlue as coprecipitant. The RNA pellet was washed with 70% ethanol and redissolved in nuclease-free water to obtain a concentration of at least 70 ng/µl. RNA yields are checked using the NanoDrop.

2.4 RNA purification and globin reduction

Whole blood consists of a relatively large proportion of globin mRNA transcripts. These transcripts dilute the mRNA population and decrease sensitivity of detecting less abundant mRNAs using microarray technology. The Globinclear kit (Ambion) was used to deplete globin mRNA from total RNA preparations. In short, biotinylated globin-capture DNA oligos were added to 4 µg of total RNA and globin mRNA were removed by streptavidin magnetic beads. The remaining globin-reduced total RNA was purified using magnetic beads and eluted in 30 µl of elution buffer. RNA integrity was tested with capillary gel-electrophoresis using RNA 6000 Chips (Agilent Technologies, Palo Alto, CA), analyzed on the Agilent 2100 Bioanalyzer. RNA was considered to be intact when showing an RNA integrity number (RIN) of seven or more. Samples were stored at -80°C until further use.

2.5 RNA amplification and labelling

Total RNA was amplified and labelled to generate complementary RNA (cRNA) using the Low Input Quick Amp Labelling (one color) kit (Agilent Technologies) according to the manufacturer's instructions. Briefly, 100-200 ng of total RNA was reverse transcribed into complementary DNA (cDNA) using T7-promotor primer and MMLV reverse transcriptase. The cDNA was transcribed into cRNA, during which it was fluorescently labelled by incorporation of cyanine (Cy)3-CTP). The single-stranded, labelled cRNA was purified with Qiagen's RNeasy mini spin columns (Qiagen, Hilden, Germany). Yield and specific activity (dye-incorporation rate) were determined using a NanoDrop Spectrophotometer (NanoDrop Technologies). cRNAs showing a specific activity of more than 8 pmol/µg cRNA were selected for further processing.

2.6 Microarray hybridization and scanning

A total amount of 1.65 µg cRNA sample was hybridized on 4x44K Whole Human Genome microarray slides (design 014850_D_20070207, Agilent Technologies) for 17 hours using the automated HS4800TM pro hybridization station (Tecan, Männedorf, Switzerland) according to the manufacturer's instructions. The arrays were scanned on an Agilent DNA microarray scanner (G2565BA) and further processed using Agilent Feature Extraction Software (Version 10.5). The use of this software included automatic grid positioning, intensity extraction (signal and background) and quality control. Details on the data processing steps used to generate the Agilent one-color output can be found in the Agilent protocol GE1-10.7-SEP09. The gProcessedSignal per probe was used for further analysis. This value is the result of Agilent's Feature Extraction algorithm that corrects raw intensity signals by subtracting background values and by applying multiplicative detrending. This algorithm is designed to compensate for slight linear variations in intensities that can occur if the processing is not homogeneous across the slide. This non-homogeneous processing results in different chemical reaction times, for example, between the sides and the center, and produces a dome effect. Prior to statistical analysis, the gProcessedSignals were normalized using quantile normalization using the R function *normalizeBetweenArrays*. This was followed by a $\log_2$ transformation of the data. Some of the probes are replicated per array and the median of the replicates was taken for these probes. This resulted in a data file with 41 000 unique probes.

3. RESULTS

3.1 Description of the study population

From 25 potential candidates for this study, 22 participants were selected who donated blood samples (two fractions) in autumn 2009 (week 46, 47 and 48) and spring of 2010 (week 17, 18 and 19). Each participant completed a questionnaire at the moment of blood collection. The results of this questionnaire are summarized in Table 1. The data from the questionnaires were not statistically analyzed because we are dealing with a small population. The information is used to characterise the population, to ensure that we are dealing with a **healthy population in a normal environmental situation**. In the case of an outlier in the microarray slides, the questionnaire could help to identify if there was a specific situation for the volunteer (such as sickness, exposure to chemicals, etc.).

Table 1 Results of the questionnaire

Health state	%	Exposure during the last 24 h	%
Fasting	12	Vegetarian meal	21
State of health (good)	93	Fish consumption	16
Disease	14	Meat consumption	75
Health problem	14	Alcohol (1 glass or less)	77
Use of medication	15	Coffee (3 cups or less)	71
Tired	33	Dietary supplements	8
Amount of sleep (6 h or more)	71	Passive smoking	11
Stress	17	Practice sport	22
		Exposure to chemicals	20
		Daily travel time	81 min

3.2 Sample processing

In total 132 samples were collected (6 samples per individual). These samples (first fraction) were processed according to the method described in Materials and Methods. The second fraction was stored at -20°C as back-up sample. Samples were randomized and each step in the sample processing was controlled as much as possible. The average RNA yield per sample was 9.94 µg (± 2.47) and the RNA integrity (RIN) was on average 8.70 (± 0.50). The average yield was 8.44 µg (± 2.71) after RNA precipitation. An aliquot of 4 µg precipitated RNA was depleted for globin mRNA, resulting in an average RNA yield of 3.00 µg ± (0.41). The quality of the samples was checked and resulted in a RIN = 8.93 ± 0.40. All samples were of good quality (RIN > 7) and were labelled with Cy3. Labelling resulted in an average specific activity or 16.73 pmol/µg (± 2.75) for each sample. Finally, samples were hybridised onto Agilent 4X44k humane whole-genome microarrays with design number 014850.

3.3 Quality control of the microarrays

A primary quality control of the microarrays was performed using the Agilent Feature Extraction software. This software generates 10 quality measurements that give an idea about the quality of the hybridization process. 128 of the 132 samples met at least 9 of the criteria. The quality of 4 samples was somewhat lower as shown by the reproducibility (coefficient of variation) of spike-ins and negative control sports. The values were slightly above the background and did not indicate a true failure in the hybridization or microarray scanning process. Nevertheless, it was concluded to redo the full microarray workflow for these samples. The second round of hybridization using the same RNA for these four samples was successful. A typical summary of the control parameters for one array is shown in Figure 1A. Figure 1B summarized the control parameters for all the microarrays and it should be clear that all samples have comparable scores for the quality parameters. The average technical reproducibility within an array, based on repeated probes on the array, was high (Figure 2). Median and standard errors of the coefficient of variation was 4.4% (± 1.1). The different control parameters indicate that the data is technical reproducible and of good and homogenous quality.

A

Metric Name	Value	Excellent	Good	Evaluate
IsGoodGrid	1.00		>1	<1
AnyColorPrcntFeatNonUn...	0.00		<1	>1
gNegCtrlAveNetSig	42.61		<40	>40
gNegCtrlAveBGSubSig	-4.23		-10 to 5	<-10 or >5
gNegCtrlSDevBGSubSig	2.53		<10	>10
gSpatialDetrendRMSFilt...	3.03		<15	>15
gNonCntrlMedCVProcSign...	4.70		0 to 8	<0 or >8
gE1aMedCVProcSignal	3.98		0 to 8	<0 or >8
absGE1E1aSlope	0.97		0.90 to 1.20	<0.90 or >1.20
DetectionLimit	1.45		0.01 to 2	<0.01 or >2

B

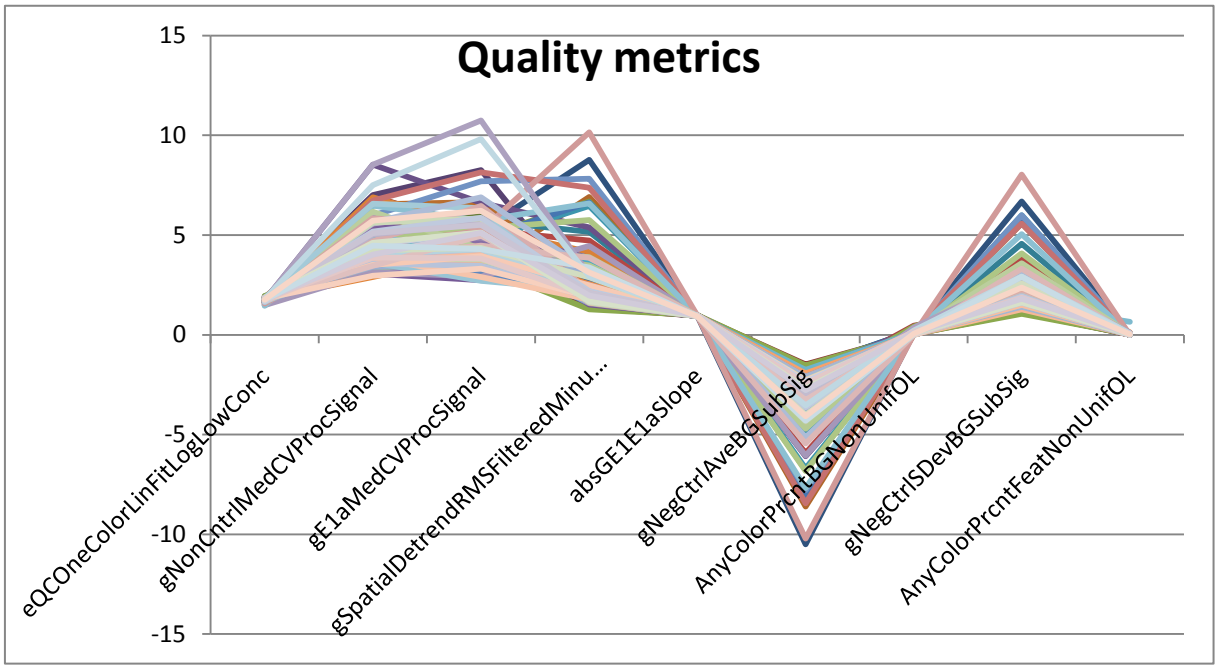


Figure 1 Quality control summary from Agilent Feature Extraction software for one microarray (A), Summary of the quality parameters for all the arrays (B). Each line represents the values for 1 array. The values for each parameter are shown in the first column of Figure 1A

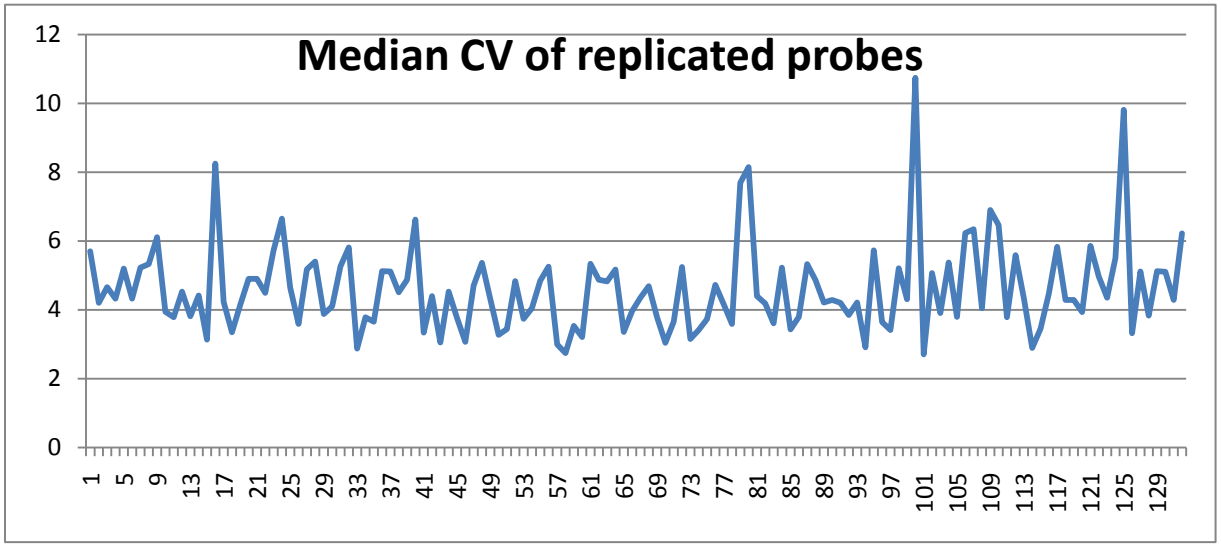


Figure 2 The technical reproducibility within an array is based on repeated probes on the array and is plotted as the median coefficient of variation of the replicated probes. The results are shown for the final set of 132 microarrays

Subsequent to the first quality control, explorative data analysis was used to detect bias introduced by hybridization. Spearman correlation coefficients were calculated between each two arrays, based on the gProcessedSignal per gene. This gProcessedSignal was generated by Feature Extraction and represents the intensity signal per gene after background correction and a spatial detrending algorithm. The Results were plotted in two different heat maps. The samples in the first plot were identified and ordered by microarray slide number (Figure 3A). In the second heat map the samples are organized according to the participants and ordered by sampling date (Figure 3B). The heat maps were constructed to identify arrays with overall gene expression patterns that are different. This could be due to technical aspects (such as hybridization date or hybridization batch). This could be visualized as lower Spearman correlation coefficients in Figure 3A. Different array correlations due to subject or time sampling aspects are potentially identified in Figure 3B. The Spearman correlation between the arrays is high (r_s>0.96). Dark red represents a perfect correlation (r_s=1) between the samples (self-self correlation). Weaker correlations are shown in blue. The three samples showed a somewhat lower correlation, but still high. The minimum correlation is 0.94. The questionnaires were consulted to check if variability between the arrays could be explained to items reported by the volunteers. The lower correlation could not be related to (i) specific conditions for the respective individuals or (ii) the quality of the processed RNA sample. Therefore, the three samples were left in the full dataset. Overall, the dataset is of good and homogenous quality based on Spearman correlation. The within-subject correlation is higher than the between-subject correlation indicating that the level of gene expression is individual-specific. This is concluded from the heat map given in Figure 3B and can be observed as stronger correlation (blocks of dark red colour) corresponding to the samples from the same individual. The effect is subtle because all the correlation coefficients are quite high.

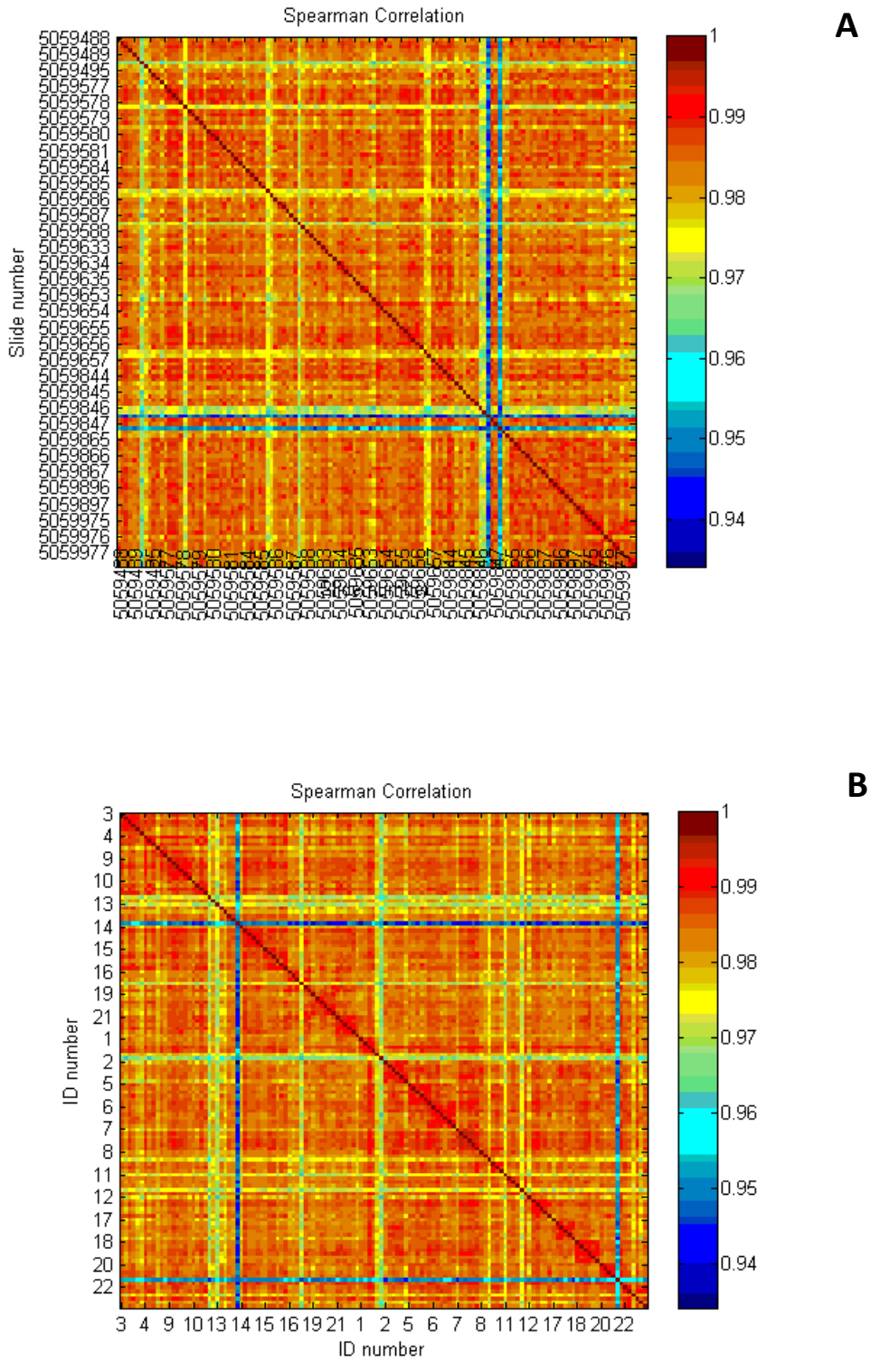
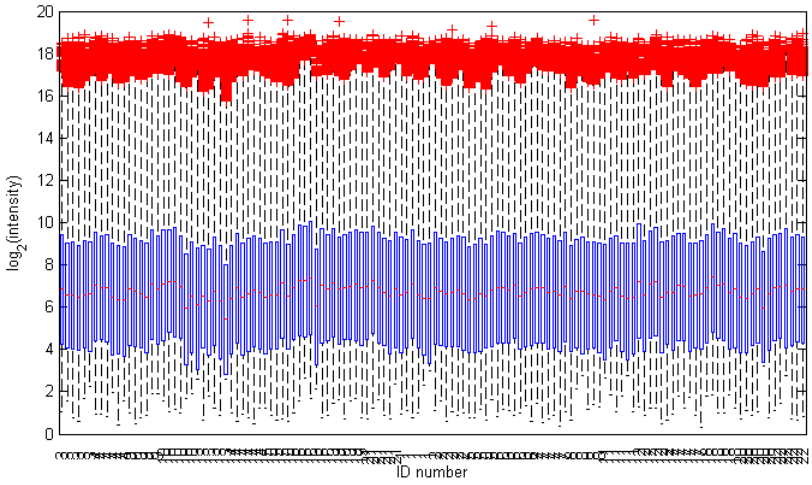


Figure 3 Heat maps based on Spearman correlation. The 132 samples were ordered by array identification (A) and by sample identification (B). The ID number refers to the different participants. The first 10 number (from 3 to 21) refer to males, the following numbers (1 to 22) refer to females. The colours indicate the strength of the correlation

Subsequently, boxplots were created for all microarray data, based on the log_2(gProcessedSignal). Figure 4A shows the box plots ordered by sample identification and the plots are ordered per participant and per time point. No trends can be observed that indicate technical bias (hybridization and or sampling aspect (individual and time point). The same figure was generated for plots arranged by array identification (not shown). Figure 4A indicates that signal intensity fluctuates between arrays because of technical aspects related to hybridization. This observation justifies an additional normalization step by means of quantile normalization on the log_2(gProcessedSignal). This process makes the signal intensity distribution between the arrays comparable and the result of this step is shown in Figure 4B. Samples are ordered by participant and per time point in this plot.

A



B

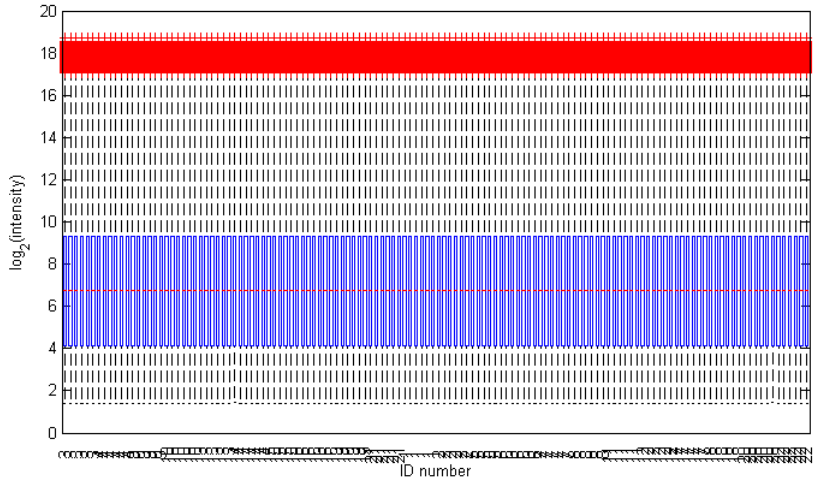


Figure 4 Box plot based on log_2(gProcessedSignal). Samples are ordered by sample identification (A), Box plots of the log_2(gProcessedSignal) after performing quantile normalisation (B)

3.4 Descriptive analysis

The results of the descriptive statistics are given in the following histograms. The results have been organized according to the two main factors that are under scrutiny (gender and season). Figure 5 shows the histogram for the signal intensity and indicates that there is a bimodal-like distribution. There is high number of probes with low intensity expression and the expression of these probes may be largely due to background signal and/or noise. The inclusion of these non-informative probes in the data set may mask subtle biological effects. The inclusion of a large number of these probes may have important consequences for multiple testing approaches: more probes in the data set requires more corrective measures and hence increase the false discovery rate. The number of non-uninformative probes was checked by applying a filtering procedure that is routinely used at the Maastricht University for Agilent microarrays. The data of WP2 (41K probes) were processed through their pipeline (more details in WP3). The main parameters used to flag spots as good or bad spots are i) size of the spot, ii) mean/median ratio of the signal from different pixels in the spot, iii) saturation of the signal and iv) intensity versus background signal (remove low intensities). About 95% of the bad spots fail at point 2 or 4 (mainly 4). Saturation is not an issue when slides are scanned on the Agilent scanner. This additional filtering procedure led to a reduced 33K probe list or about 25% of the probes did not pass the additional quality check. The histograms for the signal intensity of these 33K is shown in Figure 6 and it is clear that most of the low intensity probes have been removed. It was on visually checked that there are no differences exist between the distributions. All subsequent analyses have been performed on the 33K probe list.

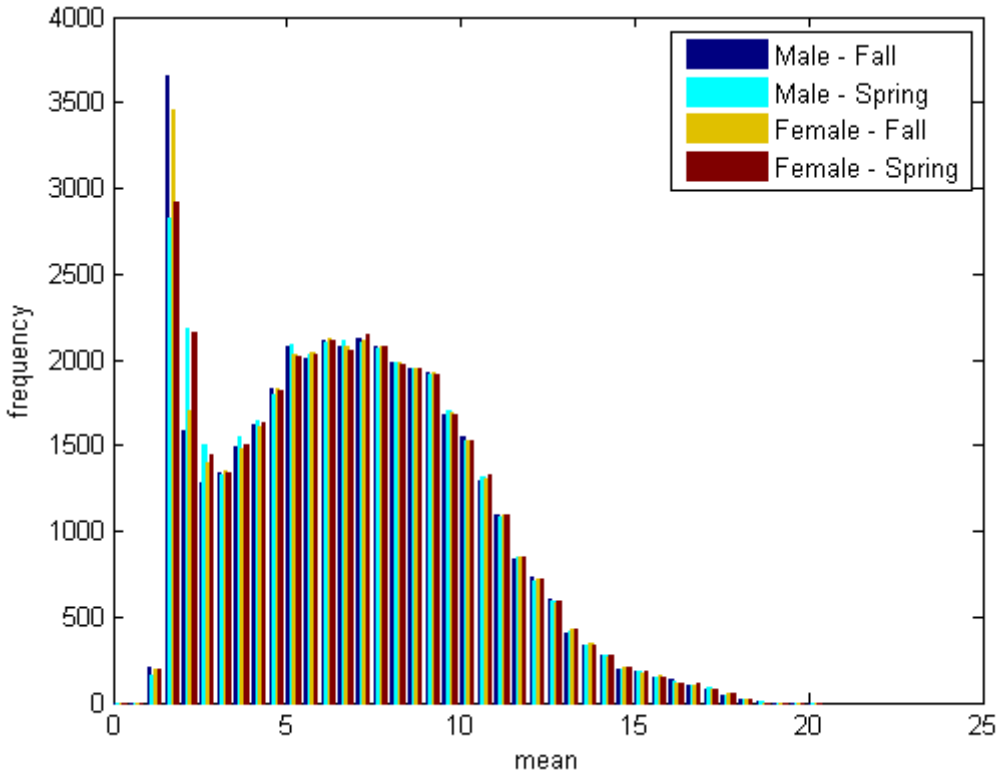


Figure 5 Histograms of signal intensity for 41K probes. The data are organized in 4 categories according to the main factors (gender and season). The signal intensity is given in $\log_2$-scale

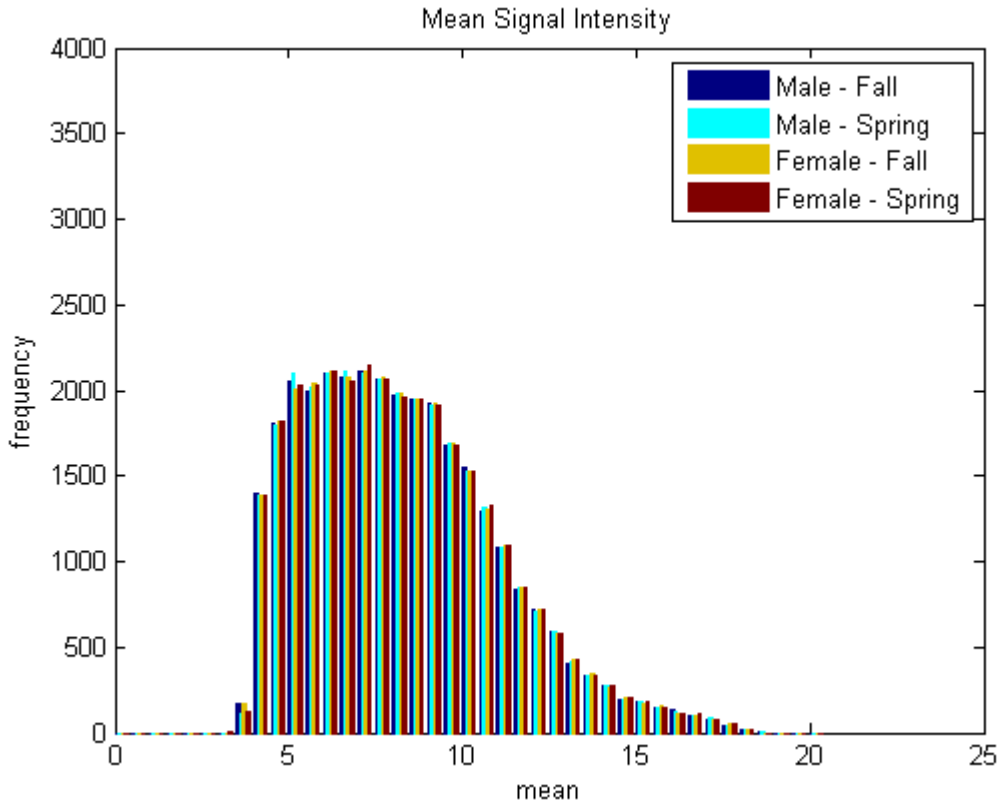


Figure 6 Histograms of signal intensity for 33K probes. The data are organized in 4 categories according to the main factors (gender and season). The signal intensity is given in $\log_2$-scale

The heat map with the between-sample Spearman correlation were again calculated using the 33K probe list and the result is shown in Figure 7. After filtering out non-informative probes the stronger correlation within the individual is more apparent and is given as the more intense red colours around the diagonal. The within-subject correlation is more apparent when non-informative probes are removed using a stringent quality control of the raw data (compare Figure 7 with Figure 3B).

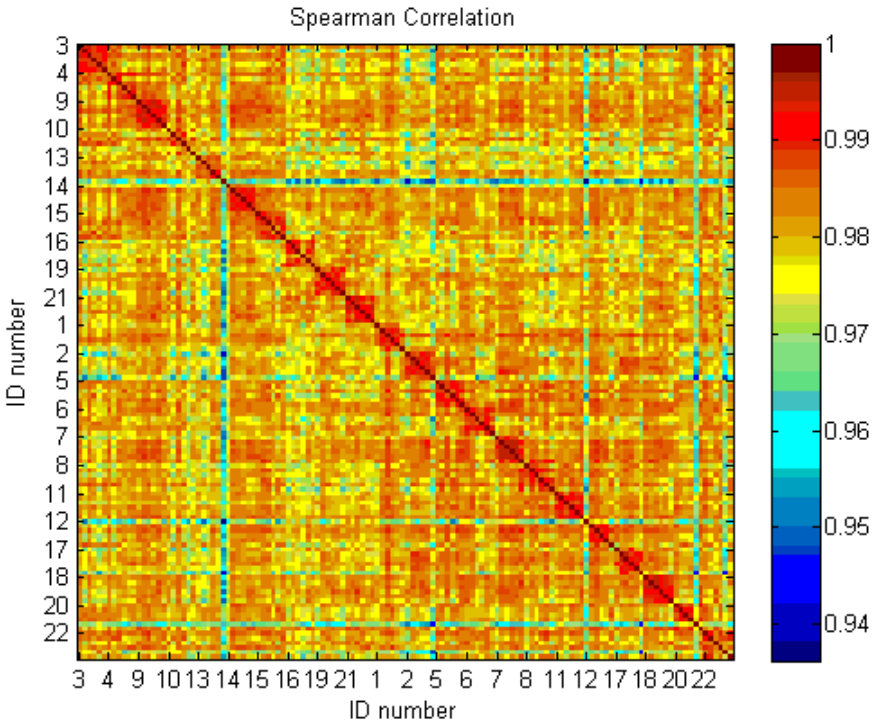


Figure 7 Heat maps based on Spearman correlation. The 132 samples were ordered by sample identification. The ID number refers to the different participants. The first 10 number (from 3 to 21) refer to males, the following numbers (1 to 22) refer to females. The colours indicate the strength of the correlation

Figure 8 represents the histograms of the $\log_2$ standard deviation. There seems to be differences in frequencies for genes with lower standard deviation. However, no clear trends according to the main factors (gender and season) could be identified. Figure 9 gives some examples to identify that the size of the standard deviation depends on the signal intensity. Higher intensity probes are more prone to variation. This has consequences for detecting differential gene expression (DGE) as it will be more challenging to identify DGE for variable genes when the effect size is small for these probes. *A priori* power calculation to identify the population size needed depends of the signal intensity, standard deviation per gene and the expected effect size per gene. To get an idea about the size of the standard deviation in function of signal intensity we have plotted intensity versus standard deviation for males and females separately for a specific time point in a season. The same plot types were obtained for other time points. Figure 9 is shown to illustrate that signal intensity per gene plays an important role for estimating standard deviation. The latter is an important parameter for power calculations (besides the effect size). Hence, it should be clear that for estimating future study populations one needs to take into account the size of the individual standard deviations.

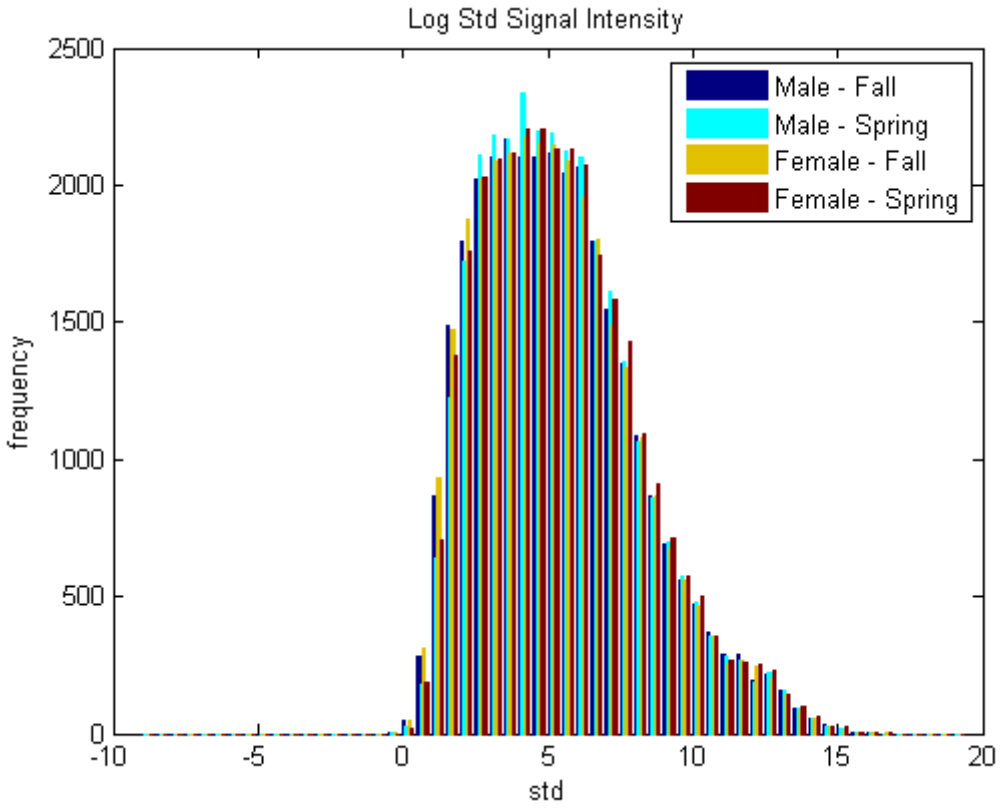


Figure 8 Histograms based on the $\log_2$ standard deviation. Samples are grouped by the factors gender and season

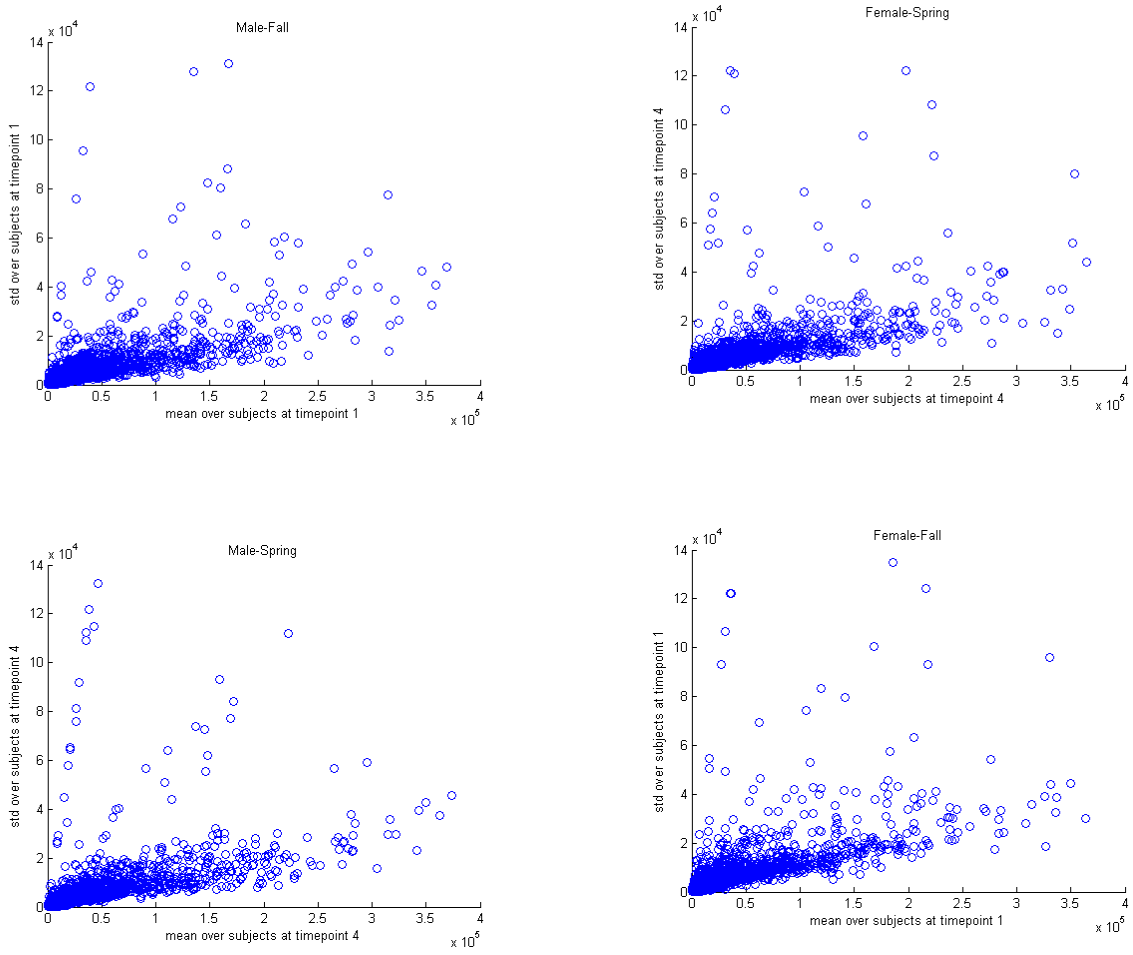


Figure 9 Illustration of the variation of standard deviation in gene expression. The size of the standard deviation depends on the signal intensity of the individual gene. This effect is plotted for male and female individuals separately for one season and one time point

Finally, we calculated the coefficient of variation for the four different categories. For this plot the average probe expression over the different volunteers was taken within a season. It is clear from Figure 10 that there is a large variation in the coefficient of variation, with a long right tail. For the bins with a CV<0.3, there seems to be a difference between two seasons *i.e.* more probes with a lower coefficient of variation in fall. One should compare the dark blue bars with the light blue and the orange versus the red bars for the left part of the histograms in Figure 10. The trend seems to be independent of the gender and therefore we grouped all individuals and calculated the average probe expression per season, calculated the coefficient of variation for each probe and constructed a histogram. The result is given in figure 11. The observed phenomenon that the coefficient of variation between subjects varies according to the season can be better observed for coefficients of variation lower than 0.3.

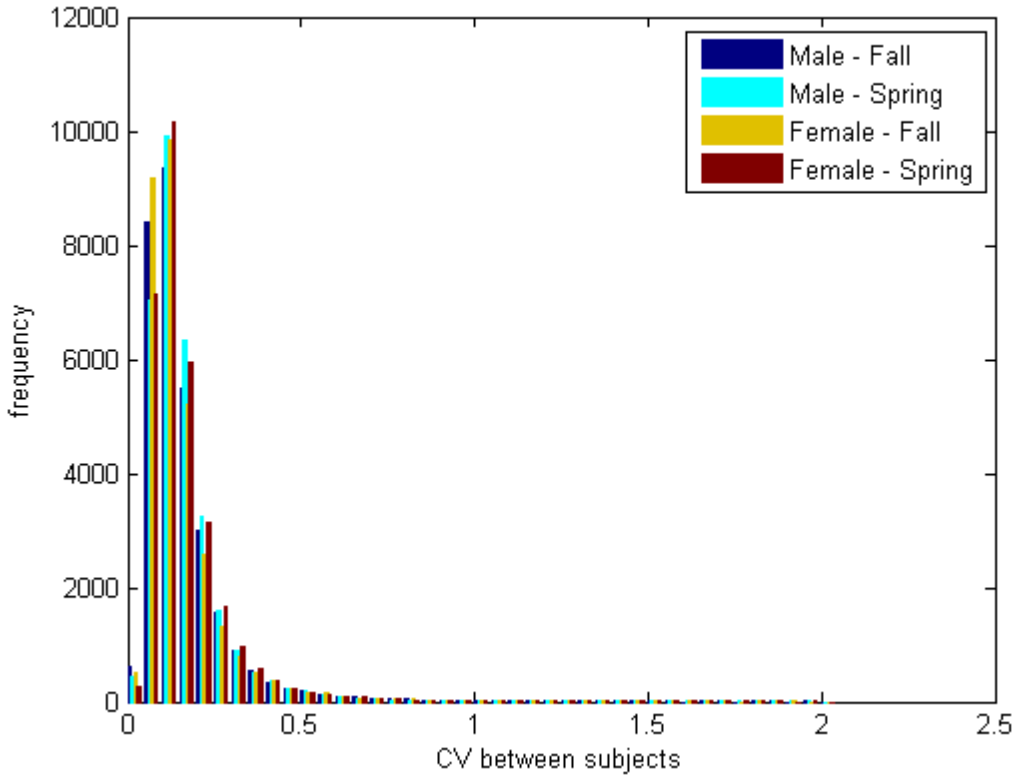


Figure 10 Distribution of the coefficient of variation between subjects for the two main factors (season and gender)

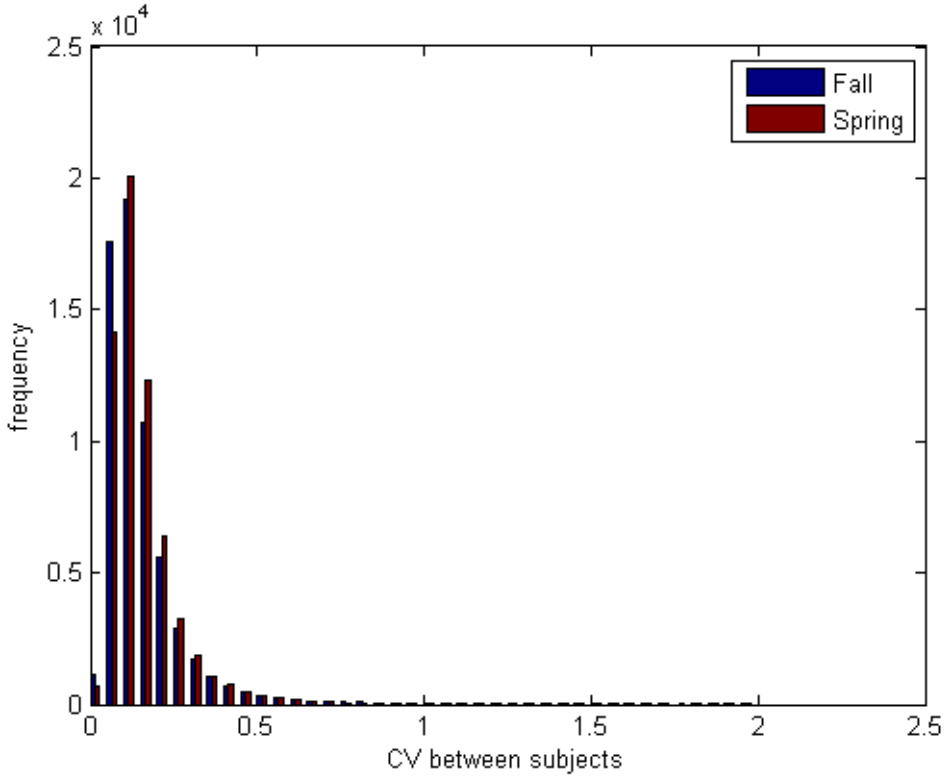
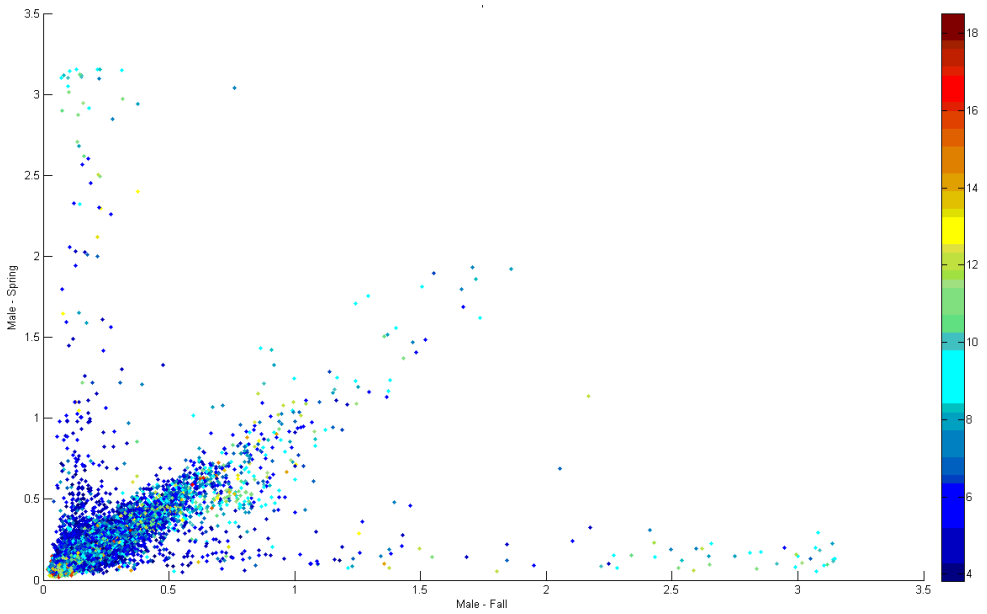


Figure 11 Distribution of the coefficient of variation between subjects when the data are grouped according to season

3.5 Exploratory data-analysis

The data were plotted in separate 2-dimensional scatter plots in which the coefficient of variation of one season was plotted against the other season for each of the 33K probes. The dots were coloured according to their average expression value over both seasons (A-value or $\frac{1}{2}(\log_2 I_{season1} + \log_2 I_{season2})$) as measure for general expression level. Alternatively, the colouring was done according to the M-value or $\log_2 I_{season1} - \log_2 I_{season2}$, which can be considered as a measure for DGE. Figure 12A and B are the plots for the male individuals. Figure 13A and B are for females. Figure 12A and Figure 13A indicate that the majority of the probes have an average expression level (given as blue dots), but a large distribution for the coefficient of variation ranging from 0.01 to 1.Figure 12B (males) and Figure 13B (females) indicate that very few probes are candidates for differential expression. Probes that are not affected by the factor season have generally an M-value around zero and are coloured green. candidate probes for differential expression have a red (upregulated) or a blue colour (downregulated). Most of these probes are situated in the top left or the bottom right corner of Figure 12B or Figure 13B. They correspond to probes with a high coefficient of variation in a one season, whereas the coefficient of variation is low in the other season. This rough but straightforward approach gave this unexpected observation of these *extreme* probes that are highly variable between both seasons. The high coefficient of variation within a particular season may be indicative for an interindividual variation and hence a subject-dependent expression. These *extreme* probes were investigated more closely by listing all probes with an arbitrary cut-off (for the coefficient of variation) above two. Subsets of probes were generated separately for male and female individuals.

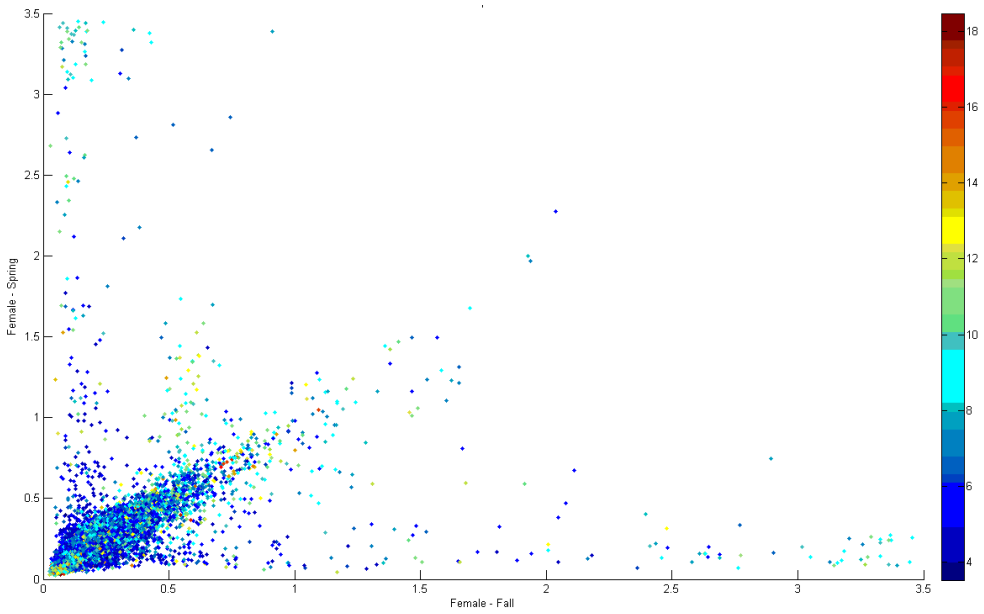
A



B

Figure 12 Scatter plot of the coefficient of variation per season for male individuals. Each dot indicates one probe. The colour code indicates the average expression intensity over the seasons (A-value) (A) or the difference in intensity (M-value) over the two seasons (B)

A



B

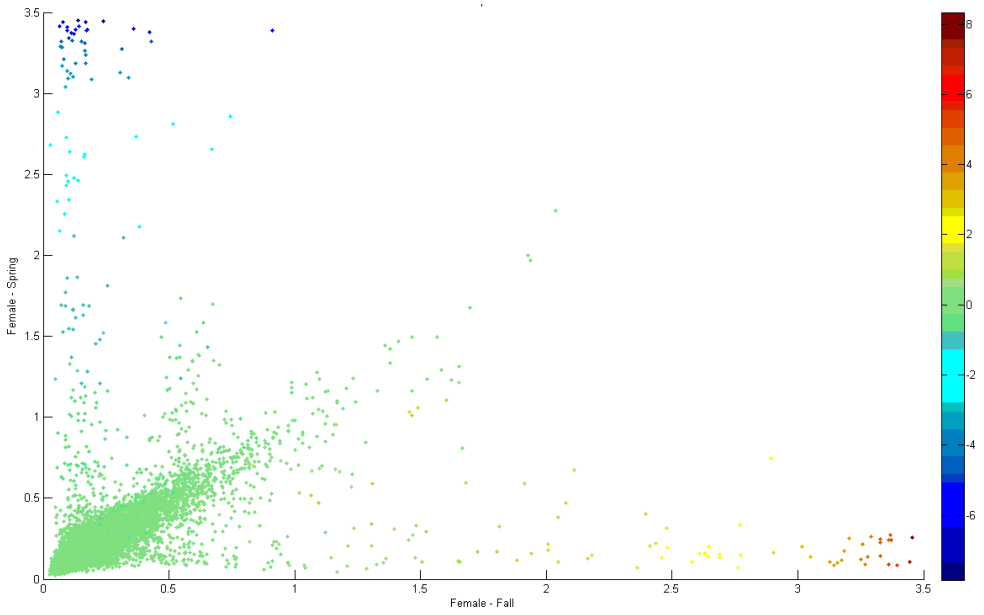


Figure 13 Scatter plot of the coefficient of variation per season for female individuals. Each dot indicates one probe. The colour code indicates the average expression intensity over the seasons (A-value) (A) or the difference in intensity (M-value) over the two seasons (B)

Table 2 lists the number of *extreme* probes (coefficient of variation above two) and the number of mapped genes per condition. A Venn diagram of those mapped genes indicated that very few genes were in common between the different conditions meaning that those genes appear to be gender and season-specific (Figure 14). In spring, male and female individuals had guanidinoacetate N-methyltransferase (GAMT) in common. In Fall, the ral guanine nucleotide dissociation stimulator-like 1 (RGL1) were in common. Female individuals had glycine-N-acyltransferase-like 2 (GLYATL2) in common between the two seasons. The list of all mapped *extreme* genes are given in Annex 3.

Table 2 Number of probes and mapped genes that were considered as *extreme* based on the coefficient of variation larger than two. The number of probes and genes are summarized over the two main factors: gender and season

Condition	Probes	Genes
Male-Fall	42	32
Male-Spring	41	33
Female-Fall	52	38
Female-Spring	62	40

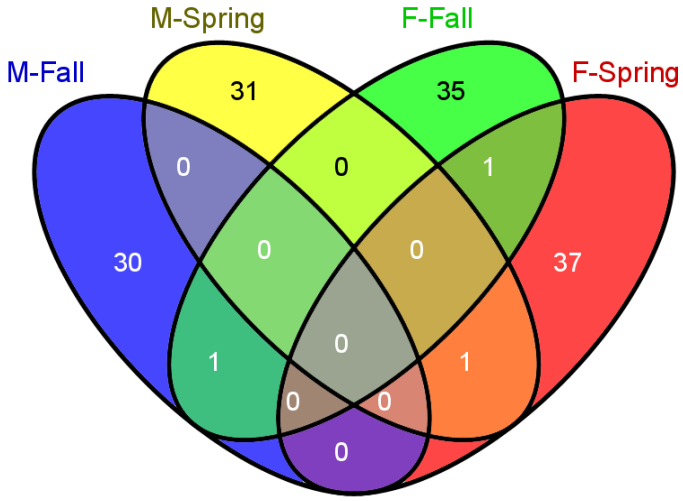


Figure 14 Venn diagram showing the unique and overlapping genes for the extreme genes that have been identified for the two main factors (gender and season). Extreme genes were selected based upon the cut-off for the coefficient of variation larger than two for the four categories according to the main factors

The extreme genes were grouped per gender and a standard Ingenuity Pathway Analysis Analysis (IPA) was done on male and female genes, separately. A summary report extracted from IPA is given in Figure 15 and Figure 16. The results in the different Figures as top-5 identified i) biological networks, ii) bio functions (organized according to 3 different ontologies: disease and disorders, molecular and cellular functions, physiological system development and function), and iii) canonical pathways. The ontologies are proprietary classes developed and maintained by Ingenuity Systems. Biological networks are established and scored according to proprietary Ingenuity Systems algorithms. Canonical pathways include major signalling pathways and metabolic pathways that are extracted from open-source databases (such as KEGG) and manually annotated pathways. Statistical analysis is performed according to classic Fisher Exact enrichment analysis with p-values<0.05 considered as significant.

Top Networks

ID	Associated Network Functions	Score
1 View	Cell Death, Immunological Disease, Cellular Development	25
2 View	Gene Expression, Cell Signaling, Cell Death	2
3 View	Cardiovascular System Development and Function, Tissue Development, Embryonic Development	2
4 View	Cellular Development, Cellular Growth and Proliferation, Hair and Skin Development and Function	2
5 View	Cell Cycle, Cell Morphology, Cellular Assembly and Organization	2

Top Bio Functions

Diseases and Disorders

Name	p-value	# Molecules
Cancer	3.38E-03 - 4.63E-02	18
Connective Tissue Disorders	3.38E-03 - 3.38E-03	1
Metabolic Disease	3.38E-03 - 1.01E-02	1
Reproductive System Disease	5.95E-03 - 2.33E-02	5
Genetic Disorder	6.74E-03 - 1.34E-02	4

Molecular and Cellular Functions

Name	p-value	# Molecules
Amino Acid Metabolism	6.74E-03 - 1.01E-02	2
Cell Cycle	6.74E-03 - 3.00E-02	2
Cell Morphology	6.74E-03 - 4.31E-02	3
Cellular Compromise	6.74E-03 - 4.95E-02	2
Cellular Development	6.74E-03 - 4.31E-02	3

Physiological System Development and Function

Name	p-value	# Molecules
Hair and Skin Development and Function	6.74E-03 - 6.74E-03	1
Connective Tissue Development and Function	1.34E-02 - 1.34E-02	1
Skeletal and Muscular System Development and Function	1.34E-02 - 1.34E-02	1
Digestive System Development and Function	1.68E-02 - 1.68E-02	1
Cell-mediated Immune Response	4.31E-02 - 4.31E-02	1

Top Canonical Pathways

Name	p-value	Ratio
B Cell Development	3.71E-03	2/30 (0.067)
One Carbon Pool by Folate	7.18E-02	1/22 (0.045)
Antiproliferative Role of TOB in T Cell Signaling	8.43E-02	1/26 (0.038)
IL-22 Signaling	8.74E-02	1/27 (0.037)
4-1BB Signaling in T Lymphocytes	9.36E-02	1/33 (0.03)

Figure 15 Summary of Ingenuity results for *extreme* genes in male individuals

Top Networks

ID	Associated Network Functions	Score
1 View	Cell Cycle, Cancer, Cellular Movement	16
2 View	Cancer, Cellular Assembly and Organization, Cellular Compromise	2
3 View	Cell Morphology, Cellular Compromise, Cell Cycle	2
4 View	Cellular Assembly and Organization, Cell Cycle, Nervous System Development and Function	2
5 View	Gastrointestinal Disease, Hematological Disease, Immunological Disease	2

Top Bio Functions

Diseases and Disorders

Name	p-value	# Molecules
Cancer	4.73E-03 - 4.63E-02	6
Cardiovascular Disease	4.73E-03 - 4.73E-03	1
Genetic Disorder	4.73E-03 - 3.35E-02	9
Hematological Disease	4.73E-03 - 2.34E-02	2
Immunological Disease	4.73E-03 - 1.41E-02	1

Molecular and Cellular Functions

Name	p-value	# Molecules
Cellular Movement	2.18E-04 - 4.63E-02	4
Cell Cycle	4.73E-03 - 3.72E-02	3
Cell Death	4.73E-03 - 4.18E-02	1
Cell Morphology	4.73E-03 - 4.63E-02	2
Cell-To-Cell Signaling and Interaction	4.73E-03 - 4.63E-02	3

Physiological System Development and Function

Name	p-value	# Molecules
Cardiovascular System Development and Function	4.73E-03 - 4.63E-02	1
Connective Tissue Development and Function	4.73E-03 - 1.88E-02	1
Embryonic Development	4.73E-03 - 3.26E-02	1
Endocrine System Development and Function	4.73E-03 - 4.63E-02	3
Immune Cell Trafficking	4.73E-03 - 2.08E-02	3

Top Canonical Pathways

Name	p-value	Ratio
Actin Cytoskeleton Signaling	2.86E-03	5/226 (0.022)
mTOR Signaling	2.96E-02	3/148 (0.02)
Propanoate Metabolism	3.68E-02	2/64 (0.031)
Basal Cell Carcinoma Signaling	4.33E-02	2/72 (0.028)
Renal Cell Carcinoma Signaling	4.44E-02	2/73 (0.027)

Figure 16 Summary of Ingenuity results for *extreme* genes in female individuals

The data in Figure 12 and Figure 13 were also summarized by filtering on the coefficient of variation and summarizing the data in histograms. The data were broken down according to their expression values. Three categories (Low (A<6), Medium (6<A<12) and High (A>12)) were used for this. The Boolean operator 'AND' was used to select probes with a coefficient of variation below a defined threshold, which was varied between 0.05 to 0.50. The latter values correspond to 5% and 50% coefficient of variation (Table 3 and 4).

Table 3 Percentage of genes with a coefficient of variation (CV) less than a defined thresholds in both seasons. The percentages are broken down according to the expression level of the probes. The last column contains the total amount of genes that meets the criteria. Results are shown for male subjects

CV	Low	Medium	High	Total number of probes
0.5	0.23	0.66	0.10	30220
0.45	0.23	0.67	0.10	29940
0.4	0.23	0.67	0.10	29497
0.35	0.23	0.67	0.10	28820
0.3	0.22	0.67	0.11	27690
0.25	0.21	0.68	0.11	25709
0.2	0.19	0.69	0.12	21807
0.15	0.16	0.71	0.13	15030
0.1	0.09	0.73	0.18	5223
0.05	0.01	0.73	0.26	122

Table 4 Percentage of genes with a coefficient of variation (CV) less than the proposed thresholds in both seasons. The last column contains the total amount of genes that meets the criteria. Results are shown for female subjects

CV	Low	Medium	High	Total number of probes
0.5	0.23	0.66	0.10	30169
0.45	0.23	0.66	0.10	29892
0.4	0.23	0.66	0.10	29432
0.35	0.23	0.67	0.10	28775
0.3	0.22	0.67	0.11	27641
0.25	0.21	0.68	0.11	25726
0.2	0.20	0.69	0.11	22271
0.15	0.16	0.71	0.13	15869
0.1	0.10	0.73	0.17	5830
0.05	0.01	0.58	0.40	89

The coefficient of variation was assessed in detail for different subset of genes. The first series that was evaluated are so-called housekeeping genes that are used for normalizing real-time PCR data. Those genes are assumed to be stable in expression and this was assessed in our data set using the coefficient of variation. A set of 19 housekeeping or reference genes that are suggested by Roche were evaluated. The results are shown in Figure 17. The data are organized according to the 2 main factors (season and gender). In many cases, there are more than four dots (4 conditions) per gene because there are multiple probes referring to the same gene on the Agilent microarray. For example, compare TFRC gene (1 probe) with PPIA (multiple probes). It is clear from Figure 15 that even housekeeping genes can have large coefficient of variation and that the most appropriate subset needs to be selected for normalizing real-time PCR data. The gene HPRT1 for example was found to be very stable, whereas the widely used GAPDH showed considerable variation. Appropriate techniques exist for selecting multiple housekeeping genes for normalization purpose. Further elaboration on this approach is beyond the scope of this document. All genes had a coefficient of variation above the 5% threshold and this corresponds exactly to the technical reproducibility mentioned in Figure 2.

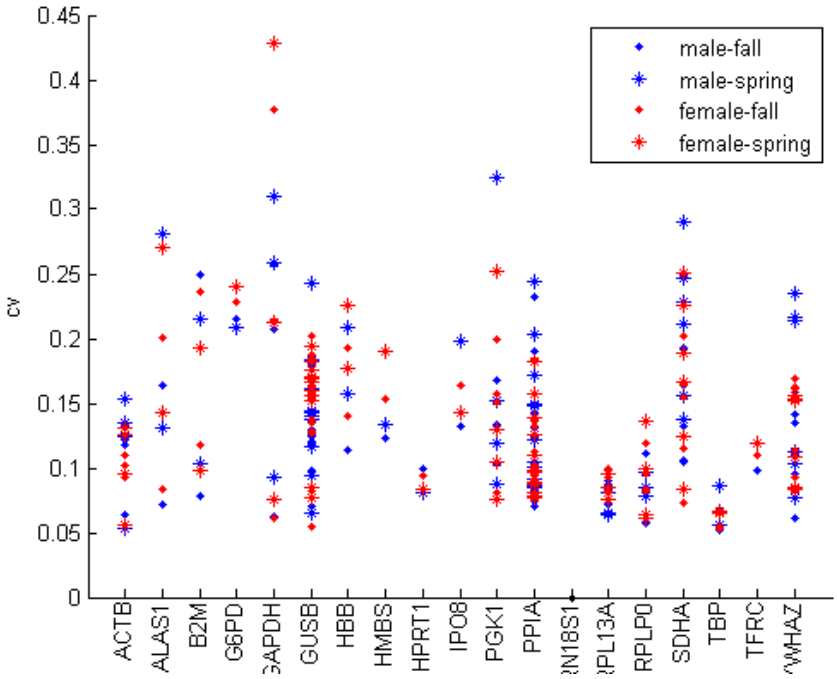


Figure 17 Coefficient of variation for 19 genes that are recommended by Roche as potential housekeeping or reference genes for real-time PCR analysis. The dots refer to the main conditions. Multiple dots per condition indicate that multiple microarray probes refer to the same gene identifier

A second subset of genes that was investigated more closely is the candidate biomarker genes that have been selected by Ghent University and Maastricht University in the frame of the first Flemish biomonitoring campaign and the additional microarray analysis that have been performed in the context of the LNE gene expression study (see WP3). A pipeline of several statistical approaches incorporating (i) p-values of regression analysis, (ii) p-values obtained by enrichment analysis using pathway analysis, (iii) expected gene expression changes based upon regression analysis of the 10th and 90th percentile population exposure values, have led to a selection of 20 candidate biomarker genes per gender that are correlated with exposure to environmental pollutants. Those genes are given in Figure 18 (males) and Figure 19 (females). Their gene description is given Annex 4. The Tables also give the coefficient of variation and the expression level of the genes according to the low, medium or high category (see higher).The candidate genes are classified in several different classes and their coefficient of variation is have a wide range of coefficient

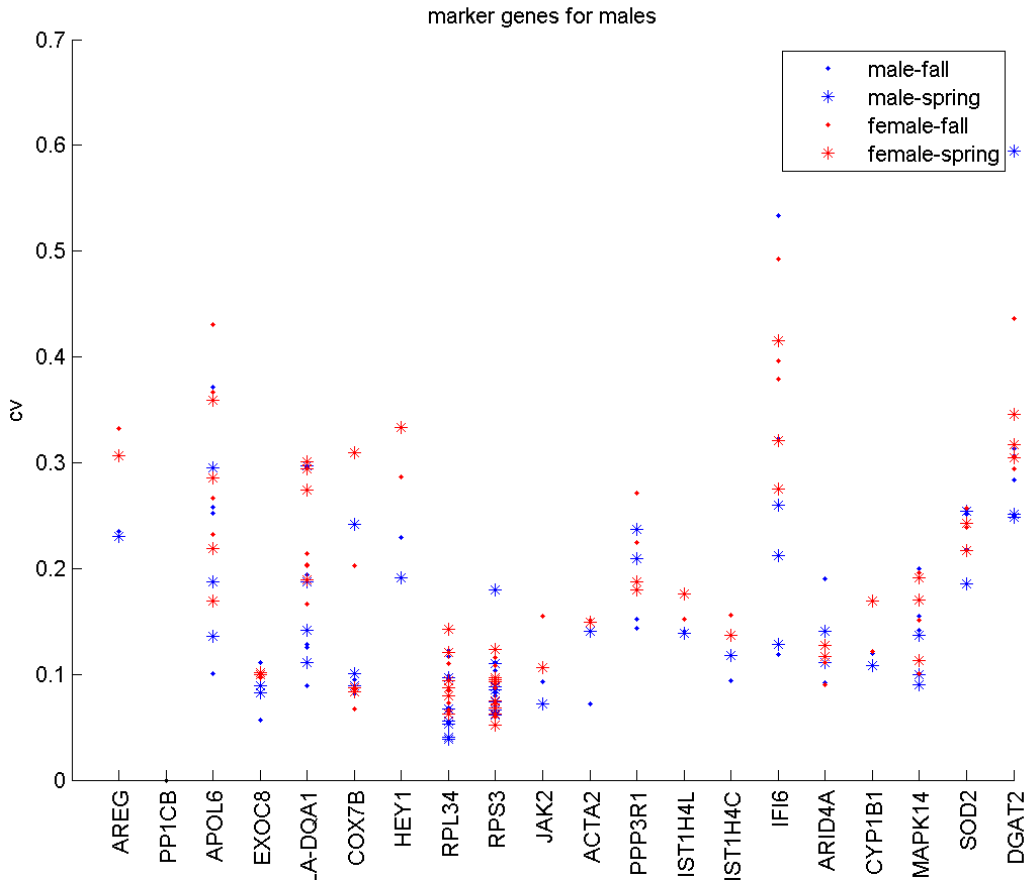


Figure 18 Coefficient of variation for 20 genes that have been selected by Maastricht University as candidate genes that can be measured as biomarker of early biological effect in male individuals in order to assess their exposure to environmental pollutants The dots refer to the main conditions. Multiple dots per condition indicate that multiple microarray probes refer to the same gene identifier

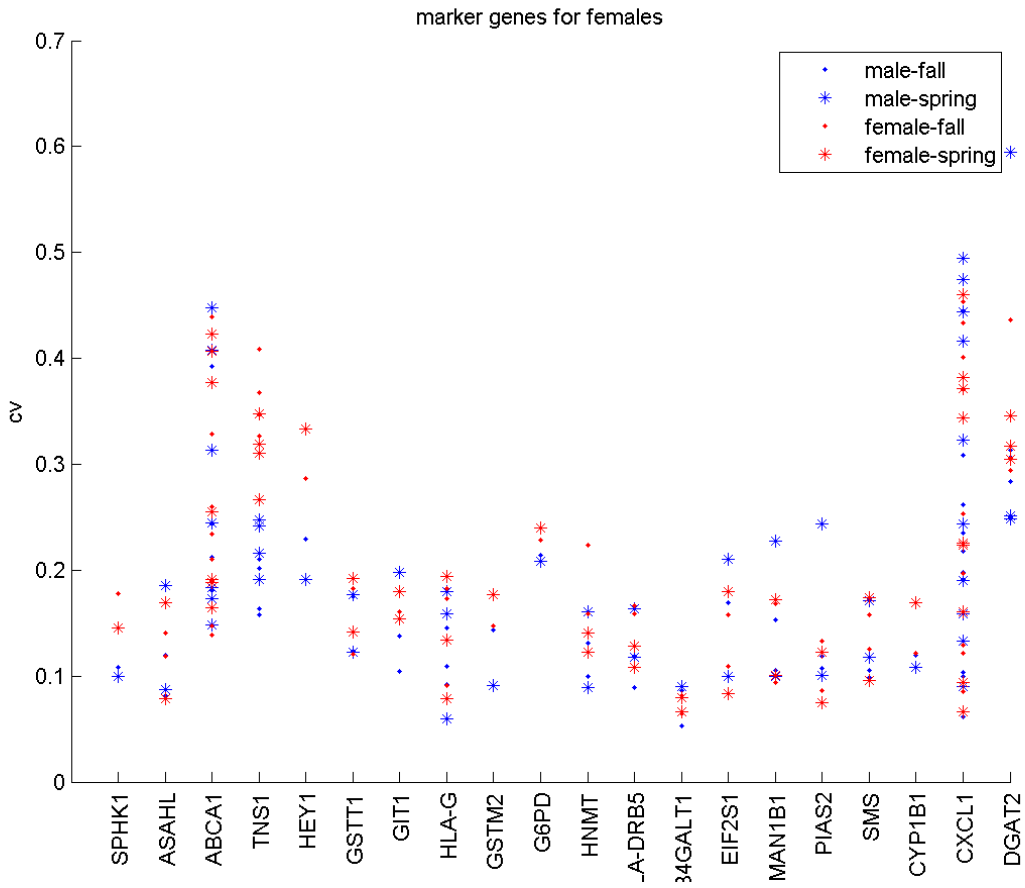


Figure 19 Coefficient of variation for 20 genes that have been selected by Maastricht University as candidate genes that can be measured as biomarker of early biological effect in female individuals in order to assess their exposure to environmental pollutants The dots refer to the main conditions. Multiple dots per condition indicate that multiple microarray probes refer to the same gene identifier

3.6 Statistical methodology

3.6.1 Framework

The study has as objectives the identification of the possible short term time trends in gene expression within a season . Analysis of differential expression between two seasons (main effect), between gender (main effect) and possible interaction. As a second series of objectives are related to the identification of genes for which the variability of the gene expression *within* subject depends of gender or season, and the identification of genes for which the variability *between* subjects depends of gender or season.

The mixed model approach has been proposed to deal with the different questions related to short term and long term variation in the data set. In a first phase it was evaluated if there was a significant time trend in the short time series (three samples per season). This step was needed for verifying that data within one season can be considered as repeated measures. The SAS Procedure MIXED was used. In every analysis, the obtained p-values were adjusted for multiplicity using either the Benjamini-Hochberg false discovery rate (FDR) correction or Q-values. A microarray probe will be called significant if its false discovery rate or q-value is less than 5%.

3.6.2 Short term (per season) analysis

The initial step in the analysis was to assess if there was a short term time effect (*i.e.* per season analysis). The analysis tests for a trend within one season and will determine if the measurements within one season can be considered as repeated measures or not. If the data within one season can be considered as repeated measures, then the mixed model can be reduced to testing 2 time points (i.e. season 1=autumn versus season 2=spring). For every season, three replicates were obtained per subject at a weekly interval. Observations from a subject are probably more alike as compared to observations from different subjects. In this regard, a mixed effects model which can model the correlation was formulated. Let Y_{ijgst} be the expression level of subject (volunteer) *i*, probename *j*, gender *g*, season *s* and trial *t*, for *i* = 1,2,…, 22, *j* = 1,2,..., 41K, *g* = 1,2, *s* = 1,2, and *t* = 1,2,3. Analysis will be done here per gene and per season. Thus dropping the *s* and *j* indices the model (1) becomes:

$$Y_{igt} = \beta_0 + \beta_1 d_{it} + \alpha_g + b_i + \varepsilon_{igt} \quad (1)$$

Y_{igt} Response for subject i, gender g, measured at time point d_{it}
β_0 Overall mean
β_1 Slope estimate for the linear time point effect
α_g Effect of female (α_1) or male (α_2)
b_i Subject I effect
ε_{igt} Random error of subject i gender g and trial t assumed to have mean 0 and variance σ_ε^2

β_0, β_1 and α_g are fixed effects while b_i is a random effect assumed to have mean 0 and variance σ_b^2

A fit of this model on the data followed by p-values adjusted for multiplicity resulted in 110 significant probes for the time effect in the autumn (first season) but none for the spring (second season). A standard pathway analysis (core analysis) was performed using Ingenuity Pathway Analysis using the 110 probes. A total of 97 probes could be mapped to known genes. The pathway results revealed a number of general pathways and biological functions that are enriched. However, these terms did not reveal a relevant biological interpretation in the frame of biomonitoring context. The season variability of these 97 genes should be taken into account if those genes also appear to be relevant in view of exposure-effect relationships. Only a small number of probes were found to have a short term time effect. For the further analysis it was therefore assumed that there was no significant short term time effect.

3.6.3 Long term analysis

- Genes with fixed gender and season effects

It was assumed that there is no short term time trend it the data set. Hence, the data have a repeated nature and a random effects model for assessing differentially expressed genes for gender and season effects can be motivated. For individual i = 1,2,…, 22, gender g = 1,2, s = 1,2, and time point per season t = 1,2,3, the following general model (2) is considered for every gene (probename):

$$Y_{igst} = \beta_0 + \tau_s + \alpha_g + (\tau\alpha)_{sg} + b_i + (b\tau)_{is} + \varepsilon_{igst} \quad (2)$$

Y_{igst}	Response for subject i, gender g, measured at time point d_{it} in a season
β_0	Overall mean
τ_s	Effect of spring (τ_1) or autumn (τ_2)
α_g	Effect of female (α_1) or male (α_2)
b_i	Subject i effect
$(b\tau)_{is}$	Season slope effect
ε_{igst}	Random error of subject i gender g, season s and trial t

β_0, τ_s and α_g are fixed effects. b_i, $(b\tau)_{is}$ and ε_{igst} are random effects with the following assumptions:

$b_i \sim N(0, \sigma^2_{b,g})$

$(b\tau)_{is} \sim N(0, \sigma^2_{a,g})$

$\varepsilon_{igst} \sim N(0, \sigma^2_{\varepsilon,g,s})$

A fit of this model on the data, followed by p-values adjustment for multiplicity, resulted in only two probes with a significant interaction effect. Therefore this interaction term was left out of the model giving rise to the following full model (3), to be used in all further analyses:

$$Y_{igst} = \beta_0 + \tau_s + \alpha_g + b_i + (b\tau)_{is} + \varepsilon_{igst} \quad (3)$$

The parameters are defined as in model (2). This model was fitted for every gene. Given that there were 41K probes and two fixed effects (season and gender) 82K different probF p-values were obtained. The model was run on individual probes and hence the raw p-values are not affected by the number of probes either 41K or the filtered 33K-list. For multiple-testing correction, we first needed to extract p-values for the 33K probes and then apply a correction on this subset. The testing was done separately for the two main effects. 191 and 1 995 probes passed the FDR p-value at 5% level of significance for the parameters gender and season, respectively. Alternatively, we also applied the less conservative q-value procedure and obtained 315 and 2 798 probes for gender and season, respectively. The threshold for selection was also 5% for the q-values. Two different methods were applied to verify the consistency of the probe selection.

The probes selected via the FDR-procedure were submitted to a Gene Ontology (GO) analysis using TopGO packages in R. Probes affected by the season term and gender term were analyzed separately. The results for the top-20 significant terms are listed in Table 5 and Table 6 (an extended list of the Top-50 is given in Annex 5). Two different analysis were performed *i.e.* classic Fisher test and the Fisher test using the weight algorithm that takes into account dependency in the GO classification tree.

Table 5 Gene Ontology analysis for gender-affected probes. Top-20 most significant biological processes are shown using TopGO. The Fisher test and two different algorithms (classic and weight) have been used to score the terms. The values shown are the uncorrected p-values per term

GO.ID	Term	Rank in classic	classic	weight
GO:0006413	translational initiation	2	4.40E-05	4.40E-05
GO:0042733	embryonic digit morphogenesis	3	0.00014	0.00014
GO:0043462	regulation of ATPase activity	4	0.0003	0.0003
GO:0006269	DNA replication, synthesis of RNA primer	5	0.00048	0.00048
GO:0051900	regulation of mitochondrial depolarizati...	6	0.00048	0.00048
GO:0007064	mitotic sister chromatid cohesion	12	0.00142	0.00142
GO:0032769	negative regulation of monooxygenase act...	13	0.00142	0.00142
GO:0030049	muscle filament sliding	17	0.00282	0.00282
GO:0007224	smoothened signaling pathway	22	0.00417	0.00417
GO:0007052	mitotic spindle organization	24	0.0052	0.0052
GO:0044419	interspecies interaction between organis...	28	0.00561	0.00561
GO:0030048	actin filament-based movement	1	4.10E-05	0.00611
GO:0000075	cell cycle checkpoint	21	0.00355	0.00738
GO:0001947	heart looping	36	0.00824	0.00824
GO:0045727	positive regulation of translation	37	0.00824	0.00824
GO:0045930	negative regulation of mitotic cell cycl...	39	0.00893	0.00893
GO:0007205	activation of protein kinase C activity ...	41	0.00964	0.00964
GO:0051384	response to glucocorticoid stimulus	42	0.00964	0.00964
GO:0045862	positive regulation of proteolysis	44	0.01037	0.01037
GO:0016568	chromatin modification	58	0.01379	0.01047

Table 6 Gene Ontology analysis for season-affected probes. Top-50 most significant biological processes are shown using TopGO. The Fisher test and two different algorithms (classic and weight) have been used to score the terms. The values shown are the uncorrected p-values per term

GO.ID	Term	Rank in classic	classic	weight
GO:0032147	activation of protein kinase activity	4	0.00011	0.00011
GO:0006793	phosphorus metabolic process	1	1.90E-05	0.00021
GO:0031122	cytoplasmic microtubule organization	10	0.0004	0.0004
GO:0030032	lamellipodium assembly	11	0.00052	0.00052
GO:0006388	tRNA splicing, via endonucleolytic cleav...	12	0.00055	0.00055
GO:0048477	oogenesis	5	0.00013	0.00064
GO:0046885	regulation of hormone biosynthetic proce...	16	0.00178	0.00178
GO:0030194	positive regulation of blood coagulation	18	0.00255	0.00255
GO:0032026	response to magnesium ion	19	0.00255	0.00255
GO:0032908	regulation of transforming growth factor...	20	0.00255	0.00255
GO:0007032	endosome organization	24	0.00275	0.00275
GO:0006298	mismatch repair	26	0.00392	0.00392
GO:0016254	preassembly of GPI anchor in ER membrane	27	0.00473	0.00473
GO:0006661	phosphatidylinositol biosynthetic proces...	32	0.00596	0.00596
GO:0051084	'de novo' posttranslational protein fold...	33	0.00646	0.00646
GO:0001936	regulation of endothelial cell prolifera...	36	0.00801	0.00801
GO:0046847	filopodium assembly	37	0.00851	0.00814
GO:0051452	intracellular pH reduction	42	0.01014	0.01014
GO:0030325	adrenal gland development	46	0.01229	0.01229
GO:0030511	positive regulation of transforming grow...	47	0.01242	0.01242

Top Networks

ID	Associated Network Functions	Score
1 View	Gene Expression, Infection Mechanism, Genetic Disorder	34
2 View	Infection Mechanism, Gene Expression, Cell-To-Cell Signaling and Interaction	32
3 View	DNA Replication, Recombination, and Repair, Cell Cycle, Cell Morphology	30
4 View	Connective Tissue Development and Function, Embryonic Development, Skeletal and Muscular Disorders	29
5 View	RNA Post-Transcriptional Modification, Cellular Assembly and Organization, Cell Death	28

Top Bio Functions

Diseases and Disorders

Name	p-value	# Molecules
Infection Mechanism	2.57E-04 - 2.21E-02	110
Infectious Disease	3.99E-04 - 4.78E-02	128
Cancer	5.10E-04 - 4.95E-02	311
Neurological Disease	5.10E-04 - 3.82E-02	29
Inflammatory Response	1.26E-03 - 2.61E-02	11

Molecular and Cellular Functions

Name	p-value	# Molecules
Gene Expression	9.52E-08 - 4.52E-02	179
Cell Cycle	5.61E-06 - 4.85E-02	119
Cell Signaling	3.29E-04 - 3.57E-02	8
Molecular Transport	3.29E-04 - 4.15E-02	27
RNA Post-Transcriptional Modification	3.29E-04 - 4.52E-02	30

Physiological System Development and Function

Name	p-value	# Molecules
Connective Tissue Development and Function	3.04E-04 - 1.37E-02	9
Cardiovascular System Development and Function	1.25E-03 - 4.83E-02	17
Embryonic Development	1.25E-03 - 3.57E-02	9
Skeletal and Muscular System Development and Function	1.25E-03 - 4.15E-02	22
Tissue Development	1.25E-03 - 4.95E-02	25

Top Canonical Pathways

Name	p-value	Ratio
Actin Cytoskeleton Signaling	8.16E-05	30/231 (0.13)
RhoA Signaling	9.36E-05	19/108 (0.176)
Integrin Signaling	6.33E-04	26/201 (0.129)
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	1.08E-03	22/189 (0.116)
Ephrin Receptor Signaling	1.6E-03	23/198 (0.116)

Figure 20 Ingenuity Pathway Analysis summary report for genes that are found to be significantly affected (FDR p-value<0.05) by the factor season using a random-effects model. The top-5 scored biological networks, functions and canonical pathways are shown

Top Networks

ID	Associated Network Functions	Score
1 View	Cellular Assembly and Organization, Cardiovascular System Development and Function, Organ Morphology	32
2 View	Drug Metabolism, Molecular Transport, Small Molecule Biochemistry	26
3 View	Gene Expression, Cellular Growth and Proliferation, Tumor Morphology	23
4 View	Cellular Assembly and Organization, DNA Replication, Recombination, and Repair, Carbohydrate Metabolism	19
5 View	Protein Synthesis, Skeletal and Muscular System Development and Function, Post-Translational Modification	17

Top Bio Functions

Diseases and Disorders

Name	p-value	# Molecules
Cancer	1.65E-04 - 4.52E-02	33
Hepatic System Disease	1.65E-04 - 1.65E-04	10
Inflammatory Disease	4.94E-03 - 4.57E-02	6
Renal and Urological Disease	4.94E-03 - 4.57E-02	5
Skeletal and Muscular Disorders	6.25E-03 - 1.98E-02	9

Molecular and Cellular Functions

Name	p-value	# Molecules
Cellular Assembly and Organization	4.38E-05 - 4.57E-02	20
Cell Cycle	9.54E-05 - 4.57E-02	14
Cell Death	9.84E-05 - 4.85E-02	10
Cell Morphology	4.32E-04 - 4.57E-02	11
Gene Expression	6.45E-04 - 4.36E-02	8

Physiological System Development and Function

Name	p-value	# Molecules
Skeletal and Muscular System Development and Function	1.53E-03 - 4.57E-02	5
Tissue Morphology	1.53E-03 - 3.28E-02	5
Cardiovascular System Development and Function	3.78E-03 - 4.21E-02	6
Tissue Development	3.78E-03 - 2.63E-02	3
Embryonic Development	6.08E-03 - 3.28E-02	6

Top Canonical Pathways

Name	p-value	Ratio
VEGF Signaling	2.58E-05	6/100 (0.06)
Pancreatic Adenocarcinoma Signaling	7.29E-04	5/120 (0.042)
PTEN Signaling	7.29E-04	5/123 (0.041)
Prostate Cancer Signaling	2.16E-03	4/98 (0.041)
Apoptosis Signaling	2.26E-03	4/88 (0.045)

Figure 21 Ingenuity Pathway Analysis summary report for genes that are found to be significantly affected (FDR p-value<0.05) by the factor gender using a random-effects model. The top-5 scored biological networks, functions and canonical pathways are shown

A Gene Ontology or GO-analysis aims at aggregating gene lists at a biological level. Combining individual genes with a similar function in the same biological bag should facilitate the data interpretation for the biologist. The different biological functions or bags are scored using statistical criteria leading to a list of statistically enriched biological terms associated with the original gene lists. However, the interpretation at this biological level is often complicated as many biological functions and interactions are still unknown. Genes can be involved in different biological functions and hence these different biological terms are interlinked. Table 6 and 7 give an overview of the most significant biological terms that are obtained for genes that are affected by the factors gender and season, respectively. No biological comprehensive insight is revealed and this suggests that this type of GO-analysis does not any robust biological systems that may be affected in the context of our experimental setup. However, it may be useful to do a manual curation with experts in the biomonitoring field as specific (but interesting) phenomena might be missed using a global screening.

Biological analysis and pathway analysis can be done with different open-source or commercial packages each with their own functionalities and reporting options. For WP2 we had the commercial package Ingenuity Pathway Analysis available for an analysis complementary to GO-analysis. Summaries of core analysis are shown in Figure 20 and Figure 21. The biological networks are constructed with the gene lists as a starting point. A high score points indicates a relevant biological interactions that are discovered based on the information databases that are used by Ingenuity Pathway Analysis. The description of the networks is quite generic and more detailed biological investigation by biomonitoring experts should be done to evaluate the significance of the observations. Furthermore, IPA uses a proprietary ontology of Bio functions (comparable to a gene ontology classification). The genes are binned in Bio functions and scored with a statistical approach. The most significantly enriched terms are mentioned in Figure 20 and 21. Many functions and terms related to infection and immune system are appearing, but the exact biological meaning is unclear at this moment. The enrichment for diseases like cancer does not imply that the analysis discovered a true relationship with cancer development. Many genes such as transcription factors, cell cycle genes, and cytokines are promiscuous of nature and are involved in many biological processes and functions. Most of the genes that are linked to the cancer category are many of these general genes. The enrichment for the term cancer is a pure statistical process and the biological meaning of the observation should be carefully evaluated before a conclusion can be made.

The list with genes that are influenced by the factor season in the random effects model was overlaid with the candidate biomarker genes that have been identified earlier by the joint effort between Ghent University and Maastricht University. The results are shown in a Venn diagram in Figure 22. There known overlap for three genes (HEY1, CYP1B1, and DGAT2) between the male biomarker list and the female biomarker list is seen. There was also overlap with the season-affected gene list. Five genes were identified for males (AREG, COX7B, PPP3R1, HIST1H4L, and HIST1H4C) and two for females (HLA-G, and HNMT).

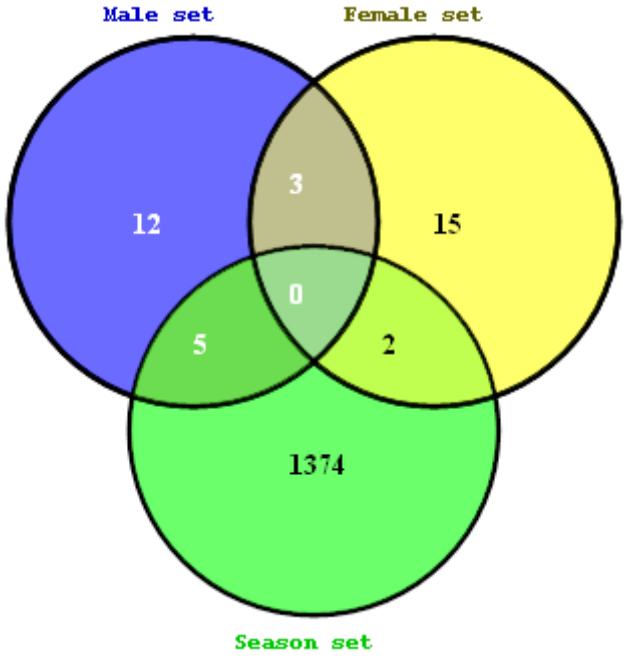


Figure 22 Venn diagram of three gene lists; i) 20 candidate male biomarkers (Male set), ii) 20 candidate female biomarkers (Female set), and iii) the genes that were found to be significantly affected by the factor season using the random effects model used in the work package

The other research objectives involve the variance-covariance structure. It is worth saying that the final mean structure is as assumed in model (3). This model will hence-forth be called the full model. Various hypotheses were built and tested. The testing was based on the Likelihood ratio test where the -2loglikelihood from a reduced model was compared with that of the full model for every probename. We allowed negative values for the σ^2 parameters in order not to put any prior restrictions on the variance and covariance structure for the outcomes Y_{igst} .

· Genes for which the variance of measurement error depends on gender and season

It was of interest to investigate if there were genes for which the variability *within* subjects depends of gender and season. In order words, which are the genes for which the following null hypothesis will be rejected:

$$H_0 : \sigma^2_{\varepsilon,1,1} = \sigma^2_{\varepsilon,2,1} = \sigma^2_{\varepsilon,1,2} = \sigma^2_{\varepsilon,2,2} \quad (4)$$

The effects of the model are as in model (3) except that $\varepsilon_{igst} \sim N(0, \sigma^2_{\varepsilon})$
Hypothesis (4) can be tested by comparing the -2LogLikelihood of model (3). It was jointly tested if the variance of the measurement error was the same for males and females for every season. Using a 3 degree of freedom likelihood ratio test, the obtained p-values were adjusted for multiplicity and 10 057 different probes were found to be significant or in other words 10 057 probenames were found to reject hypothesis (4). In hypothesis (4), we jointly test if the variance of the measurement error was the same for males and females and the same for both seasons. There could be probes for which the difference could only be for gender or for season. In this regard, the 10 057 probenames found to reject hypothesis (4) were then assessed as follows:

· Genes for which the variance of measurement error depends on gender

Identification of genes for which the variance of measurement error depends on gender implies testing hypothesis (5). We have used model (3) for this with the assumption that $\varepsilon_{igst} \sim N(0, \sigma^2_{\varepsilon,S})$

$$H_0 : \sigma^2_{\varepsilon,1,1} = \sigma^2_{\varepsilon,2,1}, \sigma^2_{\varepsilon,1,2} = \sigma^2_{\varepsilon,2,2} \quad (5)$$

In order to test (5), a 2 degree of freedom likelihood ratio test was used.
A total of 7 533 of the 10 057 probes were found to be significant.

- Genes for which the variance of measurement error depends on season

The following hypothesis (6) was tested. The assumption was that $\varepsilon_{igst} \sim N(0, \sigma^2_{\varepsilon,g})$

$$H_0 : \sigma^2_{\varepsilon,1,1} = \sigma^2_{\varepsilon,1,2}, \sigma^2_{\varepsilon,2,1} = \sigma^2_{\varepsilon,2,2} \quad (6)$$

In order to test (6), a 2 degree of freedom likelihood ratio test was used.
A total of 8 559 of the 10 057 probes were found to be significant.

- Genes for which the variability between subjects depends on gender and/or season

The null hypothesis to be tested is:

$$H_0 : \sigma^2_{a,1} = 0, \sigma^2_{a,2} = 0, \sigma^2_{b,1} = \sigma^2_{b,2} \quad (7)$$

Based on a 3 degree of freedom likelihood ratio test, only 25 probes were found to reject (7). These probes could be mapped to 23 genes and the details can be found in Table 7.

These are: A_23_P148541, A_23_P22696, A_23_P25945, A_23_P351295, A_23_P360534, A_23_P40574, A_23_P40919, A_23_P46309, A_23_P74059, A_23_P7901, A_23_P97541, A_24_P103893, A_24_P140475, A_24_P174316, A_24_P229234, A_24_P295543, A_24_P367602, A_24_P418408, A_24_P674479, A_24_P917254, A_24_P99066, A_32_P125832, A_32_P223256, A_32_P41396, A_32_P68479.

Hypothesis (7) is a joint test and it may be interesting to know for which of these 25 genes the variability between subjects depends only on season or gender. These genes were therefore reassessed as follows:

The null hypothesis to be tested for seasonal effect is (8):

$$H_0 : \sigma^2_{a,1} = 0, \sigma^2_{a,2} = 0 \quad (8)$$

A 2 degree of freedom test was used and 9 of the 25 probes were found to reject hypothesis (8). These are A_23_P351295, A_32_P125832, A_32_P68479, A_24_P917254, A_24_P367602, A_23_P7901, A_32_P223256, A_32_P41396, A_23_P46309.

The null hypothesis to be tested for gender effect is (9):

$$H_0 : \sigma^2_{a,1} = \sigma^2_{a,2}, \sigma^2_{b,1} = \sigma^2_{b,2} \quad (9)$$

A 2 degree of freedom likelihood ratio test was used and only 2 (A_24_P917254, A_23_P351295) of the 25 probes were found not to reject hypothesis (9). The other 23 rejected hypothesis (9).

Table 7 Annotation of the 23 genes for which the variability between subjects depends on gender and/or season

Probe number	Gene description	Gene Symbol
A_23_P148541	cancer/testis antigen 1A	CTAG1A
A_23_P22696	cancer/testis antigen 2	CTAG2
A_23_P25945	aarF domain containing kinase 1	ADCK1
A_23_P351295	heparan sulfate (glucosamine) 3-O-sulfotransferase 5	HS3ST5
A_23_P360534	chromosome 18 open reading frame 2	C18orf2
A_23_P40574	crystallin, beta B2	CRYBB2
A_23_P40919	G protein-coupled receptor 128	GPR128
A_23_P46309	SNHG3-RCC1 readthrough transcript	SNHG3-RCC1
A_23_P74059	natriuretic peptide precursor A	NPPA
A_23_P7901	tubulin tyrosine ligase-like family, member 2	TTLL2
A_23_P97541	complement component 4 binding protein, alpha	C4BPA
A_24_P103893	heparanase 2	HPSE2
A_24_P140475	sorbin and SH3 domain containing 2	SORBS2
A_24_P174316	KIAA1310	KIAA1310
A_24_P229234	myosin XVI	MYO16
A_24_P295543	biogenesis of lysosomal organelles complex-1, subunit 2	BLOC1S2
A_24_P367602	dual specificity phosphatase 5 pseudogene	DUSP5P
A_24_P418408	family with sequence similarity 89, member A	FAM89A
A_24_P674479	family with sequence similarity 122C	FAM122C
A_24_P99066	ring finger protein 17	RNF17
A_32_P125832	hypothetical protein LOC100128893	LOC100128893
A_32_P41396	fibroblast growth factor 13	FGF13

CONCLUSIONS AND RECOMMENDATIONS

- Detailed investigation of PubMed literature indicated that the blood gene expression levels and gene expression variability in the normal population using genome-wide transcriptomics is not well documented. Only small-scale studies have been performed using different collection techniques and microarray technologies. The probe design of those arrays is at present outdated. Comparison between those studies is difficult and it was not possible to get easily access to the data for own statistical analysis.

- Considering the effort that is being undertaken in Flemish Environment & Health (E&H) studies to introduce gene expression analysis for identifying biomarkers of exposure and early effect that probe for exposure to environmental pollutants, it was considered valuable to define gene expression variability in the normal population. Particular attention focus was put on the impact of gender and season (time-dependent variation) on gene expression.

- A pilot study has been set up with 22 healthy volunteers (equal distribution between males and females) with an age between 20-40. This population group is relevant for E&H in Flanders. Six blood samples have been collected per individual. Three samples in the Fall of the year 2009 and three samples during Spring 2010. Samples within one season allowed calculating basic statistical parameters. Long term variability and seasonal effect was estimated by sampling in two different periods.

- Blood samples have been processed and high quality RNA was obtained for microarray analysis. The total RNA was depleted with globin mRNA (non-informative mRNA that is a major part of the total RNA) to increase the detection of low intense signals. The effect of the depletion step was not tested in this project. However, numerous reports exist about the value of this globin-removal step for different microarray platforms (Affymetrix, Agilent, and Illumina).

- Agilent 4×44K human microarrays have been successfully used in WP2 and a raw data set of 132 high quality microarrays has been collected. A single-colour (Cy3-labeling) strategy has been used to allow measuring the absolute signal intensity of the probes; Furthermore, such strategy allows maximal flexibility in terms of statistical analysis.

- Many different approaches exist for treatment of raw microarray data. There is no consensus on the best strategy, but it is important to agree on the same pipeline when several research labs are collaborating. The more stringent Agilent QC filtering that is used by Maastricht University showed that about 8 000 probes (of the 41 000) were flagged. These probes were also identified as being expressed at low level in the data set of WP2. A priori filtering of non-informative probes can be recommended for future projects as this eliminates noisy data. It also has a positive impact on the multiple testing procedures because less correction needs to be done (about 20% less probes needs to be tested).

- The Spearman correlation for the overall microarray profile identified that the within-subject samples were tightly clustered over time. This is indicative for the fact that blood gene expression levels are quite constant within a subject.

- Descriptive statistics have been described for the two main factors (gender and season) that were under scrutiny in WP2. There is considerable spread in gene expression intensity, and standard deviation when one considers all probes on an Agilent array that pass the QC controls. Moreover, the standard deviation fluctuates with the signal intensity. This aspect has important consequences for future power calculations as one needs to make a distinction between the absolute probe (or gene) expression for calculating the number of individuals that are needed for identifying a specified effect size in case-control study (*e.g.* low exposed versus high exposed individuals).

- The coefficient of variation fluctuates, with 75% of the probes having a coefficient of variation lower than 0.25 in both seasons. We also identified between 40 to 60 probes that have very a high coefficient of variation in one season and a very low coefficient in the other season. Their gene description has been retrieved and their involvement in biological processes and pathways has been listed. It is unclear at present what the true biological meaning is of our findings.

- The coefficient of variation has also been extracted for two particular gene panels : i) housekeeping genes, and ii) biomarker gene sets that have been identified by Maastricht University and Ghent University. The genes showed large difference in their coefficient of variation in our data set when the data were grouped according to the main factors gender and season. The coefficient of variation was also highly dependent on the probe that was used to identify the gene. It is recommended that when a candidate biomarker is identified, one should check the level of expression of the different probes (if applicably), and the expression level for each probe.

- A statistical model has been developed for testing the impact of gender and season in the data set of WP2. Mixed models analysis per gene did not reveal significant short-term (or within season) effects. A total number of 97 mapped genes showed a short-term time trend in the autumn, but not in spring time. Biological interpretation of the data revealed no clear biological pathways or biological functions that are affected. The observed time trend in autumn might be important if those genes will be considered for use in biomonitoring programs. This time trend needs to be confirmed in a larger population before a conclusion can be drawn.

- Next, a random effects model was used to study the effect of the two main factors season and gender. No interaction effect was observed. After multiplicity correction, we identified 110 probes and 1 995 probes for which the signal intensity was affected by gender and season, respectively. Many different biological terms and biological networks have been retrieved for the two gene lists. Pathway analysis is a complex field because of the many different tools, databases, algorithms, dependency of the data, etc. It is beyond the scope to embark on this issue within the timeframe of this project.

- The lists of candidate biomarker genes (20 for males and 20 for females) identified earlier by Maastricht University and Ghent University were overlaid with the season and gender gene lists from the random effects model. Respectively five and two genes from the male and female list showed a season effect. It is suggested that the impact of this confounding effect on the gene signature profile should be investigated more in detail.

- Finally, the variance measures within the variance-covariance matrix was considered. Different likelihood ratio tests identified that for about 8 000 probes variance was dependent of season or gender. The between-subject effect on variance was limited to 25 probes and those have been described. The biological reason why those genes were affected has not been studied.

- WP2 generated valuable information on the gene expression levels and the variability in function of time for a subset of the normal population. Separate gene expression studies are aiming at the discovery of genes and gene fingerprints that correlate with exposure complex measurements (*e.g.* WP3 of this project). Those studies can identify genes that are differentially expressed between a group of people that received a high exposure versus a group with low exposure. These studies have usually one measurement for each individual and in this way the within subject (or time-dependent) variation cannot be captured. it is possible that differentially expressed genes (with a small effect size) might be false-positives because their time-dependent variability is a stronger effect than the treatment effect or exposure. It is envisioned that a systematic comparison of these differentially expressed genes with the information we have obtained from the variability study may help to identify false positive marker genes. One of the first steps that will be useful in the process of valorisation of the gene expression variability data is to store raw and processed data per gene in such a way that they can be consulted efficiently. Furthermore, criteria should be established to relate gene expression variability data to gene expression data obtained in the frame of exposure evaluations. A next step should consist of a fusion between these different datasets but this should be evaluated by experts in biostatistics.

REFERENCE LIST

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.

Amundson SA, Do KT, Shahab S, Bittner M, Meltzer P, Trent J, et al. 2000. Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. Radiat Res 154: 342-346.

Arikawa E, Sun Y, Wang J, Zhou Q, Ning B, Dial SL, et al. 2008. Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. BMC Genomics 9: 328.

Asare AL, Kolchinsky SA, Gao Z, Wang R, Raddassi K, Bourcier K, et al. 2008. Differential gene expression profiles are dependent upon method of peripheral blood collection and RNA isolation. BMC Genomics 9: 474.

Baird AE. 2006. The blood option: transcriptional profiling in clinical trials. Pharmacogenomics 7: 141-144.

Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, et al. 2005. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proc Natl Acad Sci U S A 102: 11023-11028.

Bouwens M, Afman LA, Müller M. 2007. Fasting induces changes in peripheral blood mononuclear cell gene expression profiles related to increases in fatty acid beta-oxidation: functional role of peroxisome proliferator activated receptor alpha in human peripheral blood mononuclear cells. Am J Clin Nutr 86: 1515-1523.

Eady JJ, Wortley GM, Wormstone YM, Hughes JC, Astley SB, Foxall RJ, et al. 2005. Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. Physiol Genomics 22: 402-411.

Feezor RJ, Baker HV, Mindrinos M, Hayden D, Tannahill CL, Brownstein BH, et al. 2004. Whole blood and leukocyte RNA isolation for gene expression analyses. Physiol Genomics 19: 247-254.

Field LA, Jordan RM, Hadix JA, Dunn MA, Shriver CD, Ellsworth RE, et al. 2007. Functional identity of genes detectable in expression profiling assays following globin mRNA reduction of peripheral blood samples. Clin Biochem 40: 499-502.

Forrest MS, Lan Q, Hubbard AE, Zhang L, Vermeulen R, Zhao X, et al. 2005. Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. Environ Health Perspect 113: 801-807.

Karlovich C, Duchateau-Nguyen G, Johnson A, McLoughlin P, Navarro M, Fleurbaey C, et al. 2009. A longitudinal study of gene expression in healthy individuals. BMC Med Genomics 2: 33.

Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7: 673-679.

Kim C, Paik S. 2010. Gene-expression-based prognostic assays for breast cancer. Nat Rev Clin Oncol 7: 340-347.

Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, et al. 2004. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol Biomarkers Prev 13: 445-453.

Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. 2006. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J Lab Clin Med 147: 126-132.

Liu J, Walter E, Stenger D, Thach D. 2006. Effects of globin mRNA reduction methods on gene expression profiles from whole blood. J Mol Diagn 8: 551-558.

Meaburn EL, Fernandes C, Craig IW, Plomin R, Schalkwyk LC. 2009. Assessing individual differences in genome-wide gene expression in human whole blood: reliability over four hours and stability over 10 months. Twin Res Hum Genet 12: 372-380.

Mohr S, Liew CC. 2007. The peripheral-blood transcriptome: new insights into disease and risk assessment. Trends Mol Med 13: 422-432.

Palmer C, Diehn M, Alizadeh AA, Brown PO. 2006. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC Genomics 7: 115.

Patino WD, Mian OY, Kang JG, Matoba S, Bartlett LD, Holbrook B, et al. 2005. Circulating transcriptome reveals markers of atherosclerosis. Proc Natl Acad Sci U S A 102: 3423-3428.

Peretz A, Peck EC, Bammler TK, Beyer RP, Sullivan JH, Trenga CA, et al. 2007. Diesel exhaust inhalation and assessment of peripheral blood mononuclear cell gene transcription effects: an exploratory study of healthy human volunteers. Inhal Toxicol 19: 1107-1119.

Radich JP, Mao M, Stepaniants S, Biery M, Castle J, Ward T, et al. 2004. Individual-specific variation of gene expression in peripheral blood leukocytes. Genomics 83: 980-988.

Rockett JC, Burczynski ME, Fornace AJ, Herrmann PC, Krawetz SA, Dix DJ. 2004. Surrogate tissue analysis: monitoring toxicant exposure and health status of inaccessible tissues through the analysis of accessible tissues and cells. Toxicol Appl Pharmacol 194: 189-199.

Sullivan PF, Fan C, Perou CM. 2006. Evaluating the comparability of gene expression in blood and brain. Am J Med Genet B Neuropsychiatr Genet 141B: 261-268.

Tian Z, Palmer N, Schmid P, Yao H, Galdzicki M, Berger B, et al. 2009. A practical platform for blood biomarker study by using global gene expression profiling of peripheral whole blood. PLoS One 4: e5157.

Valk PJ, Verhaak RG, Beijen MA, Erpelinck CA, Barjesteh van Waalwijk van Doorn-Khosrovani, Boer JM, et al. 2004. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 350: 1617-1628.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009.

van Erk MJ, Blom WA, van Ommen B, Hendriks HF. 2006. High-protein and high-carbohydrate breakfasts differentially change the transcriptome of human blood cells. Am J Clin Nutr 84: 1233-1241.

van Leeuwen DM, Gottschalk RW, van Herwijnen MH, Moonen EJ, Kleinjans JC, van Delft JH. 2005. Differential gene expression in human peripheral blood mononuclear cells induced by cigarette smoke and its constituents. Toxicol Sci 86: 200-210.

van Leeuwen DM, van Herwijnen MH, Pedersen M, Knudsen LE, Kirsch-Volders M, Sram RJ, et al. 2006. Genome-wide differential gene expression in children exposed to air pollution in the Czech Republic. Mutat Res 600: 12-22.

Wang Z, Neuburg D, Li C, Su L, Kim JY, Chen JC, et al. 2005. Global gene expression profiling in whole-blood samples from individuals exposed to metal fumes. Environ Health Perspect 113: 233-241.

Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, et al. 2003. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100: 1896-1901.

ANNEX 1 QUESTIONNAIRE USED FOR THE LNE-GENE EXPRESSION STUDY

ONDERZOEK IN HET KADER VAN EEN OPDRACHT VAN DE VLAAMSE OVERHEID NAAR DE TIJDSGEBONDEN GENEXPRESSIE VARIATIE IN HUMAAN BLOED

Studie van de dienst Toxicologie van de Vlaamse Instelling voor Technologisch Onderzoek (Vito) in samenwerking met het Provinciaal Instituut voor Hygiëne en het Studiecentrum voor Kernenergie

VRAGENLIJST

Gegevens uit deze vragenlijst dienen om informatie te bekomen over uw gezondheidstoestand en blootstelling van de periode vlak voor de staalname.

Datum van vandaag: ……/……./……..
Labelnummer:

GEZONDHEID

1 Hoe is uw algemene gezondheidstoestand nu?

- ☐ zeer goed
- ☐ goed
- ☐ gaat wel (redelijk)
- ☐ slecht
- ☐ zeer slecht

2 Bent u de laatste 7 dagen ziek geweest? o ja o nee

3 Indien ja, welke ziekte(n) heeft u de laatste 7 dagen gehad?

- ☐ verkoudheid
- ☐ griep
- ☐ keelontsteking
- ☐ oorontsteking
- ☐ andere: _____

4 Heeft u momenteel gezondheidsproblemen? Indien ja welke?

- ☐ hoge bloedddruk
- ☐ astma
- ☐ allergie
- ☐ migraine of ernstige hoofdpijn
- ☐ andere: _____

5 Heeft u de laatste 24 h medicatie genomen? o ja, welke_____
o nee

6 Voelt u zich vermoeid? o ja o nee

7 Hoeveel uur heeft u deze nacht geslapen? ___uren

8 Heeft u last van stress? o ja o nee

ENKEL VOOR VROUWEN

9 Wanneer zijn uw laatste maandstonden begonnen? datum_____

BLOOTSTELLING GEDURENDE DE VOORBIJE DAG

10 Eet u vegetarisch? o ja o nee

11 Heeft u de laatste 24 h vis gegeten? o ja o nee

12 Heeft u de afgelopen 24 h vlees gegeten? o ja o nee

13 Heeft u de laatste 24 h alcoholhoudende dranken genuttigd? o ja o nee
___glazen bier
___glazen wijn
___glazen sterk alcoholische drank

14 Hoeveel koppen koffie dronk u de laatste 24 h? ___koppen

15 Heeft u de laatste 24 h voedingssupplementen genomen (vitamines, ijzer,;…)?
o ja o nee

Noteer de namen van de producten._____

16 Gebruikte u de laatste 24 h nicotine in de vorm van pleisters, kauwgom, spray's?
o ja
o nee

17 Hoeveel uren heeft u de laatste 24 h doorgebracht in een ruimte waar gerookt werd door anderen?
___uren

18 Heeft u gedurende de laatste 24 h lichamelijke inspanningen verricht waarbij u bezweet geraakte?
o ja
o nee

19 Kwam u de laatste 24 uur in contact met de vermelde producten?

scheikundige producten in het laboratorium	o ja o nee
pesticiden, houtbewaarmiddelen	o ja o nee
kleurstoffen, verven of oplosmiddelen (thinner, in lijm,…)	o ja o nee
haarverzorgingsproducten (niet: shampoo of lotions, wel: bleekmiddelen, kleuringsproducten,…)	o ja o nee

20 Hoe intens nam u deel aan het verkeer gedurende de voorbije 24 h?

Hoeveel minuten zat u in een auto ,autobus of tram (gemiddeld) ___minuten

Hoeveel minuten hiervan zat u in een auto, autobus of tram die in een file stond (gemiddeld) ___minuten

Hoeveel minuten fietste u of ging u te voet langs straten met een (vrij) druk verkeer (gemiddeld) ___minuten

21 In welke omgeving is uw huis gelegen?
o landelijke regio
o stad, stedelijke rand

Bedankt voor het invullen van de vragenlijst !

ANNEX 2 APPROVAL

Universiteit Antwerpen
Faculteit Geneeskunde

UA CDE - S2 23A - UNIVERSITEITSPLEIN,1 B2610 WILRIJK
VITO
Dr. Ir. Patrick De Boever
Afdeling Toxicologie
Retieseweg
2440 Geel

commissie medische ethiek
campus drie eiken – lokaal S2.23
Universiteitsplein 1 B-2610 Antwerpen(Wilrijk)

T +32(0)3 820 25 40
F +32(0)3 820 25 01
kristin.deby@ua.ac.be
www.ua.ac.be

UW KENMERK

ONS KENMERK
PC-KD

DATUM
9 oktober 2009

BIJLAGE

Betreft: **Beslissing commissie medische ethiek UA A09 21**
gelieve bij briefwisseling steeds het dossiernummer te vermelden

Geachte,

De commissie medische ethiek UA heeft tijdens de vergadering van 5 oktober 2009 uw project: "Tijdsgebonden variatie in genexpressie in humaan perifeer bloed" **goedgekeurd.**

De leden willen nog hetvolgende opmerken:

- Het bijgevoegde verzekeringsattest is nog steeds niet de foutloze aansprakelijkheidsverzekering die art 29 van de wet van 29/5/2004 voorschrijft.
- De leden stellen voor in de informatiebrief *expressie* te vervangen door *uitdrukking*.
- Er zijn verschillende spellingfouten in de informatiebrief, aangeduid in het document in bijlage.

Met vriendelijke groeten,

Prof. Dr. P. Cras
Ondervoorzitter

ANNEX 3 TABLES OF GENES THAT WERE DESIGNATED EXTREME FOR THE DIFFERENT CONDITIONS BASED ON THE FACTORS (GENDER AND SEASON). THE SELECTION WAS DONE BASED ON A COEFFICIENT OF VARIATION FOR THE MICROARRAY PROBES LARGER THAN 2

Table 1 Extreme genes observed for male individuals in Fall

Gene description	Gene symbol
helicase with zinc finger	HELZ
Yes-associated protein 1, 65kDa	YAP1
ral guanine nucleotide dissociation stimulator-like 1	RGL1
BCL2-antagonist/killer 1	BAK1
major facilitator superfamily domain containing 9	MFSD9
transmembrane protein 176B	TMEM176B
Sec61 alpha 2 subunit (S. cerevisiae)	SEC61A2
dehydrodolichyl diphosphate synthase	DHDDS
protein phosphatase 1, regulatory (inhibitor) subunit 13B	PPP1R13B
ankyrin repeat and KH domain containing 1	ANKHD1
X-prolyl aminopeptidase (aminopeptidase P) 1, soluble	XPNPEP1
chromosome X open reading frame 24	CXorf24
chromosome 1 open reading frame 163	C1orf163
zinc finger protein 644	ZNF644
recombination activating gene 1	RAG1
NDC80 homolog, kinetochore complex component (S. cerevisiae)	NDC80
zinc finger protein 138	ZNF138
MAM domain containing 4	MAMDC4
pantothenate kinase 2	PANK2
solute carrier family 25 (mitochondrial carrier; ornithine transporter) member 2	SLC25A2
family with sequence similarity 73, member B	FAM73B
MAP7 domain containing 3	MAP7D3
protein tyrosine phosphatase type IVA, member 1	PTP4A1
phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	GART
family with sequence similarity 131, member A	FAM131A
chromosome 6 open reading frame 162	C6orf162
BMS1 homolog, ribosome assembly protein (yeast) pseudogene	LOC96610
similar to programmed cell death 2	LOC728739
anaphase promoting complex subunit 13	ANAPC13
zinc finger protein 143	ZNF143
chromosome 1 open reading frame 104	C1orf104
ribosomal L1 domain containing 1	RSL1D1

Table 2 Extreme genes observed for male individuals in Spring

Gene description	Gene symbol
CKLF-like MARVEL transmembrane domain containing 5	CMTM5
lectin, galactoside-binding, soluble, 7B	LGALS7B
family with sequence similarity 35, member A	FAM35A
hook homolog 2 (Drosophila)	HOOK2
ATP/GTP binding protein-like 5	AGBL5
zinc finger protein 600	ZNF600
major histocompatibility complex, class I-related	MR1
GINS complex subunit 1 (Psf1 homolog)	GINS1
SET domain containing 5	SETD5
sorting nexin 9	SNX9
sialophorin	SPN
agouti related protein homolog (mouse)	AGRP
tumor necrosis factor (ligand) superfamily, member 9	TNFSF9
small proline-rich protein 1A	SPRR1A
family with sequence similarity 26, member F	FAM26F
guanylate binding protein 2, interferon-inducible	GBP2
ubiquitin-like modifier activating enzyme 6	UBA6
dual specificity phosphatase 5 pseudogene	DUSP5P
guanidinoacetate N-methyltransferase	GAMT
methionyl aminopeptidase 1	METAP1
DEAH (Asp-Glu-Ala-His) box polypeptide 34	DHX34
lectin, galactoside-binding, soluble, 7	LGALS7
variable charge, X-linked 3A	VCX3A
cullin 1	CUL1
interleukin 10 receptor, beta	IL10RB
required for meiotic nuclear division 1 homolog (S. cerevisiae)	RMND1
lectin, galactoside-binding, soluble, 7B	LGALS7B
death-associated protein kinase 1	DAPK1
bone morphogenetic protein 1	BMP1
solute carrier family 9 (sodium/hydrogen exchanger), member 1	SLC9A1
RRN3 RNA polymerase I transcription factor homolog (S. cerevisiae) pseudogene	LOC730092
immunoglobulin heavy constant gamma 1 (G1m marker)	IGHG1
signal sequence receptor, alpha	SSR1

Table 3 Extreme genes observed for female individuals in Fall

Gene description	Gene symbol
cysteine-rich secretory protein LCCL domain containing 2	CRISPLD2
ral guanine nucleotide dissociation stimulator-like 1	RGL1
ATPase, Ca++ transporting, type 2C, member 2	ATP2C2
diaphanous homolog 3 (Drosophila)	DIAPH3
cytochrome P450, family 1, subfamily A, polypeptide 1	CYP1A1
dystrobrevin, alpha	DTNA
solute carrier family 39 (zinc transporter), member 1	SLC39A1
eukaryotic translation initiation factor 2-alpha kinase 1	EIF2AK1
CD177 molecule	CD177
RNA binding motif protein 11	RBM11
arylsulfatase D	ARSD
ADAM metallopeptidase domain 19 (meltrin beta)	ADAM19
transcription factor 7-like 2 (T-cell specific, HMG-box)	TCF7L2
ALS2 C-terminal like	ALS2CL
glycine-N-acyltransferase-like 2	GLYATL2
lactate dehydrogenase C	LDHC
eukaryotic translation initiation factor 4B	EIF4B
phytanoyl-CoA dioxygenase domain containing 1	PHYHD1
NCK-associated protein 1	NCKAP1
ubiquitin D	UBD
slingshot homolog 1 (Drosophila)	SSH1
ubiquitin domain containing 2	UBTD2
G protein-coupled receptor 120	GPR120
transmembrane protein 86A	TMEM86A
osteoclast stimulating factor 1	OSTF1
acyl-Coenzyme A dehydrogenase family, member 10	ACAD10
retinoblastoma binding protein 5	RBBP5
eukaryotic translation initiation factor 4E	EIF4E
RAN binding protein 2	RANBP2
chemokine (C-C motif) receptor 3	CCR3
immunoglobulin lambda locus	IGL@
FAM13A opposite strand (non-protein coding)	FAM13AOS
Fas (TNFRSF6) binding factor 1	FBF1
COX11 homolog, cytochrome c oxidase assembly protein (yeast)	COX11
G-2 and S-phase expressed 1	GTSE1
zinc finger protein 175	ZNF175
hect domain and RLD 4	HERC4
RNA binding motif protein 43	RBM43

Table 4 Extreme genes observed for female individuals in Spring

Gene description	Gene symbol
Duffy blood group, chemokine receptor	DARC
C2 calcium-dependent domain containing 3	C2CD3
KIAA0528	KIAA0528
chromosome 21 open reading frame 15	C21orf15
membrane-spanning 4-domains, subfamily A, member 3 (hematopoietic cell-specific)	MS4A3
hedgehog interacting protein	HHIP
inositol hexakisphosphate kinase 3	IP6K3
POM (POM121 homolog, rat) and ZP3 fusion	POMZP3
cell division cycle 37 homolog (S. cerevisiae)-like 1	CDC37L1
deleted in lymphocytic leukemia 2 (non-protein coding)	DLEU2
chromosome 1 open reading frame 217	C1orf217
molybdenum cofactor synthesis 3	MOCS3
zinc finger protein 547	ZNF547
oxoeicosanoid (OXE) receptor 1	OXER1
Smith-Magenis syndrome chromosome region, candidate 5	SMCR5
NADPH oxidase organizer 1	NOXO1
SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1	SMARCA1
glycine-N-acyltransferase-like 2	GLYATL2
B double prime 1, subunit of RNA polymerase III transcription initiation factor IIIB	BDP1
DNA cross-link repair 1C (PSO2 homolog, S. cerevisiae)	DCLRE1C
ATPase, H+/K+ transporting, nongastric, alpha polypeptide	ATP12A
regenerating islet-derived 1 alpha	REG1A
transcription factor CP2	TFCP2
INO80 complex subunit C	INO80C
guanidinoacetate N-methyltransferase	GAMT
syntaxin binding protein 1	STXBP1
family with sequence similarity 69, member A	FAM69A
coiled-coil domain containing 45	CCDC45
transmembrane protein 161B	TMEM161B
zinc finger protein 530	ZNF530
centromere protein I	CENPI
FSHD region gene 2	FRG2
PTPRF interacting protein, binding protein 1 (liprin beta 1)	PPFIBP1
phosphatidylinositol transfer protein, alpha	PITPNA
golgi associated, ARF binding protein 2	GGA2
coiled-coil domain containing 101 pseudogene	LOC388242
hypothetical protein LOC285708	LOC285708
phosphatidylinositol-5-phosphate 4-kinase, type II, gamma	PIP4K2C
TIP41, TOR signaling pathway regulator-like (S. cerevisiae)	TIPRL

tectonic family member 3	TCTN3

ANNEX 4 OVERVIEW OF THE 20 MARKER GENES FOR MALE AND FEMALE INDIVIDUALS, RESPECTIVELY. THE GENES HAVE BEEN SELECTED BY JOINT EFFORT BETWEEN MAASTRICHT UNIVERSITY AND GHENT UNIVERSITY ANS ARE CANDIDATES FOR EVALUATING EXPOSURE TO ENVIRONMENTAL POLLUTION IN THE GENERAL POPULATION

Table 5 Overview of the 20 candidate marker genes for male individuals. Gene symbol, gene description, class of the coefficient of variation over both seasons (CV) and the category of average gene expression over both season. There are three categories: LOW: A-value<6, MEDIUM: <6A<12, and HIGH: A>12

Gene symbol	Gene description	CV	Expression
EXOC8	exocyst complex component 8	0.1	MEDIUM
COX7B	cytochrome c oxidase subunit VIIb	0.1	MEDIUM/HIGH
RPL34	ribosomal protein L34	0.1	HIGH
RPS3	ribosomal protein S3	0.1	HIGH
ARID4A	AT rich interactive domain 4A (RBP1-like)	0.1	MEDIUM
APOL6	apolipoprotein L, 6	0.2	MEDIUM
JAK2	Janus kinase 2 (a protein tyrosine kinase)	0.2	MEDIUM
HIST1H4C	histone cluster 1, H4c	0.2	HIGH
MAPK14	mitogen-activated protein kinase 14	0.2	MEDIUM/HIGH
PPP3R1	protein phosphatase 3 (formerly 2B), regulatory subunit B, alpha isoform	0.25	MEDIUM
HIST1H4L	histone cluster 1, H4l	0.25	HIGH
DGAT2	diacylglycerol O-acyltransferase homolog 2 (mouse)	0.25	MEDIUM
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1	0.3	MEDIUM
PP1CB	catalytic subunit of protein phosphatase 1	0.4	MEDIUM
IFI6	interferon, alpha-inducible protein 6	0.4	MEDIUM
SOD2	superoxide dismutase 2, mitochondrial	0.4	MEDIUM
ACTA2	actin, alpha 2, smooth muscle, aorta	0.45	MEDIUM
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	0.5	MEDIUM
HEY1	hairy/enhancer-of-split related with YRPW motif 1	> 0.50	
AREG	amphiregulin (schwannoma-derived growth factor)	> 0.50	

Table 6 Overview of the 20 candidate marker genes for female individuals. Gene symbol, gene description, class of the coefficient of variation over both seasons (CV) and the category of average gene expression over both season. There are three categories: LOW: A-value<6, MEDIUM: <6A<12, and HIGH: A>12

Gene symbol	Gene description	CV	Expression
SPHK1	sphingosine kinase 1	0.3	MEDIUM
G6PD	glucose-6-phosphate dehydrogenase	0.1	MEDIUM
MAN1B1	mannosidase, alpha, class 1B, member 1	0.1	MEDIUM
PIAS2	protein inhibitor of activated STAT, 2	0.1	MEDIUM
SMS	spermine synthase	0.1	MEDIUM
HLA-G	major histocompatibility complex, class I, G	0.15	HIGH
B4GALT1	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 1	0.15	HIGH
EIF2S1	eukaryotic translation initiation factor 2, subunit 1 alpha, 35kDa	0.15	MEDIUM
GSTT1	glutathione S-transferase theta 1	0.2	LOW
HNMT	histamine N-methyltransferase	0.2	MEDIUM
CXCL1	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	0.25	MEDIUM
ASAHL	N-acylethanolamine acid amidase	0.3	MEDIUM/HIGH
HLA-DRB5	major histocompatibility complex, class II, DR beta 5	0.3	HIGH
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	0.35	LOW/MEDIUM
GSTM2	glutathione S-transferase M2 (muscle)	0.35	MEDIUM
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1	0.35	MEDIUM
DGAT2	diacylglycerol O-acyltransferase homolog 2 (mouse)	0.35	MEDIUM
GIT1	G protein-coupled receptor kinase interactor 1	0.4	LOW
TNS1	tensin 1	> 0.50	
HEY1	hairy/enhancer-of-split related with YRPW motif 1	> 0.50	

ANNEX 5 TABLES OF THE TOP-50 GENE ONTOLOGY TERMS THAT ARE SIGNIFICANTLY ENRICHED (P<0.05) USING FISHER EXACT TEST AND TWO DIFFERENT SCORING ALGORITHMS

Table 5 Gene Ontology analysis for gender-affected probes. Top-50 most significant biological processes are shown using TopGO. The Fisher test and two different algorithms (classic and weight) have been used to score the terms. The values shown are the uncorrected p-values per term (column 4 and 5). The Rank of the term obtained with the weight algorithm is compared to its rank in the classic algorithm (column 3). GO identification code and short description are given in column 1 and 2)

GO.ID	Term	Rank in classic	classic	weight
GO:0006413	translational initiation	2	4.40E-05	4.40E-05
GO:0042733	embryonic digit morphogenesis	3	0.00014	0.00014
GO:0043462	regulation of ATPase activity	4	0.0003	0.0003
GO:0006269	DNA replication, synthesis of RNA primer	5	0.00048	0.00048
GO:0051900	regulation of mitochondrial depolarizati...	6	0.00048	0.00048
GO:0007064	mitotic sister chromatid cohesion	12	0.00142	0.00142
GO:0032769	negative regulation of monooxygenase act...	13	0.00142	0.00142
GO:0030049	muscle filament sliding	17	0.00282	0.00282
GO:0007224	smoothened signaling pathway	22	0.00417	0.00417
GO:0007052	mitotic spindle organization	24	0.0052	0.0052
GO:0044419	interspecies interaction between organis...	28	0.00561	0.00561
GO:0030048	actin filament-based movement	1	4.10E-05	0.00611
GO:0000075	cell cycle checkpoint	21	0.00355	0.00738
GO:0001947	heart looping	36	0.00824	0.00824
GO:0045727	positive regulation of translation	37	0.00824	0.00824
GO:0045930	negative regulation of mitotic cell cycl...	39	0.00893	0.00893
GO:0007205	activation of protein kinase C activity ...	41	0.00964	0.00964
GO:0051384	response to glucocorticoid stimulus	42	0.00964	0.00964
GO:0045862	positive regulation of proteolysis	44	0.01037	0.01037
GO:0016568	chromatin modification	58	0.01379	0.01047
GO:0055010	ventricular cardiac muscle morphogenesis	49	0.01271	0.01271
GO:0042493	response to drug	50	0.01321	0.01321
GO:0001841	neural tube formation	52	0.01353	0.01353
GO:0007368	determination of left/right symmetry	53	0.01353	0.01353
GO:0060048	cardiac muscle contraction	54	0.01353	0.01353
GO:0009314	response to radiation	20	0.00333	0.01532
GO:0008633	activation of pro-apoptotic gene product...	61	0.018	0.018
GO:0042632	cholesterol homeostasis	62	0.018	0.018

GO:0030258	lipid modification	73	0.02295	0.02295
GO:0002326	B cell lineage commitment	80	0.02829	0.02829
GO:0010523	negative regulation of calcium ion trans...	81	0.02829	0.02829
GO:0009791	post-embryonic development	87	0.0285	0.0285

Table 5 Continued

GO:0009953	dorsal/ventral pattern formation	88	0.0285	0.0285
GO:0043524	negative regulation of neuron apoptosis	90	0.03084	0.03084
GO:0030301	cholesterol transport	92	0.03203	0.03203
GO:0006534	cysteine metabolic process	96	0.03385	0.03385
GO:0031103	axon regeneration	97	0.03385	0.03385
GO:0033033	negative regulation of myeloid cell apop...	98	0.03385	0.03385
GO:0034380	high-density lipoprotein particle assemb...	99	0.03385	0.03385
GO:0042989	sequestering of actin monomers	100	0.03385	0.03385
GO:0045634	regulation of melanocyte differentiation	101	0.03385	0.03385
GO:0060556	regulation of vitamin D biosynthetic pro...	102	0.03385	0.03385
GO:0001658	branching involved in ureteric bud morph...	109	0.03939	0.03939
GO:0003065	positive regulation of heart rate by epi...	110	0.03939	0.03939
GO:0010224	response to UV-B	111	0.03939	0.03939
GO:0032469	endoplasmic reticulum calcium ion homeos...	112	0.03939	0.03939
GO:0043330	response to exogenous dsRNA	113	0.03939	0.03939
GO:0046668	regulation of retinal cell programmed ce...	114	0.03939	0.03939
GO:0002320	lymphoid progenitor cell differentiation	124	0.04489	0.04489
GO:0010894	negative regulation of steroid biosynthe...	125	0.04489	0.04489

Table 6 Gene Ontology analysis for season-affected probes. Top-50 most significant biological processes are shown using TopGO. The Fisher test and two different algorithms (classic and weight) have been used to score the terms. The values shown are the uncorrected p-values per term (column 4 and 5). The Rank of the term obtained with the weight algorithm is compared to its rank in the classic algorithm (column 3). GO identification code and short description are given in column 1 and 2)

GO.ID	Term	Rank in classic	classic	weight
GO:0032147	activation of protein kinase activity	4	0.00011	0.00011
GO:0006793	phosphorus metabolic process	1	1.90E-05	0.00021
GO:0031122	cytoplasmic microtubule organization	10	0.0004	0.0004
GO:0030032	lamellipodium assembly	11	0.00052	0.00052
GO:0006388	tRNA splicing, via endonucleolytic cleav...	12	0.00055	0.00055
GO:0048477	oogenesis	5	0.00013	0.00064
GO:0046885	regulation of hormone biosynthetic proce...	16	0.00178	0.00178
GO:0030194	positive regulation of blood coagulation	18	0.00255	0.00255
GO:0032026	response to magnesium ion	19	0.00255	0.00255
GO:0032908	regulation of transforming growth factor...	20	0.00255	0.00255
GO:0007032	endosome organization	24	0.00275	0.00275
GO:0006298	mismatch repair	26	0.00392	0.00392
GO:0016254	preassembly of GPI anchor in ER membrane	27	0.00473	0.00473
GO:0006661	phosphatidylinositol biosynthetic proces...	32	0.00596	0.00596
GO:0051084	'de novo' posttranslational protein fold...	33	0.00646	0.00646
GO:0001936	regulation of endothelial cell prolifera...	36	0.00801	0.00801
GO:0046847	filopodium assembly	37	0.00851	0.00814
GO:0051452	intracellular pH reduction	42	0.01014	0.01014
GO:0030325	adrenal gland development	46	0.01229	0.01229
GO:0030511	positive regulation of transforming grow...	47	0.01242	0.01242
GO:0045022	early endosome to late endosome transpor...	58	0.01776	0.01776
GO:0007159	leukocyte adhesion	61	0.01894	0.01894
GO:0042554	superoxide anion generation	68	0.02205	0.02205
GO:0006729	tetrahydrobiopterin biosynthetic process	71	0.02386	0.02386
GO:0022614	membrane to membrane docking	72	0.02386	0.02386
GO:0031053	primary microRNA processing	73	0.02386	0.02386
GO:0050817	coagulation	51	0.01281	0.02637
GO:0032313	regulation of Rab GTPase activity	82	0.02776	0.02776
GO:0001756	somitogenesis	87	0.02906	0.02906
GO:0002690	positive regulation of leukocyte chemota...	88	0.03124	0.03124
GO:0020027	hemoglobin metabolic process	89	0.03124	0.03124
GO:0030219	megakaryocyte differentiation	90	0.03124	0.03124
GO:0043353	enucleate erythrocyte differentiation	91	0.03124	0.03124
GO:0000910	cytokinesis	41	0.0094	0.03145

Table 6 Continued

GO:0006397	mRNA processing	39	0.00923	0.03221
GO:0043966	histone H3 acetylation	96	0.03239	0.03239
GO:0006944	membrane fusion	97	0.03335	0.03335
GO:0051402	neuron apoptosis	98	0.03335	0.03335
GO:0001556	oocyte maturation	103	0.03765	0.03765
GO:0002002	regulation of angiotensin levels in bloo...	104	0.03765	0.03765
GO:0002246	healing during inflammatory response	105	0.03765	0.03765
GO:0006704	glucocorticoid biosynthetic process	106	0.03765	0.03765
GO:0014805	smooth muscle adaptation	107	0.03765	0.03765
GO:0032909	regulation of transforming growth factor...	108	0.03765	0.03765
GO:0040036	regulation of fibroblast growth factor r...	109	0.03765	0.03765
GO:0042109	lymphotoxin A biosynthetic process	110	0.03765	0.03765
GO:0042789	mRNA transcription from RNA polymerase I...	111	0.03765	0.03765
GO:0045628	regulation of T-helper 2 cell differenti...	112	0.03765	0.03765
GO:0050665	hydrogen peroxide biosynthetic process	113	0.03765	0.03765
GO:0001841	neural tube formation	54	0.01407	0.03828

WP3: Exposure-effect relations

1. INTRODUCTION

The first Flemish Environment and Health Survey or FLEHS (2002-2006) aimed to measure internal exposure to pollutants in areas differing in pollution burden and to assess whether place of residence or observed differences in internal concentrations of pollutants were associated with biological health effects. In this study, it was shown that gene expression research in biomonitoring studies of exposed populations leads to relevant results (http://www.milieu-en-gezondheid.be/; Van Leeuwen *et al.* 2008). In the study the expression of eight cancer-related genes, i.e. CYP1B1, SOD2, ATF4, MAPK14, CXCL1, PINK1, DGAT2 and TIGD3, were investigated in 398 adults from the Flemish population of "elderly" (age 50-65 years). The expression was determined by quantitative PCR and by which the relationship with several biomarkers for exposure and effect was investigated. The results were positive, i.e. relations were found between gene expression results on the one hand and exposure and effect markers on the other hand and thereby confirmed the usefulness of gene expression analysis in population studies. At the same time the limited number of genes studied raises the question whether important effects on the gene expression level have been missed. Therefore, in the current research a subpopulation (n=100) from the previous conducted study within the scope of "Steunpunt Milieu & Gezondheid" (n=398 for the gene expression analysis based on RT-PCR) was used to obtain global gene expression profiles (all active genes) by using DNA microarray technology.

The general aim of this study is to investigate the relationship between gene expression profiles and biomarkers of exposure and effect of pollutants at exposure levels that are relevant for the general population in Flanders. For this study the data of exposure and effect markers obtained in the study of the 1st generation Steunpunt Milieu & Gezondheid will be used as starting point. Pollutant and dose dependant gene expression changes were analysed on the level of individual genes as well as on the level of genetic networks/pathways in order to obtain more insight on the biological relevance of these modifications. Furthermore, associations with important effect markers, i.e. tumor markers, micronuclei frequencies, COMET signals and oxidative DNA damage, also obtained from the 1st generation Steunpunt Milieu & Gezondheid, were analysed.
Finally, the correspondence between the gene expression data obtained from the microarray analysis with those from the RT-PCR data of the 1st generation Steunpunt Milieu & Gezondheid was established. From previous research within the Steunpunt it was shown that gender specific effects are to be expected and thus the results were analyzed for the complete test-population as well as for males and females separately.

2. METHODS

The selection of the adults, the measurement of the pollutants and biological effect markers and expression analysis in the study of the 1st generation Steunpunt Milieu & Gezondheid (http://www.milieu-en-gezondheid.be/; Van Leeuwen *et al.* 2008) will be shortly described, followed by the methods of the current study.

2.1 Selection of the test-population

The original study population consisted of 398 subjects from eight different regions of residence in Flanders with different environmental burdens. Criteria for selection were age 50–65 years and living in their region of residence >5 years. From these adults a subpopulation 100 non-smoking adults were selected. Table 1 provides an overview of the number of non-smoking men and women, divided per age class and high and low exposure (based on z-score; see below), from which RNA was isolated in the previous study of the Steunpunt and now analyzed through microarray technology.

Table 1: Number of men and women with high (H) and low (L) exposure (based on z-scores) per age category.

		Men		Women		
		H	L	H	L	
Age	50-55	7	8	8	8	**31**
	55-60	8	8	8	8	**32**
	60-65	9	9	9	10	**37**
		24	**25**	**25**	**26**	

The men and women were selected based on their previously measured exposure levels of multiple pollutants or their metabolites. The levels of these pollutants/metabolites were measured in whole blood, serum, or urine by various methods: heavy metals (cadmium (Cd) and lead (Pb)) in whole blood; dioxins and furans (CALUX), *p,p´*-dichlorodiphenyldichloroethylene (*p,p´*-DDE) and non-dioxin-like polychlorinated biphenyls (PCBs; 138+153+180) in serum; 1-OH-pyrene (a metabolite of polycyclic aromatic hydrocarbons (PAHs)) and *t,t*-muconic acid (ttMA; a metabolite of benzene) in urine (see De Coster *et al.* 2008). The integral exposure is examined based on the sum of the z-scores of each pollutant. For each subject, an index of internal exposure (*I*ex) was calculated, defined as the arithmetic mean of the *z*-scores for concentrations of urinary cadmium, blood lead, serum marker PCBs (PCB 138+153+180), serum dioxin-like activity (CALUX assay), urinary t,t'-muconic acid (marker of benzene exposure) and urinary 1-hydroxypyrene (marker of PAH exposure): [$I\text{ex} = z_{\text{urinary cadmium}} + z_{\text{blood lead}} + z_{\text{PCBs}} + z_{\text{dioxin-like activity}} + z_{\text{urinary t-t-muconic acid}} + z_{\text{urinary 1-hydroxypyrene}} + z_{\text{urinary 1-hydroxypyrene}}$], where $z = (X-\mu)/\sigma$, and x is the raw score to be standardized, μ is the mean of the population, and σ is the standard deviation of the population. We chose to compare the 50 highest and 50 lowest exposed individuals (according to the index I_{ex}) evenly distributed over sex- and age-categories, in order to maximize the power of our analysis of the influence of exposure to environmental pollutants on gene expression levels. Using this approach a group of high exposure and a group of low exposure (average of pollutant content) was determined (see table 1).

2.2 Biologiscal effect markers

The biological effect markers consist of several tumor markers (prostate specific antigen), carcinoembryonic antigen (CEA) and p53, a biomarker of chromosomal damage (micronuclei), a biomarker of DNA damage (COMET assay) and a biomarker of oxidatiove DNA damage (8-hydroxydeoxyguanosine, HDG). Details on methods of measurement can be found in De Coster et al. (2008).

2.3 RNA isolation and globine reductie

Total RNA was isolated from whole blood from PAXgene Blood RNA vacutainers using the PAXgene Blood RNA system (PreAnalytix, Qiagen, Hilden, Germany), according to the manufacturer's instructions.
A globin reduction assay was performed in order to remove a-specific hemoglobin mRNA. The GLOBINclear™ Kit by Ambion (Austin, USA) was used according to the manufacturer's instructions. After the reduction, the RNA samples were again submitted to integrity and purity measurement. RNA integrity was assessed using the BioAnalyzer (Agilent, Palo Alto, USA) and purity was measured spectrophotometrically.

2.4 RT-PCR

The expression of eight cancer-related genes, i.e. CYP1B1, SOD2, ATF4, MAPK14, CXCL1, PINK1, DGAT2 and TIGD3, was determined by quantitative PCR in the 100 adults and their relationship with several biomarkers for exposure and effect was investigated. Table 2 shows an overview of the eight cancer-related genes.

Table 2: Overview of genes used in RT-PCR (adopted from *Leeuwen et al.* 2008)

Gene name (abbreviation)[a]	GenBank accession no.[a]	Biological summary[a]	Primers
Cytochrome P450 1B1 (*CYP1B1*)	NM_000104	Catalysis of many reactions involved in drug and xenobiotic metabolism (e.g., metabolism of procarcinogens)	5´-AGTGCAGGCAGAATTGGATCA-3´ (forward) 5´-GCGCATGGCTTCATAAAGGA-3´ (reverse)
Activating transcription factor 4 (*ATF4*)	NM_001675	Encodes a transcription factor that belongs to a family of DNA-binding proteins, including the AP-1 and CREB families	5´-CTCCAGCGACAAGGCTAAGG-3´ (forward) 5´-GTTGTTGGAGGGACTGACCAA-3´ (reverse)
Superoxide dismutase 2 (*SOD2*)	NM_000636	Associated with oxidative stress; converts superoxide to hydrogen peroxide and diatomic oxygen	5´-ATCAGGATCCACTGCAAGGAA-3´ (forward) 5´-CGTGCTCCCACACATCAATC-3´ (reverse)
Mitogen-activated protein kinase 14 (*MAPK14*)	NM_001315	Activated by various environmental stressors and proinflammatory cytokines; integration point for multiple biochemical signals and involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation, and development	5´-TGAAGACTGTGAGCTGAAGATTCTG-3´ (forward) 5´-CCACGTAGCCTGTCATTTCATC-3´ (reverse)
Chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha) (*CXCL1*)	NM_001511	Regulates cell trafficking of various types of leukocytes and has a role in development, homeostasis, and function of the immune system	5´-CCACTGCGCCCAAACC-3´ (forward) 5´-GCAGGATTGAGGCAAGCTTT-3´ (reverse)
PTEN-induced putative kinase-1 (*PINK1*)	NM_032409	Encodes a serine/threonine protein kinase that localizes to mitochondria; it is thought to protect cells from stress-induced mitochondrial dysfunction	5´-AGCAGTCACTTACAGAAAATCCAAGA-3´ (forward) 5´-GGTGAAGGCGCGGAGAA-3´ (reverse)
Diacylglycerol *O*-acyltransferase homolog 2 (mouse) (*DGAT2*)	NM_032564	Responsible for triglyceride synthesis	5´-GCACAGAGGCCACAGAAGTG-3´ (forward) 5´-CCCTCAACACAGGCATTCG-3´ (reverse)
Tigger transposable element derived 3 (*TIGD3*)	NM_145719	Belongs to the tigger subfamily of the pogo superfamily of DNA-mediated transposons in humans; exact function of gene is not known	5´-GTGCTGGAACTCCTGGATGAG-3´ (forward) 5´-TTGCAGATGCGCGAGATCT-3´ (reverse)

[a]National Center for Biotechnology Information (2008).

2.5 Microarray preparation and hybridization

0.2 microgram of total RNA from each samples was used to synthesize dye-labeled cRNA (Cy3) following the Agilent one-color Quick-Amp labeling protocol (Agilent Technologies). Individual samples were hybridized on Agilent 4x44K Whole Human Genome microarrays (i.e. 100 arrays on 25 Agilent 4x44K slides).

2.6 Data analysis

Microarrays were scanned on an Agilent Microarray Scanner (Agilent Technologies, Amstelveen, The Netherlands). Raw data on pixel intensities were extracted from the scan images using Agilent Feature Extraction Software (Agilent). Raw data were pre-processed using an in house developed quality control pipeline in R as follows: local background correction, flagging of bad spots, controls and spots with too low intensity, log2 transformation, quantile normalization. From the processed data files genes were omitted showing more than 30% flagged data per group (low and high exposed separately), after which the data files were transferred to the Gene Expression Pattern Analysis Suite, GEPAS (GEPAS 2010; Montaner et al. 2006) for further pre-processing, including merging replicates (based on average), and imputing missing values by means of K-nearest neighbor imputation (K=15). Filtering for flat peaks was used with root mean square value 0.25.
The filtered data, containing 28634 genes will be further used for statistical analyses:

T-test:
The t-test was used to analyze the microarray data for identifying significantly ($p<0.05$) modulated genes between the high and low exposed men or women.

CLEAR-test (Valls et al. 2008):
The CLEAR-test method was also used to analyze the microarray data for identifying differentially expressed genes between the high and low exposed individuals. This method combines the *z*-test, which focuses on large changes, with a χ^2 test to evaluate variability, and has the advantage of not reporting genes with small changes and low variances as differentially expressed.

Correlation analyses:
Pearson correlation and regression analysis of associations of gene expression levels with biomarkers of exposure (pollutants) and effect (tumor markers, chromosome- and DNA-damage) was performed using GEPAS and EXCEL (Microsoft, Seattle, WA, USA).

Pathway analysis:
For the pathway analysis the differentially expressed genes and correlated genes from the various analyses were uploaded onto MetaCore (GeneGo, San Diego, CA) for identifying the involvement of these genes in specific cellular pathway maps by overrepresentation analyses compared to the total amount of objects involved in the particular maps. In the analysis the filtered data set was used as background list for the various gene lists. Pathway maps with a *p* value <0.05 were considered significantly modulated.

RT-PCR vs Microarray:
The gene expression changes obtained from the microarray analysis were compared with those from the RT-PCR of the 1st generation Steunpunt Milieu & Gezondheid through Pearson correlation analysis in STATISTICA (Statsoft, Tulsa, OK, USA).

3. RESULTS & DISCUSSION

3.1 Exposure markers

The following table gives an overview of the exposure markers for the selected men and women. Significant differences between the individual pollutant values of the high and low exposed groups were tested for men and women, separately, using a t-test.

Table 3: Concentrations of environmental pollutants of high and low exposed individuals for men and women.

Men		High exposure			Low exposure			
		n	average	SD	n	average	SD	p-value t-test
Cd urine	µg/g crt	24	0.67	0.57	25	0.48	0.23	0.1425
Pb blood	µg/L	24	56.48	32.42	25	34.38	11.59	0.0024
marker PCBs	ng/g lipid	24	474.97	204.18	25	326.74	65.31	0.0012
PCB118	ng/g lipid	24	27.21	14.65	25	29.94	12.55	0.4871
p,p'-DDE	ng/g lipid	24	588.83	359.97	25	618.89	837.75	0.8720
Hexachlorobenzene	ng/g lipid	24	49.79	22.05	25	46.26	16.86	0.5305
TEQ	pg/g lipid	23	37.04	28.61	22	12.91	9.36	0.0005
ttMA (~Benzene)	mg/g crt	21	0.16	0.16	23	0.08	0.07	0.0447
OH-pyr (~PAHs)	µg/g crt	24	0.40	0.61	25	0.13	0.10	0.0276

Women		High exposure			Low exposure			
		n	average	SD	n	average	SD	p-value t-test
Cd urine	µg/g crt	25	1.07	0.47	26	0.51	0.16	<0.0001
Pb blood	µg/L	25	54.63	21.00	26	28.36	13.99	<0.0001
marker PCBs	ng/g lipid	25	399.40	156.31	26	324.37	69.67	0.0304
PCB118	ng/g lipid	25	34.23	14.76	26	32.04	12.42	0.5676
p,p'-DDE	ng/g lipid	25	661.71	537.03	26	740.41	531.07	0.6011
Hexachlorobenzene	ng/g lipid	25	76.60	37.21	26	91.69	39.56	0.1673
TEQ	pg/g lipid	25	33.97	22.62	25	11.89	9.81	<0.0001
ttMA (~Benzene)	mg/g crt	24	0.23	0.19	22	0.08	0.07	0.0006
OH-pyr (~PAHs)	µg/g crt	25	0.40	0.52	26	0.15	0.20	0.0240

3.2 T-test analysis

Table 4 gives an overview of the total number of significantly modulated genes for the high and low exposed men and women as well as for all high and low exposed adults. The "fraction" indicates the number of significantly modulated genes in relation to the total number of genes on the array after filtering (28634 genes).

Table 4: Number of genes differentially expressed in t-test between high and low exposed individuals for the three datasets.

	UP	DOWN	TOTAL	fraction
Men	320	50	370	1.3%
Women	902	2016	2918	10.2%
All	621	576	1197	4.2%

The huge difference quantitatively speaking, i.e. the number of differentially expressed genes, between men and women is evident. This difference could be gender specific, as men and women react differently to the various pollutants. This was also observed in the previous research within the Steunpunt.
Choosing a more stringent evaluation, in which only effects are taken into account that exceed the level of coincidence (5%), resulted in statistically significant differences between high and low exposed women only.
Furthermore, the differentially expressed genes for women with high exposure levels are mostly down-regulated, whereas for men with high exposure levels the differentially expressed genes are up-regulated.

3.3 CLEAR-test analysis

The result of the CLEAR-test analysis, based on four different aspects, i.e. differential expression, variability, outliers, and sample heterogeneity, is shown in Table 5. Divided over four categories (differentially expressed, differentially expressed highly variability, highly variability and none significant) only the differentially expressed genes (with and without high variance) were used in the pathway analysis.
Similar as with the t-test analysis more differentially expressed genes were found for women.

Table 5: Results of the CLEAR-test between high and low exposed individuals showing the number of genes in the different categories.

	Men	Women	All
diff.expr	22	920	297
diff.expr.high.var	115	665	566
high.var	2314	1473	2050
non.significant	26183	25576	25721

3.4 Correlation analysis

The results of the correlation analyses between the gene expression and the exposure and effect markers in men (n=49), women (n=51) and all individuals (n=100) are shown in Table 6. The number of positive and negative correlations is indicated and for comparison the t-test results are also shown. The "fraction" indicates the number of significantly modulated genes in relation to the total number of genes on the array after filtering (28634 genes).

Table 6: Number of genes significantly correlating with the exposure and effect markers.

		Mannen					Vrouwen					Alle				
		n	UP	DOWN	TOTAL	fraction	n	UP	DOWN	TOTAL	fraction	n	UP	DOWN	TOTAL	fraction
Exposures	t-test HvsL	49	320	50	370	1.3%	51	902	2016	2918	10.2%	100	621	576	1197	4.2%
	Cd	49	5053	9735	14788	51.6%	51	958	834	1792	6.3%	100	3399	5733	9132	31.9%
	Pb	49	137	356	493	1.7%	51	831	865	1696	5.9%	100	501	516	1017	3.6%
	PCBs	49	216	77	293	1.0%	51	775	1217	1992	7.0%	100	199	179	378	1.3%
	PCB118	49	118	280	398	1.4%	51	635	1115	1750	6.1%	100	198	778	976	3.4%
	DDE	49	442	230	672	2.3%	51	163	295	458	1.6%	100	187	140	327	1.1%
	HCB	49	954	1170	2124	7.4%	51	821	707	1528	5.3%	100	374	476	850	3.0%
	TEQ	45	710	247	957	3.3%	50	429	670	1099	3.8%	95	998	434	1432	5.0%
	TTMA	44	2778	1027	3805	13.3%	46	317	1355	1672	5.8%	90	880	1086	1966	6.9%
	HPYR	49	605	76	681	2.4%	51	258	551	809	2.8%	100	1150	271	1421	5.0%
Effects	PSA	49	583	476	1059	3.7%	-	-	-	-	-	-	-	-	-	-
	CEA	38	1529	681	2210	7.7%	38	886	532	1418	5.0%	76	1194	419	1613	5.6%
	p53 (SL)	42	261	676	937	3.3%	43	601	733	1334	4.7%	85	507	883	1390	4.9%
	micronuclei	36	241	371	612	2.1%	38	320	321	641	2.2%	74	291	561	852	3.0%
	Comet assay (median)	31	318	125	443	1.5%	36	633	441	1074	3.8%	67	575	580	1155	4.0%
	HDG	49	3054	5023	8077	28.2%	51	383	178	561	2.0%	100	1024	2001	3025	10.6%

Once again huge differences between men and women are evident, especially when applying a more stringent evaluation in which only correlations are taken into account that exceed the level of coincidence (5%). Men react, in comparison to women, stronger on cadmium exposure as well as on the exposure to benzene (i.e. the benzene metabolite t,t'-mucon acid, TTMA). Women react stronger on exposures of lead and PCBs.
Furthermore, for men a high correlation with the urinary excretion of 8-hydroxy-guanine (HDG), a marker for occurrence and repair of oxidative DNA damage, is seen.

Besides differences the correlation analysis also revealed similarities between men and women in the way the expression levels of some genes correlate with the exposure levels of certain pollutants and with the effect markers. This is shown in table 7. Especially for lead, PCBs, HCB and to a lesser extent for cadmium and HDG the correlation between exposure and gene expression is significantly different for men and women showing an opposite sign for the correlation coefficient. In contrast, the correlation coefficient for PAHs, p53, micronuclei and to a lesser extent for TEQ has an equal sign. Those genes that have the same sign for the correlation coefficient could be indicators for gender independent effects of exposure.

Table 7: Number of genes that are significantly correlated for both men and women.

	equal sign of correlation coefficient		opposite sign of correlation coefficient		p-value binomial distribution
Cd	303	46%	357	**54%**	0.00341
Pb	6	18%	28	**82%**	0.00008
SUMPCB	5	12%	38	**88%**	<0.000001
PCB118	18	62%	11	38%	0.06444
DDE	3	27%	8	73%	0.08057
HCB	36	26%	102	**74%**	<0.000001
TEQ	24	**62%**	15	38%	0.04573
TTMA	80	28%	204	**72%**	<0.000001
hPYR	38	**90%**	4	10%	<0.000001
CEA	53	54%	46	46%	0.06259
p53	29	**85%**	5	15%	0.00002
micronuclei	8	**100%**	0	0%	0.00391
Comet assay	6	60%	4	40%	0.20508
HDG	31	39%	48	**61%**	0.01450

3.5 Genetic network/pathway analysis

In the following section the results of the genetic network/pathway analysis conducted by the pathway finding tool MetaCore are described. The number of modulated networks for men and women related to the exposure levels of the pollutants and effect markers is shown in Table 8.

The “fraction” indicates the number of significantly modulated pathways in relation to the total number of pathways (650) on MetaCore.

Table 8: Results of MetaCore pathway analysis of genes correlating to exposure and effect markers.

	# pathways				
marker	**men**	**fraction**	**women**	**fraction**	**overlapping**
cd	27	4.15%	46	7.08%	0
hpyr	12	1.85%	9	1.38%	0
pb	18	2.77%	10	1.54%	0
pcbs	15	2.31%	4	0.62%	0
teq	20	3.08%	7	1.08%	0
ttma	60	9.23%	60	9.23%	7
dde	31	4.77%	12	1.85%	0
hcb	22	3.38%	17	2.62%	4
pcb118	14	2.15%	15	2.31%	0
micronuclei	18	2.77%	17	2.62%	0
cea	28	4.31%	12	1.85%	0
hdg	30	4.62%	13	2.00%	1
comet assay	15	2.31%	18	2.77%	0
p53	23	3.54%	34	5.23%	3
psa (alleen mannen)	8	1.23%			
ttest	16	2.46%	22	3.38%	2
cleartest	6	0.92%	8	1.23%	0

Markers with common pathways for men and women are highlighted.

Table 9 and 10 show the significantly altered genetic networks for high exposed men and women, respectively.

Table 9: Significantly modified pathways for high exposed men.

#	Name	pValue	Network objects
1	Development_Regulation of CDK5 in CNS	2.816e-3	3/16
2	Development_MAG-dependent inhibition of neurite outgrowth	3.998e-3	3/18
3	Regulation of metabolism_Bile acids regulation of glucose and lipid metabolism via FXR	6.265e-3	3/21
4	Development_ERK5 in cell proliferation and neuronal survival	7.160e-3	3/22
5	Apoptosis and survival_Role of CDK5 in neuronal death and survival	7.160e-3	3/22
6	Transport_RAB3 regulation pathway	1.109e-2	2/9
7	Development_Neurotrophin family signaling	1.147e-2	3/26
8	Regulation of lipid metabolism_FXR-dependent negative-feedback regulation of bile acids concentration	1.370e-2	2/10
9	Transport_FXR-regulated cholesterol and bile acids cellular transport	1.654e-2	2/11
10	G-protein signaling_RhoB regulation pathway	1.962e-2	2/12
11	Oxidative stress_Angiotensin II-induced production of ROS	2.291e-2	2/13
12	Apoptosis and survival_NGF signaling pathway	2.291e-2	2/13
13	Nitrogen metabolism	2.642e-2	2/14
14	Nitrogen metabolism/ Rodent version	2.642e-2	2/14
15	NGF activation of NF-kB	3.403e-2	2/16
16	Ascorbate metabolism	3.657e-2	1/2

Table 10: Significantly modified pathways for high exposed women.

#	Name	pValue	Network objects
1	N-Glycan biosynthesis p2	3.149e-3	7/18
2	Apoptosis and survival_Regulation of Apoptosis by Mitochondrial Proteins	3.777e-3	9/28
3	Riboflavin metabolism	5.147e-3	4/7
4	Development_Growth hormone signaling via STATs and PLC/IP3	1.069e-2	8/27
5	Apoptosis and survival_Role of CDK5 in neuronal death and survival	1.107e-2	7/22
6	Apoptosis and survival_Granzyme B signaling	1.346e-2	8/28
7	Neurophysiological process_Dopamine D2 receptor signaling in CNS	1.525e-2	4/9
8	Apoptosis and survival_FAS signaling cascades	1.593e-2	10/40
9	Apoptosis and survival_Caspase cascade	1.672e-2	8/29
10	Chemotaxis_Leukocyte chemotaxis	1.892e-2	10/41
11	Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim	2.486e-2	8/31
12	Immune response_CD16 signaling in NK cells	2.584e-2	9/37
13	Signal transduction_AKT signaling	2.810e-2	7/26
14	G-protein signaling_Rac3 regulation pathway	3.291e-2	4/11
15	Transport_FXR-regulated cholesterol and bile acids cellular transport	3.291e-2	4/11
16	Cytoskeleton remodeling_CDC42 in cellular processes	3.333e-2	5/16
17	Apoptosis and survival_Role of IAP-proteins in apoptosis	3.419e-2	7/27
18	Development_Notch Signaling Pathway	3.544e-2	8/33
19	Apoptosis and survival_Endoplasmic reticulum stress response pathway	4.128e-2	9/40
20	Muscle contraction_nNOS Signaling in Skeletal Muscle	4.486e-2	4/12
21	Retinol metabolism	4.729e-2	6/23
22	Apoptosis and survival_Ceramides signaling pathway	4.885e-2	7/29

The two highlighted pathways are deregulated in both high exposed men and women. The following figures show these pathways in more detail (next page):

Figure 1: Apoptosis and survival_Role of CDK5 in neuronal death and survival

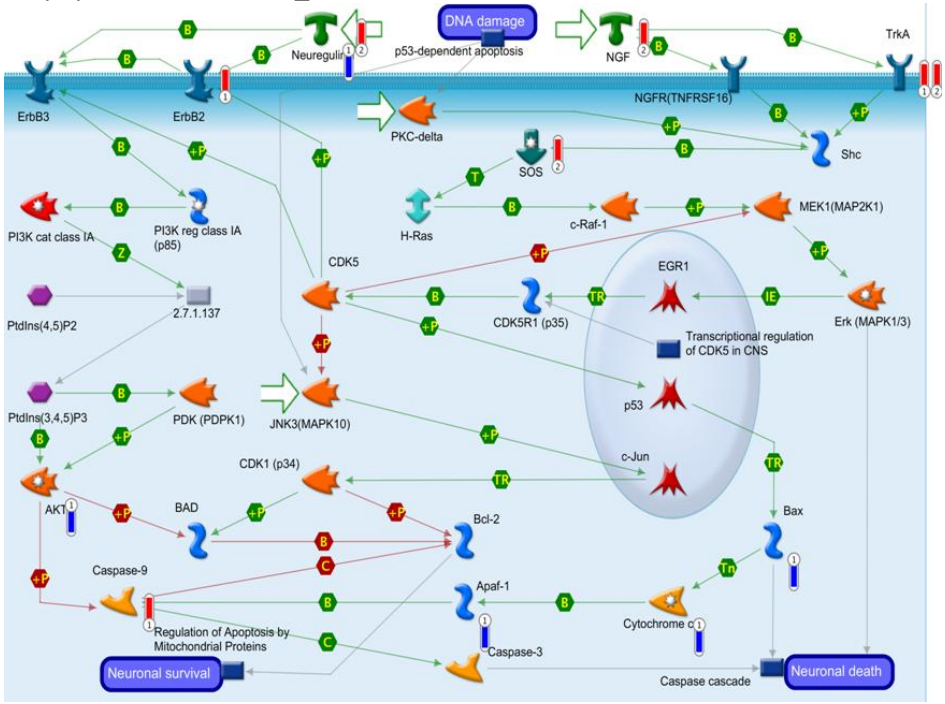
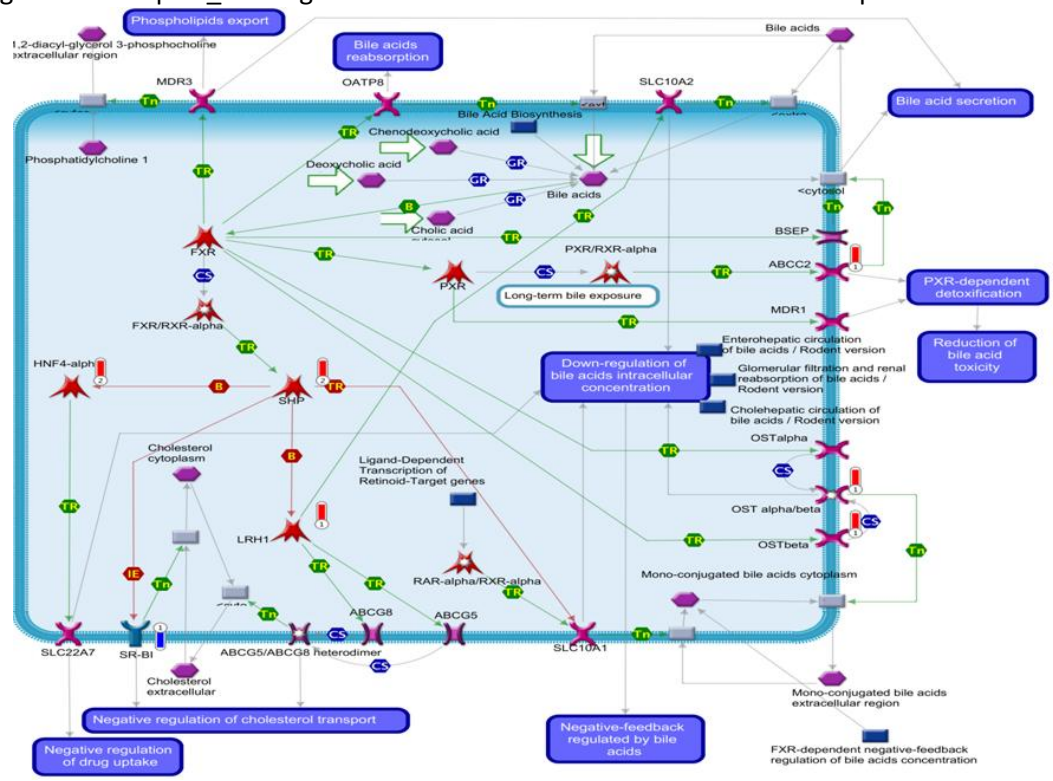


Figure 2: Transport_FXR-regulated cholesterol and bile acids cellular transport



The thermometers show the gene expression of the women (1) en men (2). Up-regulation is indicated by red and down-regulation by blue.

Since men and women are clearly different in their genomic response on the exposure to in particular hormone deregulating pollutants, specific focus was placed on hormone-related pathways from the MetaCore analysis. The analysis of men showed “Estrogen biosynthesis”

in relation with Cd exposure. In women “Estradiol metabolism” was found in relation with Cd exposure and the formation of micronuclei as well as “Estrone metabolism” in relation to exposure to PCB118.

Table 11 and 12 show the significantly altered genetic networks for the effect marker HDG in men and women, respectively.

Table 11: Significantly modified pathways for effect marker HDG in men.

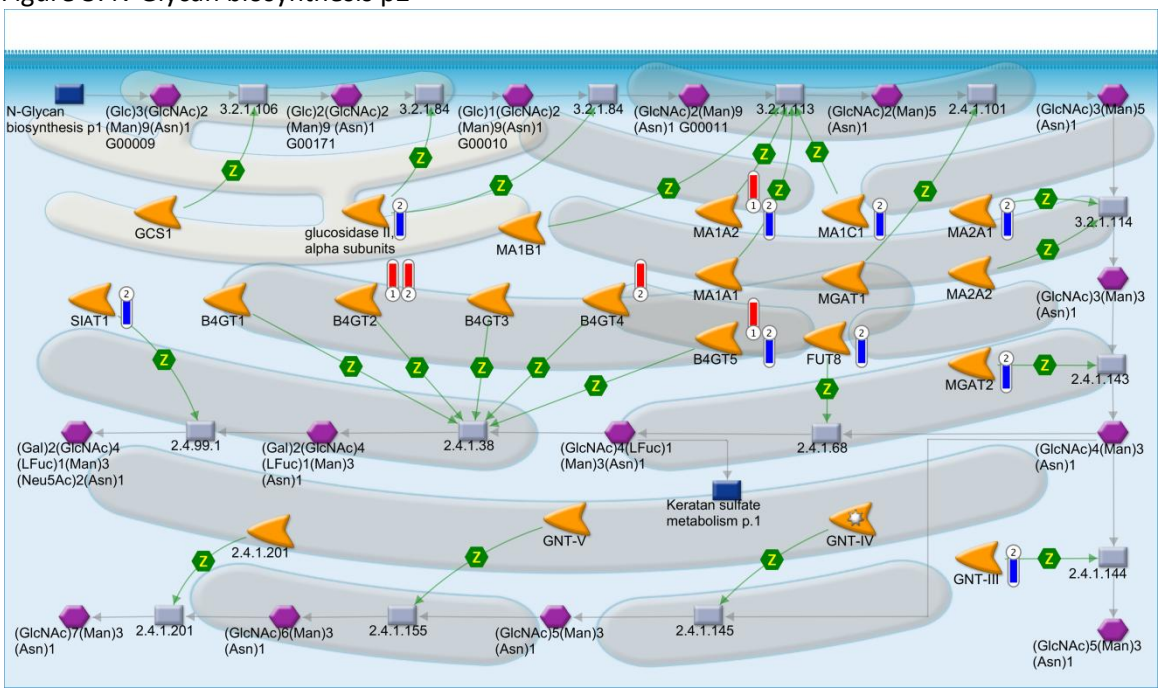
#	Name	pValue	Network objects
1	Neurophysiological process_Melatonin signaling	2.290e-3	9/11
2	Immune response_CCR5 signaling in macrophages and T lymphocytes	2.639e-3	21/35
3	Transcription_P53 signaling pathway	4.936e-3	19/32
4	Apoptosis and survival_Beta-2 adrenergic receptor anti-apoptotic action	5.351e-3	8/10
5	Delta508-CFTR traffic / Sorting endosome formation in CF	7.025e-3	12/18
6	Transcription_Transcription regulation of aminoacid metabolism	1.227e-2	13/21
7	CFTR folding and maturation (norm and CF)	1.351e-2	8/11
8	Immune response_CXCR4 signaling via second messenger	1.354e-2	11/17
9	Normal wtCFTR traffic / Sorting endosome formation	1.398e-2	9/13
10	Immune response_IFN gamma signaling pathway	1.754e-2	19/35
11	Development_Growth hormone signaling via PI3K/AKT and MAPK cascades	1.754e-2	19/35
12	Propionate metabolism p.2	2.207e-2	12/20
13	Signal transduction_Calcium signaling	2.207e-2	12/20
14	N-Glycan biosynthesis p2	2.378e-2	11/18
15	Transcription_Transcription factor Tubby signaling pathways	2.395e-2	5/6
16	Muscle contraction_ GPCRs in the regulation of smooth muscle tone	2.618e-2	15/27
17	PDGF activation of prostacyclin synthesis	2.739e-2	6/8
18	G-protein signaling_RhoB regulation pathway	2.800e-2	8/12
19	Transcription_CREM signaling in testis	2.800e-2	8/12
20	Translation_Opioid receptors in regulation of translation	2.800e-2	8/12
21	Neurophysiological process _Visual perception	2.800e-2	8/12
22	Development_Ligand-dependent activation of the ESR1/AP-1 pathway	2.836e-2	7/10
23	Transcription_Ligand-dependent activation of the ESR1/SP pathway	3.514e-2	12/21
24	Development_PACAP signaling in neural cells	3.866e-2	11/19
25	Cell cycle_Role of 14-3-3 proteins in cell cycle regulation	4.243e-2	10/17
26	UMP biosynthesis	4.496e-2	3/3
27	Transport_RAN regulation pathway	4.640e-2	9/15
28	DNA damage_Role of Brca1 and Brca2 in DNA repair	4.744e-2	13/24
29	Development_Hedgehog signaling	4.819e-2	16/31
30	Cell cycle_The metaphase checkpoint	4.819e-2	16/31

Table 12: Significantly modified pathways for effect marker HDG in women.

#	Name	pValue	Network objects
1	Keratan sulfate metabolism p.1	2.417e-3	3/11
2	Signal transduction_Activin A signaling regulation	4.131e-3	4/26
3	Keratan sulfate metabolism p.2	6.183e-3	3/15
4	Immune response_MIF - the neuroendocrine-macrophage connector	6.183e-3	3/15
5	N-Glycan biosynthesis p2	1.048e-2	3/18
6	IMP biosynthesis	2.143e-2	2/9
7	HIV-1 signaling via CCR5 in macrophages and T lymphocytes	2.324e-2	3/24
8	Regulation of lipid metabolism_G-alpha(q) regulation of lipid metabolism	2.634e-2	2/10
9	Immune response_Histamine signaling in dendritic cells	2.877e-2	3/26
10	G-protein signaling_Rac3 regulation pathway	3.165e-2	2/11
11	Development_EGFR signaling via PIP3	3.735e-2	2/12
12	wtCFTR and delta508 traffic / Clathrin coated vesicles formation (norm and CF)	4.341e-2	2/13
13	Cytoskeleton remodeling_Thyroliberin in cytoskeleton remodeling	4.980e-2	2/14

The highlighted pathway in Table 11 and 12 is deregulated in both men and women. The following figures show these pathways in more detail (next page):

Figure 3: N-Glycan biosynthesis p2



The thermometers show the gene expression of the women (1) en men (2). Up-regulation is indicated by red and down-regulation by blue.

Table 13 and 14 show the significantly altered genetic networks for the effect marker p53 in men and women, respectively.

Table 13: Significantly modified pathways for effect marker p53 in men.

#	Name	pValue	Network objects
1	Cytoskeleton remodeling_Role of Activin A in cytoskeleton remodeling	7.396e-4	5/16
2	Cytoskeleton remodeling_ESR1 action on cytoskeleton remodeling and cell migration	1.388e-3	4/11
3	Muscle contraction_EDG5-mediated smooth muscle contraction	7.953e-3	4/17
4	Apoptosis and survival_Ceramides signaling pathway	1.202e-2	5/29
5	Chemotaxis_Inhibitory action of lipoxins on IL-8- and Leukotriene B4-induced neutrophil migration	1.444e-2	4/20
6	Cell adhesion_Integrin-mediated cell adhesion and migration	1.444e-2	4/20
7	Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim	1.589e-2	5/31
8	Cytokine production by Th17 cells in CF (Mouse model)	1.589e-2	5/31
9	Inhibitory action of Lipoxins on neutrophil migration	1.718e-2	4/21
10	DNA damage_Mismatch repair	1.823e-2	3/12
11	Development_EGFR signaling via PIP3	1.823e-2	3/12
12	Immune response_CD40 signaling	2.083e-2	6/45
13	Development_NOTCH-induced EMT	2.286e-2	3/13
14	Development_Thrombospondin-1 signaling	2.286e-2	3/13
15	Cytokine production by Th17 cells in CF	2.356e-2	4/23
16	Immune response_IL-6 signaling pathway	2.722e-2	4/24
17	Muscle contraction_ACM regulation of smooth muscle contraction	2.722e-2	4/24
18	Immune response_MIF in innate immunity response	2.722e-2	4/24
19	Bacterial infections in CF airways	2.900e-2	5/36
20	Blood coagulation_GPCRs in platelet aggregation	3.573e-2	5/38
21	Signal transduction_JNK pathway	3.573e-2	5/38
22	G-protein signaling_EDG5 signaling	4.022e-2	3/16
23	Cell cycle_Influence of Ras and Rho proteins on G1/S Transition	4.334e-2	5/40

Table 14: Significantly modified pathways for effect marker p53 in women.

#	Name	pValue	Network objects
1	HIV-1 signaling via CCR5 in macrophages and T lymphocytes	3.075e-4	7/24
2	Immune response_NF-AT signaling and leukocyte interactions	2.246e-3	5/17
3	Muscle contraction_ GPCRs in the regulation of smooth muscle tone	3.876e-3	6/27
4	Neurophysiological process_Thyroliberin in cell hyperpolarization and excitability	5.322e-3	4/13
5	Transcription_Transcription regulation of aminoacid metabolism	6.101e-3	5/21
6	Neurophysiological process_Long-term depression in cerebellum	7.116e-3	4/14
7	Immune response_IL-3 activation and signaling pathway	7.528e-3	5/22
8	Immune response_PGE2 signaling in immune response	7.528e-3	5/22
9	Immune response_Regulation of T cell function by CTLA-4	1.105e-2	5/24
10	Muscle contraction_EDG5-mediated smooth muscle contraction	1.474e-2	4/17
11	Immune response_HTR2A-induced activation of cPLA2	1.474e-2	4/17
12	Heme metabolism	1.555e-2	5/26
13	Immune response_Histamine signaling in dendritic cells	1.555e-2	5/26
14	Immune response_Signaling pathway mediated by IL-6 and IL-1	1.811e-2	4/18
15	Triacylglycerol metabolism p.1	1.821e-2	5/27
16	Development_Growth hormone signaling via STATs and PLC/IP3	1.821e-2	5/27
17	Blood coagulation_GPCRs in platelet aggregation	2.120e-2	6/38
18	Immune response _CCR3 signaling in eosinophils	2.120e-2	6/38
19	Cytoskeleton remodeling_Role of PKA in cytoskeleton reorganisation	2.191e-2	4/19
20	Cardiac Hypertrophy_Ca(2+)-dependent NF-AT signaling in Cardiac Hypertrophy	2.617e-2	4/20
21	Immune response_IFN alpha/beta signaling pathway	2.617e-2	4/20
22	Development_EDG5 and EDG3 in cell proliferation and differentiation	2.617e-2	4/20
23	Immune response_CD28 signaling	2.791e-2	5/30
24	wtCFTR and delta508-CFTR traffic / Generic schema (norm and CF)	2.791e-2	5/30
25	Neurophysiological process_ACM regulation of nerve impulse	3.089e-2	4/21
26	G-protein signaling_RAC1 in cellular process	3.089e-2	4/21
27	Transcription_PPAR Pathway	3.609e-2	4/22
28	Development_GH-RH signaling	3.638e-2	3/13
29	Blood coagulation_Blood coagulation	4.175e-2	4/23
30	Transport_Clathrin-coated vesicle cycle	4.244e-2	7/56
31	Cytoskeleton remodeling_Alpha-1A adrenergic receptor-dependent inhibition of PI3K	4.361e-2	2/6
32	Neurophysiological process_Glutamate regulation of Dopamine D1A receptor signaling	4.437e-2	3/14
33	Cytoskeleton remodeling_Thyroliberin in cytoskeleton remodeling	4.437e-2	3/14
34	Muscle contraction_ACM regulation of smooth muscle contraction	4.789e-2	4/24

The highlighted pathways in Table 13 and 14 are deregulated in both men and women. The following figures show these pathways in more detail (next page):

Figure 4: Blood coagulation_GPCRs in platelet aggregation

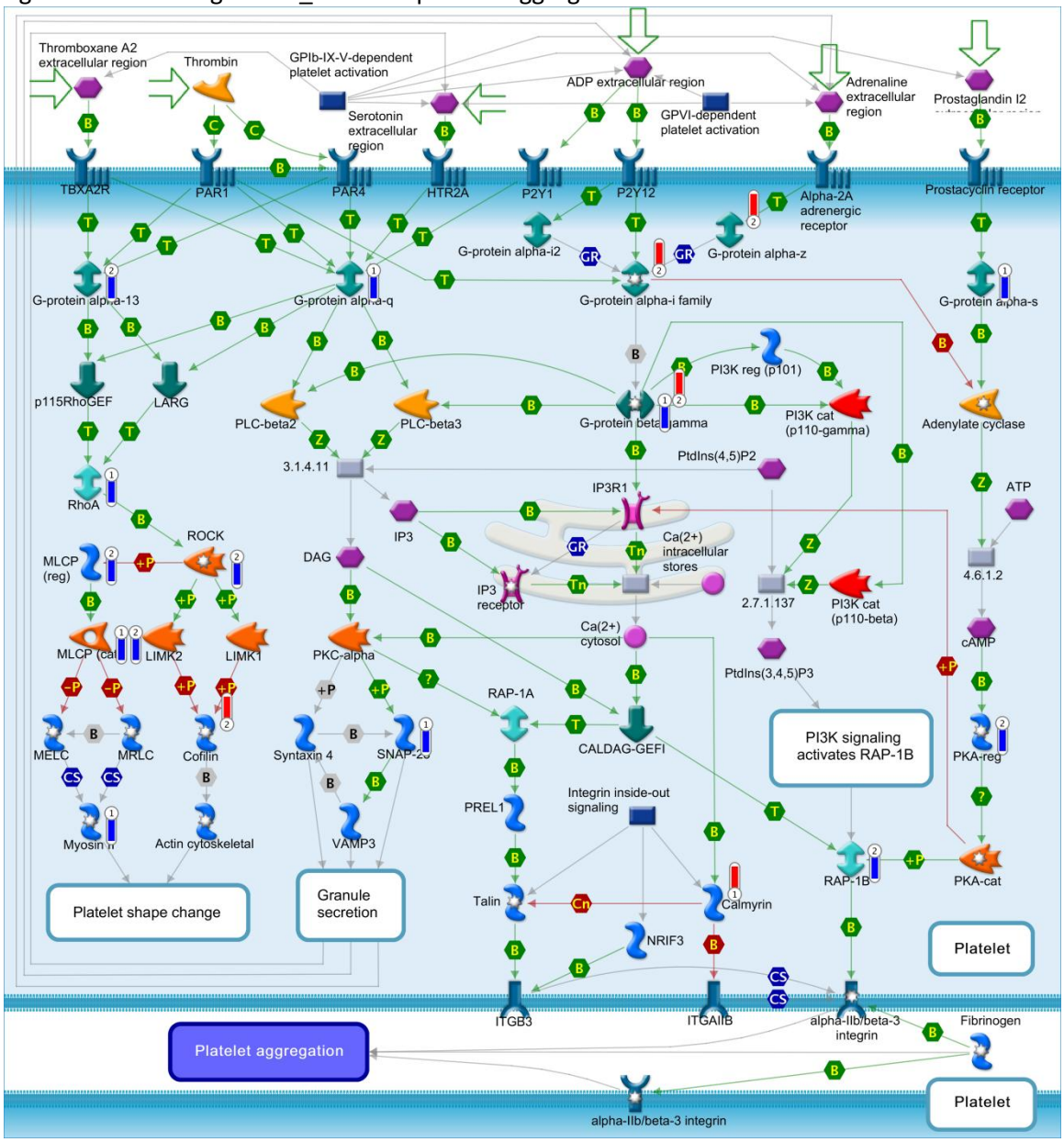


Figure 5: Muscle contraction_ACM regulation of smooth muscle contraction

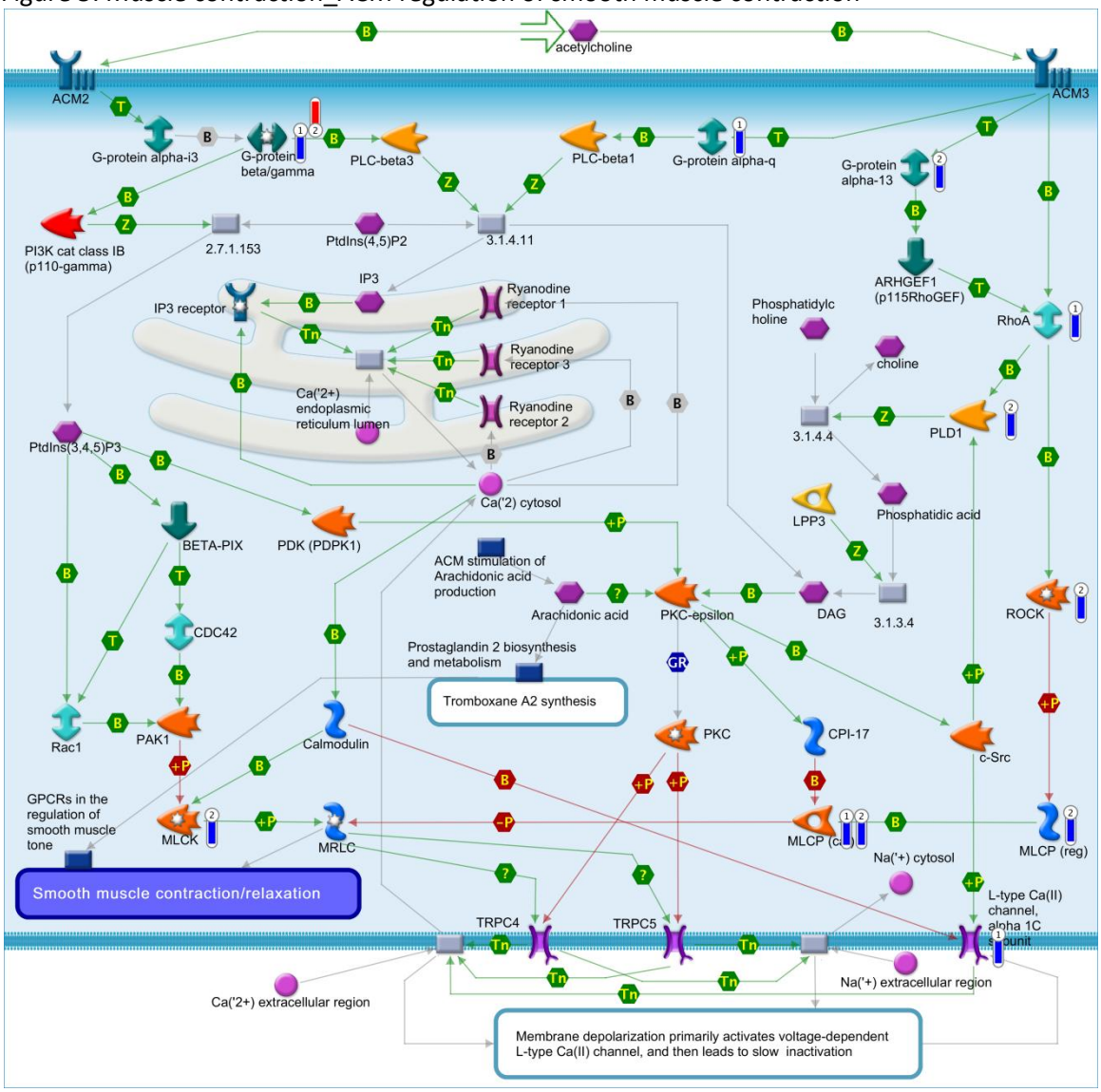
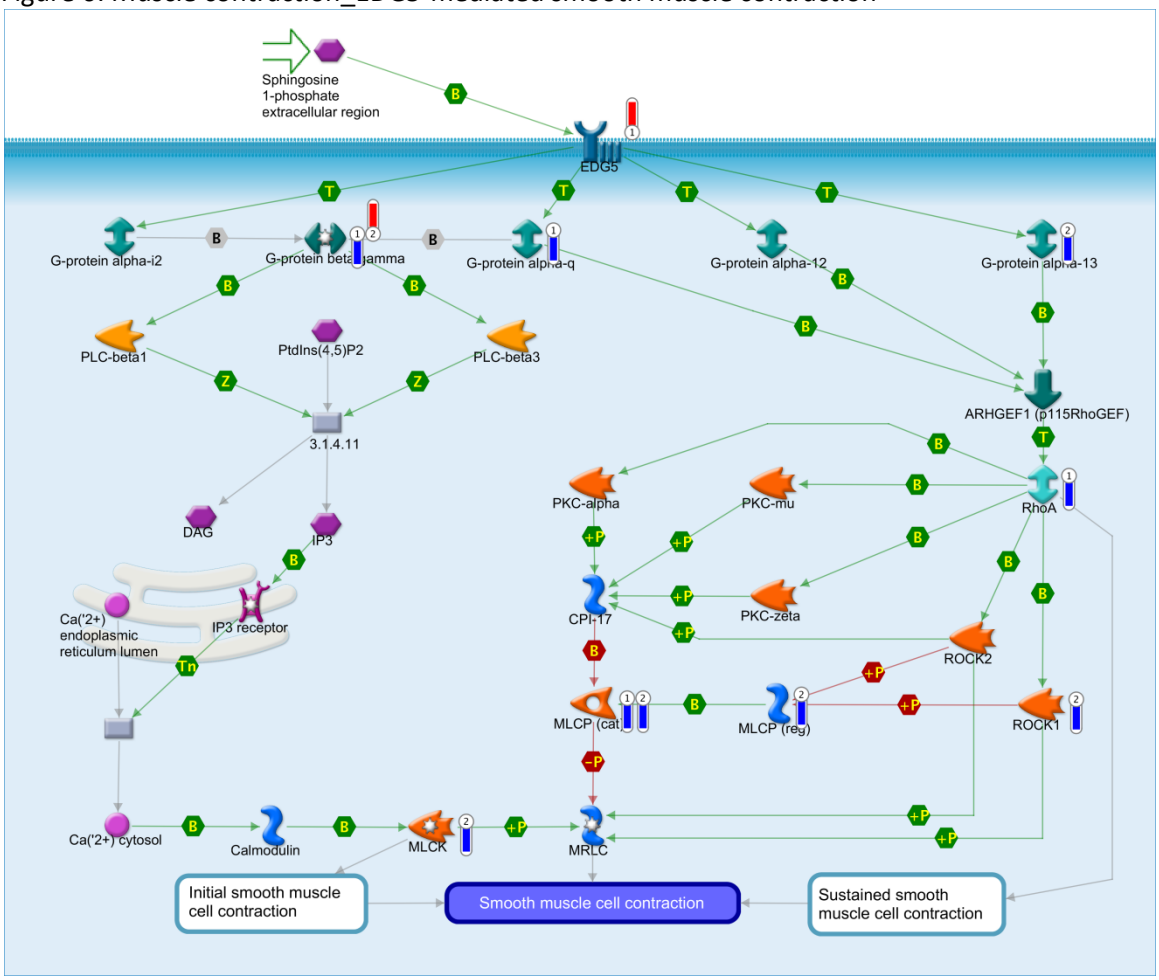


Figure 6: Muscle contraction_EDG5-mediated smooth muscle contraction



The thermometers show the gene expression of the women (1) en men (2). Up-regulation is indicated by red and down-regulation by blue.

The pathway analysis is in general limited because only part of the correlated or significant genes can be used, i.e. only those genes that are present on the pathways. Therefore, a GO processes analyses is performed in MetaCore. In these analyses all correlated or significant genes are used. Table 16 shows the significant GO processes for men and women in relation to exposure levels of pollutants and effect markers.

Table 15: Resultaten MetaCore GO processes analyse

	# GO processes					
marker	**men**		**women**		**overlapping**	
	P<0.05	**P<0.05 & FDR<0.1**	**P<0.05**	**P<0.05 & FDR<0.1**	**P<0.05**	**P<0.05 & FDR<0.1**
cd	256	54	264	32	57	10
hpyr	162	0	218	12	6	0
pb	152	0	118	1	3	0
pcbs	181	0	186	9	6	0
teq	301	1	332	5	18	0
ttma	355	17	401	72	45	0
dde	284	0	232	1	2	0
hcb	229	0	254	14	14	0
pcb118	159	0	131	1	4	0
micronuclei	221	14	178	0	3	0
cea	165	0	172	0	4	0
hdg	249	50	288	7	23	0
comet assay	174	0	186	12	3	0
p53	275	54	181	0	23	0
psa (alleen mannen)	205	0				
ttest	203	0	222	0	7	0
cleartest	140	0	107	46	1	0

For Cd there is an overlap in significant GO processes (P<0.05 andn FDR <0.1) between men and women. These processes are related to RNA metabolic processes.

An overview of the most appearing GO processes for pollutants and effect markers is shown in Table 16.

Table 16: The most appearing GO processes from the MetaCore analyses

	# GO processes	
marker	**men**	**women**
cd	RNA metabolic process transcription cell cycle	chromatin assembly RNA metabolic process transcription
hpyr	regulation of immune system	chromatin assembly
pb	protein metabolic process phosphorylation	protein processing
pcbs	immune response	translation macromolecule biosynthetic process
teq	RNA metabolic process	neurological system process
ttma	RNA metabolic process translation cellular response to stress	signaling
dde	translation cell cycle regulation of immune system proces	negative regulation of translation
hcb	homeostatic process	negative regulation of gene expression
pcb118	regulation of immune system	homeostatic process
micronuclei	chromatin assembly	various
cea	transport	regulation of immune system process
hdg	RNA metabolic process transcription	translation protein metabolic process
comet assay	various	signaling
p53	macromolecule metabolic process	regulation of protein modification process
psa (alleen mannen)	various	
ttest	transport	various
cleartest	various	neurological nervous process signaling

3.6 RT-PCR vs Microarray

The previously obtained gene expression data of the RT-PCR analysis of the 1st generation Steunpunt Milieu & Gezondheid (van Leeuwen et al. 2008) was compared with those from the microarray analysis. For the 8 genes (MAPK14, SOD2, CYP1B1, PINK1, TIGD3, CXCL1, DGAT2, ATF4) investigated in the RT-PCR analysis the gene expression value in the microarray analysis exceeded the detection limit in all 100 individuals. The comparison showed for men a significant correlation between RT PCR values and microarray values for MAPK14 ($p<0.0001$), SOD2 ($p<0.0001$), CXCL1 ($p=0.035$), DGAT2 ($p<0.0001$), CYP1B1 ($p=0.008$) and PINK1 ($p<0.0001$), but not for ATF4 ($p=0.5$) and TIGD3 ($p= 0.22$). For women a significant correlation was found for MAPK14 ($p=0.039$), SOD2 ($p<0.0001$), CXCL1 ($p=0.034$), DGAT2 ($p<0.0001$), PINK ($p<0.0001$) and TIGD3 ($p<0.0001$), but not for ATF4 ($p=0.21$) and CYP1B1 ($p=0.20$).
With 5 genes highly correlated for both men and women (MAPK14, SOD2, CXCL1, DGAT2, PINK1) and 2 genes highly correlated in either men (CYP1B1) or women (TIGD3) there is good correspondence between the RT-PCR and microarray data. Differences between both analyses could be the result of differences in the probes used, i.e. the probes on the microarray may be different than the RT-PCR fragment.

CONCLUSIONS

In the 1st Generation Steunpunt study 8 genes were selected for RT-PCR analysis and these have been used as biomarker in 398 Flemish adults. The microarray results of the current study in a subgroup of 100 Flemish adults showed that for both men and women 6 of these 8 genes significantly correlate. This is a significant indication of the technical reliability of the in this report presented microarray results.

Dose-response relations between exposure to the pollutants and gene expression changes measured by microarray technology show complex biological reactions. This complexity is not unexpected as the exposure also consist of complex mixtures of environmental contaminants.

It should be noted that these reactions occur in representatives of the ordinary Flemish population and thus there is no question about extreme exposures.

The main result of this study regards the shown difference in gender with regard to the exposure of the investigated pollutants. This is also observed for the effect markers. The gender differences in the genomic reaction pattern are substantial. In future biomonitoring studies it is advisable to take these gender differences explicitly into account.

The most found deregulated networks in men with respect to pollutants and effect markers (indicated between brackets and the effect markers highlighted) are:

- Transcription_P53 signaling pathway (cd, hpyr, dde, hdg)
- Transcription_Transcription regulation of aminoacid metabolism (teq, dde, hcb, hdg)
- Cell adhesion_ECM remodeling (teq, micronuclei, comet assay)
- Immune response_IL-13 signaling via JAK-STAT (hcb, cea)
- Immune response_PGE2 signaling in immune response (hpyr, pcb118)
- Immune response_T cell receptor signaling pathway (ttma, hcb, comet assay)
- Muscle contraction_ GPCRs in the regulation of smooth muscle tone (cd, hdg, comet assay)
- Muscle contraction_EDG5-mediated smooth muscle contraction (cd, cea, p53)
- PDGF activation of prostacyclin synthesis (cd, teq, hdg)
- Transcription_CREM signaling in testis (pcb118, hdg)

The most found deregulated networks in women with respect to pollutants and effect markers (indicated between brackets and the effect markers highlighted) are:

- N-Glycan biosynthesis p2 (cd, pb, hcb, hdg)
- Development_TGF-beta-induction of EMT via ROS (cd, hpyr, ttma, dde)
- Retinol metabolism (pcbs, teq, pcb118)
- Translation _Regulation of EIF4F activity (cd, ttma, dde, micronuclei)
- Apoptosis and survival_Ceramides signaling pathway (cd, ttma)
- Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim (cd)
- Development_Growth hormone signaling via STATs and PLC/IP3 (p53)
- G-protein signaling_Rac3 regulation pathway (cd, hdg)
- HIV-1 signaling via CCR5 in macrophages and T lymphocytes (ttma, hdg, p53)
- Keratan sulfate metabolism p.2 (cd, dde, hdg)

- Neurophysiological process_ACM regulation of nerve impulse (micronuclei, comet assay, p53)
- Neurophysiological process_Glutamate regulation of Dopamine D1A receptor signaling (dde, pcb118, p53)
- Normal and pathological TGF-beta-mediated regulation of cell proliferation (cd, ttma, dde)

In this final comparison it is clearly shown that mainly immunological responses in men and apoptotic reactions and neurophysiologic processes in women are influenced.

REFERENCES

De Coster S, Koppen G, Bracke M, Schroijen C, Den Hond E, Nelen V, Van de Mieroop E, Bruckers L, Bilau M, Baeyens W, Schoeters G, van Larebeke N. Pollutant effects on genotoxic parameters and tumor-associated protein levels in adults: a cross sectional study. Environ Health. 2008 7:26.

Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, Minguez P, Vera J, Mukherjee S, Valls J, Pujana MA, Alloza E, Herrero J, Al-Shahrour F, Dopazo J. Next station in microarray data analysis: GEPAS. Nucleic Acids Res. 2006 Jul 1;34

Valls J, Grau M, Solé X, Hernández P, Montaner D, Dopazo J, Peinado MA, Capellá G, Moreno V, Pujana MA. CLEAR-test: combining inference for differential expression and variability in microarray data analysis. J Biomed Inform. 2008 Feb;41(1):33-45.

van Leeuwen DM, Gottschalk RW, Schoeters G, van Larebeke NA, Nelen V, Baeyens WF, Kleinjans JC, van Delft JH. Transcriptome analysis in peripheral blood of humans exposed to environmental carcinogens: a promising new biomarker in environmental health studies. Environ Health Perspect. 2008 Nov;116(11):1519-25.

WP4: Conclusies en aanbevelingen

1. Staalname

De studie omvatte de analyses van perifere bloedstalen die verzameld werden bij 2 verschillende populaties van volwassenen. In 2004 werden bloedstalen verzameld voor genexpressieanalyse bij volwassenen tussen 50 en 65 jaar die werden gerekruteerd in het kader van het steunpunt Milieu en gezondheid (WP3). In 2009-2010 waren 22 volwassenen met een leeftijd tussen 20 en 40 jaar bereid om 6 opeenvolgende keren (3 maal in de winter, 3 maal in de lente) een bloedstaal te doneren(WP2). De staalname binnen een seizoen waren ongeveer een week van elkaar gescheiden. Een klein bloedvolume (3mL) werd gecollecteerd in de geschikte tubes en volstond om de analyses uit te voeren. De goede kwaliteit van het RNA dat na isolatie bekomen werd in beide studies toont aan dat de stalen langdurig kunnen bewaard worden bij -80°C.

De resultaten van de huidige studie toonden aan dat zowel bij de recent gecollecteerde stalen als bij de eerder gecollecteerde stalen, hoog kwalitatief RNA kan bekomen worden uit perifeer bloed bij biomonitoringsstudies van de algemene bevolking.
Betrouwbare en reproduceerbare methodes voor behandeling van bloedstalen en voor de bewaarcondities zijn beschikbaar aan de UM en VITO

2. Genexpresssiemetingen

Het genereren van microarray data kan gebeuren met verschillende commerciële platforms. De overeenkomst van resultaten van het gebruik van verschillende platforms zijn goed. Dit werd uitvoerig gedocumenteerd in de literatuur (Patterson *et al.* 2006, Shi *et al.* 2006).

Er werd in dit project gestreefd naar maximale harmonisatie en de verschillen in aanpak werden gedocumenteerd. Beide laboratoria werkten met het Agilent platform.

	VITO	UM
Staalname periode	2009-2010	2005
Opvangen van de stalen	Tempus tubes (3mL)	PAX gene tubes 2.5 mL
Extractie van RNA	RNA opbrengst van 5,5 tot 17,4 µg/ staal	Niet beschikbaar
Concentratie van RNA	precipitatie	vriesdrogen
Globine opzuivering van RNA	AMBION protocol	AMBION protocol
Kwaliteitsanalyse van RNA	BioAnalyzer 2100	BioAnalyzer 2100
RNA kwaliteitseisen	RIN>8	RIN>6
Hybridisatie en wasstappen	Automatische methode (TECAN)	Manueel
Arrays	Agilent G4112F Human Genome 4x44K array	Agilent G4112F Human Genome 4x44K array
scanners	AGILENT scanner Scanning gebeurd in XDR-mode met 2 laser intensiteiten (16 bit) en generatie van 2 TIF files die aan elkaar gekoppeld worden ofwel in 20-bit scanning mode met een hogere saturatiewaarde. Feature extraction voor spotherkenning en normalisatie.	Genepix scanner en software voor batch normalisatie Starting van feature extraction files naar TIF files
Pathway analyse	Ingenuity	Metacore

Gen expressie via microarrays (WP3) van 8 geselecteerde genen gemeten via RT-PCR in een voorgaande studie (2008) toonde voor 6 van de 8 genen een significante correlatie met de nieuwe resultaten op basis van whole genome arrays. Dit illustreert de betrouwbaarheid van de technologie. De reden waarom er geen bevestiging werd bekomen voor 2 genen is onduidelijk op dit moment. Een mogelijke oorzaak kan gezocht worden in het verschil in probe design om deze genen te detecteren. De 8 genen werden een paar jaar geleden geïdentificeerd met een specifieke microarray design en nadien bevestigd met real-time PCR. De huidige Agilent microarrays maakt gebruik van een andere probe design

De natte procedures zijn vergelijkbaar en de ruwe data zijn robuust en uitwisselbaar voor cross lab vergelijking.

3. Dataverwerking

Er bestaan verschillende procedures voor de verwerking van ruwe gegevens. **WP2** heeft vooraf enkel basis pre-processing gedaan van de ruwe data (132 arrays) via Feature Extraction op whole array niveau. Een eerste inspectie van de data wees uit dat een groot aantal signalen een lage intensiteit bezat en dus waarschijnlijk niet informatief waren. In parallel werd dezelfde filtering procedure toegepast als in werkpakket 3. Deze procedure identificeerde eveneens bijna 8 000 probes die niet voldeden aan de kwaliteitscriteria omdat de signaalsterkte te laag was. Er werd beslist om te verder te werken met enkel de kwaliteitsvolle probes (ongeveer 33 000 probes). De data werden genormaliseerd en een mixed model toegepast voor statistische verwerking. Na de statistische analyse (mixed models) werd een correctie doorgevoerd voor multipliciteit (dmv false discovery rate). WP 3 begon met een veel uitgebreidere pre-processing van de data (100 arrays). Deze was stringenter en deed QC op individuele spots (pixels, mean/median ratio, saturation, intensities >> background). Alle arrays werden naast elkaar gezet en flags werden toegekend aan 'slechte' spots. Als 30% van de spots geflagd was dan werd het gen uit de analyse verwijderd. Op deze manier bleven er minder spots over om statistische analyse te doen. Daarna werd normalisatie uitgevoerd (Lowess and quantile), GEPASS (uitfilteren van flat-peaks) . Er werd niet gecorrigeerd voor multipliciteit. Statistische analyse van verschillen tussen de hoog-blootgestelde en laag-blootgestelde groep gebeurde met t-testen en correlatie met blootstelling. Genlijsten zijn beschikbaar en biologische interpretatie is gebeurd.

Er zijn verschillende goede methodes voor datafiltering en statistische analyse van complexe microarray dataset. De inzichten hieromtrent evolueren en in de internationale wetenschappelijke wereld is er geen consensus betreffende een eenduidige strategie. Verschillende methodes kunnen gebruikt worden op voorwaarde dat ze conceptueel aanvaardbaar zijn en gedragen zijn in de wetenschappelijke wereld. Het proces van data processing is een onderdeel van een volledige pijplijn om microarray te gebruiken voor identificatie van genexpressie biomerkers in de context van milieu en gezondheid. Een relevante biologische conclusie en een validatie van de gevonden biomerkers in een onafhankelijke studiepopulatie zijn de finale criteria voor evaluatie van de pijplijn voor data-analyse. In een goed en betrouwbaar scenario zou uiteraard goede convergentie moeten bestaan voor de verschillende methodes. Microarray data analyse en biologische interpretatie met bioinformatica is een complex veld en wordt vaak onderschat bij het opmaken van een projectvoorstel en bijhorend budget. Voor toekomstige projecten is het belangrijk om voldoende tijd en middelen te voorzien opdat de gegenereerde data maximaal kunnen benut worden. Indien onderzoeksprojecten met meerdere labo's worden opgezet is het aan te bevelen dat bij aanvang van het project duidelijke afspraken gemaakt worden over een aantal procedures (zoals formaat van data files, kwaliteitscriteria voor data, gebruik van data of niveau van probe of gen, procedures voor correctie voor multipliciteit, enz.). Dergelijke afspraken zorgen er voor dat data in verschillende werkpakketten vergelijkbaar zijn en optimaal kunnen uitgewisseld worden.

4. Pathway analyse

Voor een biologische interpretatie van de genlijsten en aanrijking van specifieke biologische pathways zijn er diverse commerciële tools en open-source initiatieven beschikbaar. In deze studie werden Ingenuity en Metacore gebruikt. De inzichten en tools voor analyse zijn volop in ontwikkeling en worden best aangevuld met eigen specificaties, rapportage en visualisaties.
Statistische interpretatie is niet eenduidig wegens de complexe aard van biologische pathways en netwerken (bijv. door interconnectiviteit). De veelheid aan informatie en pathways die naar voor komen bemoeilijkt de biologische interpretatie en de doorvertaling naar gezondheidseffecten. De statistische analyse dient voornamelijk voor prioritisatie van biologische termen en/of pathways die verdere aandacht vragen.

In de context van humane biomonitoring is pathway analyse nuttig om data te aggregeren op een biologisch niveau en die informatie kan gebruikt worden om hypotheses te genereren naar mogelijke gezondheidsaspecten. De experimentele set-up laat niet toe het dynamisch patroon en de interactie van biologische netwerken en bijgevolg de biologische mechanismes ten volle te omvatten. Verandering is genexpressie hoeft niet noodzakelijk vertaald te worden in een fysiologische ontregeling omdat er andere controle mechanismen die zorgen voor een robuust systeem. Complementaire informatie op niveau van epigenetics, proteomics and/of metabolomics niveau zijn belangrijk om de relevantie van veranderingen in genexpressie te bevatten. Bovendien is vanuit fysiologisch standpunt een fenotypische ankering belangrijk.

5. Interpretatie van de gegevens

In WP1 werd een uitgebreid literatuur onderzoek gedaan. Er zijn momenteel verschillende internationale databanken vb comparative toxicogenomics database (http://www.mdibl.org/ctd.php) beschikbaar die informatie bevatten over de relatie tussen genen en ziektes en over genexpressieprofielen die geassocieerd zijn met fenotypes en met chemische blootstellingen. Naast informatie over het werkingsmechanisme van chemische stoffen , inter- en intraspecies variabiliteit , geven genexpressieprofielen een fingerprint van gecombineerde blootstellingen en worden zij beschouwd als een gevoelig “early warning” signaal. In een milieucontext is het gebruik van genexpressiemerkers nog volop in ontwikkeling. Er zijn reeds commerciële toepassingen voor de klinische diagnose van kanker (Mammaprint en PCA3). Eén van de uitdagingen momenteel is het onderscheid te maken tussen tijdelijke genexpressie veranderingen die gecompenseerd worden door robuuste biologische systemen en een effectieve verstoring die een functioneel (fysiologisch) effect kan uitlokken. Het is daarom belangrijk om genexpressiemetingen samen uit te voeren met andere effectmetingen. WP3 is daar verder op ingegaan. Een andere conclusie uit het literatuuronderzoek is dat het effect van persoonlijke kenmerken (inter- en intra-individuele variabiliteit, geslacht en leeftijd), levensstijl (voeding, roken) en staalname karakteristieken (tijdstip van de dag, seizoen) nog onvoldoende gekend is. In WP2 werd informatie bekomen over intra- en interindividuele variabiliteit.

Variabiliteit in gen expressie van perifere bloedcellen in een gezonde populatie van volwassenen:
We beschikken nu over gegevens met betrekking tot de korte en lang termijn intra- and interindividuele variabiliteit in genexpressie van perifere bloedcellen afkomstig van jonge gezonde volwassenen (n=22). Korte termijn variabiliteit werd bepaald via drie herhaalde metingen over een tijdspanne van één maand, lange termijn variabiliteit bevat metingen in 2 verschillende seizoenen (herfst en lente) met een tussenpauze van ongeveer 6 maanden. Dit wil zeggen dat we nu variatiecoëfficiënten hebben van meer dan 30 000 individuele genen waaronder ook de veel gebruikte huishoudgenen die als ijking gebruikt worden om variabiliteit van de methode te bepalen. 75% van de genen hebben een variatiecoëfficiënt kleiner dan 0.25. De bekomen informatie over de korte en lange termijn variabiliteit van individuele genen is een basis voor berekeningen van statistische power bij vervolgstudies waarbij de genexpressie van individuele genen als eindpunt wordt gebruikt. *A priori* power berekening om het aantal stalen te berekenen om bijvoorbeeld een significant verschil te vinden tussen 2 situaties hangt af van de genexpressie intensiteit, de standaard afwijking per gen, en het verwachte effect. Uit WP3 blijkt dat predictieve merkers waarschijnlijk uit een set van genen zullen bestaan.

De variabiliteit binnen een individu is kleiner dan de variabiliteit tussen individuen, wat aangeeft dat de genexpressie gemeten in perifere bloedcellen individu specifiek is. Mixed model analyse per gen gaf voor de meeste genen geen significante korte termijn trend aan binnen een seizoen. Alleen in de herfst werd voor 110 probes een significante korte termijn tijdstrend aangetoond.

Geslacht en seizoen hebben een impact op genexpressie: 191 en 1 995 probes werden significant beïnvloed door respectievelijk geslacht en seizoen. De waarden werden bekomen na een correctie voor multipliciteit via de Benjamini-Hochberg methode. Een cut-off van 5% werd gebruikt om significante genen te identificeren. Dit is een veel gebruikte methode, maar geen absolute methode aangezien er andere manieren zijn om voor multipliciteit te corrigeren. Een andere correctiemethode gebaseerd op q-waarden leverden 315 and 2 798 probes op voor de factoren geslacht en seizoen. Uit de biological pathway analyse via Ingenuity blijkt dat vooral de genen die in verband staan met ontsteking en het immuun systeem veranderen in expressie. De eerste resultaten wijzen in de richting van immuun-gerelateerde processen die kunnen beïnvloed zijn. In de literatuur is beschreven dat het immuunsysteem meer onder druk staat in de winter. Zowel stressors als immuun respons zijn seizoensafhankelijk (Nelson et al. 2004). WP2 was vooral opgezet om de variabiliteit van genexpressie metingen in een populatie te beschrijven. Een aantal genen blijken bovendien significant beïnvloed te worden door seizoen of geslacht. Genen die significant beïnvloed werden in deze WP2-studie of die een variabele expressie vertoonden, verdienen speciale aandacht indien uit blootstelling-effect onderzoek zou blijken dat dit kandidaat biomerker genen zijn. Genen die gemarkeerd zijn in WP2 zouden bijv. gemakkelijker aanleiding kunnen geven tot vals-positieve resultaten.

Bij ongeveer 8 000 probes was de variantie afhankelijk van seizoen of geslacht . Bij 25 probes hing de variabiliteit tussen individuen af van geslacht en van seizoen.

Relatie tussen genexpressieprofielen en biomerkers van blootstelling en effect

Dosis-afhankelijke genexpressie veranderingen zijn goed meetbaar in bloedcellen van volwassen Vlamingen die aan een reeks van polluenten zijn blootgesteld (n=100)

Een groot aantal genen bleek een wijziging in expressie te ondergaan in associatie met cadmium, hexachlorobenzeen en benzeen bij mannen en in associatie met cadmium, lood, merker PCBs, PCB 118, hexachlorobenzeen en benzeen bij vrouwen.

De blootstellingen blijken biologische processen te beïnvloeden. Bij mannen zijn dit eerder immuunresponsen, bij vrouwen zijn dit vooral apoptotische en neurofysiologische processen. Op dit moment is het nog niet geweten wat de fenotype consequenties zijn van de veranderingen in genexpressie. De biologische pathway analyse is een eerste vertaling van individuele genen naar biologische processen die verdere moeten geïnterpreteerd worden in hun fysiologische context.

Ook de dosis antwoord relaties verschillen tussen mannen en vrouwen; dit wijst op de noodzaak om in toekomstig biomonitoringsonderzoek de geslachten te onderscheiden.

Uit bovenstaande volgt dat genexpressie analyse een robuuste, gevoelige en uiterst informatieve biomerker is voor het bepalen van moleculaire effecten in blootgestelde populaties.

Wellicht is het meten van de genexpressie het gevoeligste en meest complete meetinstrument dat momenteel beschikbaar is om ontregelde biologische activiteit te meten.

6. Aanbevelingen

- De ruwe data, en verwerkte data, alsook statistische parameters per gen dienen in een databank opgeslagen te worden met garantie voor anonimiteit. De databank zou consulteerbaar moeten zijn en beschikbaar voor verdere onderzoeken. Een werkdocument dient opgemaakt te worden met een stappenplan om de aanmaak en beheer van zo'n database te realiseren.
- Verschillende procedures worden gehanteerd om ruwe gegevens te behandelen. Inzichten evolueren nog. Het is belangrijk om dezelfde werkwijze te hanteren in laboratoria die samenwerken.
- De informatie over korte en lange termijn variabiliteit in genexpressie kan gebruikt worden om de statistische power in te schatten wil men genexpressie van geselecteerde genen tussen 2 groepen vergelijken.
- Het voornaamste resultaat bekomen in WP3 heeft betrekking op de aangetoonde geslachtsverschillen in de genomische respons op blootstelling aan de onderzochte polluenten. Dit is ook waargenomen voor de effect-merkers. De geslachtsverschillen in de genomische reactiepatronen blijken substantieel te zijn. Bij toekomstige biomonitoring studies dient daarom met geslachtsverschillen expliciet rekening gehouden te worden. Van belang is erop te wijzen dat deze reacties optreden in vertegenwoordigers van de algemene Vlaamse bevolking: we spreken dus niet van extreme blootstellingen.
- De meest aangetroffen gedereguleerde netwerken bij mannen m.b.t. polluenten en effectmerkers (aangegeven tussen haakjes, waarbij effectmerkers geel gearceerd zijn) hebben betrekking op onderstaande processen. In een finale vergelijking werd aangetoond dat voornamelijk immunologische processen bij mannen en apoptotische reacties en neurophysiologische processen bij vrouwen werd beïnvloed. Deze processen en pathways hexachlorobenzenebiologisch relevant kunnen zijn. Deze informatie kan dan verder gebruikt worden om kandidaat biomerkers te selecteren voor verdere opvolging.
  - Transcription_P53 signaling pathway (cadmium, hydroxypyrene[1], pp'-DDE[2], hdg[3])
  - Transcription_Transcription regulation of aminoacid metabolism (TEQ,[4] pp'DDE, HCB[5] hdg)
  - Cell adhesion_ECM remodeling (TEQ, micronuclei, comet assay)
  - Immune response_IL-13 signaling via JAK-STAT (HCB, cea[6])
  - Immune response_PGE2 signaling in immune response (hpyr, PCB118[7])
  - Immune response_T cell receptor signaling pathway (ttma[8],HCB, comet assay)
  - Muscle contraction_ GPCRs in the regulation of smooth muscle tone (Cd, hdg, comet assay)
  - Muscle contraction_EDG5-mediated smooth muscle contraction (Cd, cea, p53)
  - PDGF activation of prostacyclin synthesis (Cd, TEQ, hdg)
  - Transcription_CREM signaling in testis (PCB118, hdg)

[1] 1-hydroxypyrene is een metaboliet van pyreen dat behoort tot de polyaromatische koolwaterstoffen(hpyr)
[2] *p,p´*-dichlorodiphenyldichloroethylene
[3] 8-hydroxy-guanine is een merker van DNA herstel
[4] TEQ: 2,3,7,8 dioxine toxiciteits equivalenten
[5] Hexachlorobenzene
[6] Cea:tumormerker carcino embryonaal antigen
[7] PCB118 : polychlorinated biphenyl 118
[8] Ttma: tt muconic acid is een benzeenmetaboliet

De meest aangetroffen gedereguleerde netwerken bij vrouwen m.b.t. polluenten en effectmerkers (aangegeven tussen haakjes, waarbij effectmerkers geel gearceerd zijn) hebben betrekking op

- N-Glycan biosynthesis p2 (Cd, lood, HCB, hdg)
- Development_TGF-beta-induction of EMT via ROS (Cd, hpyr, ttma, pp"DDE)
- Retinol metabolism (PCBs, TEQ, PCB118)
- Translation _Regulation of EIF4F activity (Cd, ttma,pp" DDE, micronuclei)
- Apoptosis and survival_Ceramides signaling pathway (Cd, ttma)
- Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim (Cd)
- Development_Growth hormone signaling via STATs and PLC/IP3 (p53)
- G-protein signaling_Rac3 regulation pathway (Cd, hdg)
- HIV-1 signaling via CCR5 in macrophages and T lymphocytes (ttma, hdg, p53)
- Keratan sulfate metabolism p.2 (Cd, pp"DDE, hdg)
- Neurophysiological process_ACM regulation of nerve impulse (micronuclei, comet assay, p53)
- Neurophysiological process_Glutamate regulation of Dopamine D1A receptor signaling (pp"DDE,PCB118, p53)
- Normal and pathological TGF-beta-mediated regulation of cell proliferation (Cd, ttma,pp'DDE)
- Profielen van veranderingen in genexpressie kunnen nu al ingesloten worden in humane biomonitoring in beleidsvoorbereidende context omdat ze een gevoelige merker zijn die op een unieke manier blootstellingspatronen van mengsels weergeeft. Op dit moment zijn de genexpressie merkers niet in staat om een uitspraak te doen over effectieve gezondheidsproblemen. Hiervoor dient complementair proteïnes en metabolieten gemeten te worden. De data dienen ook verdere in relatie gebracht worden met mogelijke klinische observaties. Op dit moment geven de genexpressies wel een duidelijke correlatie met complexe blootstelling. Dit betekent dat een complexe blootstelling een biologisch signaal induceert (ter hoogte van genexpressie). Dit vraagt verdere opvolging en bijgevolg fungeert genexpressie als een waarschuwingssignaal dat in het kader van preventieve gezondheidszorg zeer waardevol kan zijn. Een verdere opvolging kan gebeuren met zowel microarray technologie als real-time PCR (waarvoor verschillende methodes (nCounter, Fluidigm, etc.) ter beschikking zijn afhankelijk van de gewenste doorvoer (aantal stalen en aantal biomerker genen dat moet opgevolgd worden). Een ruw kost berekening voor VITO leert dat de kost om microarray data te genereren ongeveer 300 EUR bedraagt/staal (200 EUR producten en 100 EUR personeelskost). Indien meerdere stalen in parallel verwerkt worden zal de prijs dalen. De kost voor een Q-PCR experiment kan ruw geschat worden op 150 EUR/staal indien 10 genen in parallel getest worden. De vraag welke technologie dient gebruikt te worden voor verdere opvolging hangt niet alleen af van het budget dat ter beschikking is, maar eveneens van de capaciteit (beschikbaar personeel en middelen) voor data-analyse en interpretatie. Deze kosten zullen zwaarder doorwegen indien gekozen wordt voor microarray in plaats van real-time PCR. Anderzijds heeft microarray het voordeel dat er een massale hoeveelheid data wordt gegenereerd die voor verschilende vraagstellingen kan gebruikt worden. Microarray technologie is eerder hypothese-genererend, terwijl real-time PCR hypothese-testend is. Indien de keuze van kandidaat genen niet goed vooraf kan gedefinieerd worden, dan kan het interessanter zijn om microarray voor te stellen.

- Verder onderzoek zou zich moeten richten op het vertalen van deze moleculaire effecten naar morbiditeit en risico's op morbiditeit. Dit kan onder meer gebeuren door de populaties die bestudeerd werden in het kader van milieu- en gezondheidsstudies, en waarvan genexpressieresultaten beschikbaar zijn, verder op te volgen. Op dit moment zijn de genexpressieprofielen heel waardevol, maar ze geven vooral de link aan met blootstellingsmetingen. Dit kan reeds als een belangrijk waarschuwingssignaal beschouwd worden.
- Indien er een relatie kan gevonden worden met morbiditeit, dan dienen de meest belangrijke/voorspellende genen geïdentificeerd te worden. Vervolgens kan een low density screen voor gen expressie analyse ontwikkeld worden die op een gemakkelijker in bevolkingsonderzoek kan ingezet worden om de blootstelling aan mengsels en hun biologische consequenties in kaart te brengen.

7. Specifieke Beleidsvragen

Bieden meting en interpretatie de mogelijkheid om beleidsevaluatie te doen (bv. nulmeting en vervolgens periodiek zien of het de goede kant uit gaat)?

De methodes voor genexpressie zijn goed onder controle in zowel VITO als Universiteit Maastricht. De gegenereerde analyses zijn robuust en betrouwbaar. De huidige studie heeft de variabiliteit van genexpressie gedocumenteerd in een gezonde volwassene populatie. Hieruit blijkt dat genexpressie een vrij stabiele parameter is. De korte- en lange-termijn variabiliteit in genexpressie is beschreven. De genen die beïnvloed worden door een seizoen of die geslachtsafhankelijk zijn werden beschreven. Bovendien werd een inventaris gemaakt van genen met een sterk variabele expressie. Het inventariseren van deze genen is waardevol. Hoog variabele genen zijn waarschijnlijk minder geschikt om later als biomerker gebruikt te worden. Kandidaat merker genen die via andere manieren worden geïdentificeerd, worden best gecontroleerd op hun mogelijk geslacht- of seizoenseffect, alsook hun variabiliteit van expressie. De gegevens worden best gestockeerd in een database die als nulpunt kan fungeren voor latere onderzoeken bij eenzelfde doelgroep.

Werkpakket 3 observeerde significante veranderingen in genexpressie die gecorreleerd is met blootstelling aan polluenten. Een longitudinale studie kan meer informatie geven over de relevantie van de metingen. Enerzijds is er de optie om de metingen te correleren met potentiële morbiditeit. Anderzijds kunnen additionele metingen informatie geven over mogelijke persistente effecten op niveau van genexpressie. Bevestiging van de genexpressieprofielen over verschillende tijdspunten heen, zou de waarde van genexpressie in biomonitoring verder bevestigen. Een kwantificeerbare link leggen met morbiditeit/mortaliteit is momenteel te vroeg. Volgende strategieën zijn daarvoor aangewezen: 1) opname van genexpressiemetingen in prospectieve cohort studies, 2) verband onderzoeken van de relatie tussen genexpressieprofielen en gevaloriseerde effectmerkers of klinische parameters.

Kunnen stijgingen of dalingen van bepaalde expressies (minstens) vertaald worden naar een (semi-)kwantificeerbare mate van morbiditeit of mortaliteit?

Kwantificeerbare link leggen met morbiditeit/mortaliteit is momenteel te vroeg . Volgende strategieën zijn daarvoor aangewezen: 1) opname van genexpressiemetingen in prospectieve cohort studies, 2) verband onderzoeken van de relatie tussen genexpressieprofielen en gevaloriseerde effectmerkers of klinische parameters.

Kunnen meting en interpretatie aanleiding geven tot beleidsvoorbereiding en zo ja voor:

- acties voor blootstellingspreventie op doelgroepniveau?
- acties voor sensibilisering op doelgroepniveau?
- acties of op vlak van regiobrede normstelling (emissie - immissie)?

Profielen van veranderingen in genexpressie kunnen ingesloten worden in humane biomonitoring in beleidsvoorbereidende context omdat ze een gevoelige merker zijn die op een unieke manier blootstellingspatronen van mengsels weergeeft met een eerste vertaling naar mogelijk biologische

impact. De genexpressie metingen hebben een signaalfunctie die verdere opvolging vraagt. De waarde van dergelijke merkers kan slechts tot uiting komen wanneer genexpressiemetingen systematisch parallel worden uitgevoerd met andere biomerker- en effectmetingen zodat de relatieve gevoeligheid en vroege signaalfunctie kan worden bepaald. We kunnen verwachten, dat naar analogie met de klinische geneeskunde, genexpressie in de milieugezondheidscontext zal kunnen worden toegepast als biomerker.