Peter VAN LOO — IDENTIFICATION OF REGULATORY REGIONS AND DISEASE CAUSING GENES AND MECHANISMS — Mei 2008

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT GENEESKUNDE
DEPARTEMENT MENSELIJKE ERFELIJKHEID
Herestraat 49, bus 602, B–3000 Leuven

SYSTEMS BIOLOGY: IDENTIFICATION OF REGULATORY REGIONS AND DISEASE CAUSING GENES AND MECHANISMS

Promotor:
Prof. P. Marynen

Co-promotoren:
Prof. C. De Wolf-Peeters
Prof. B. De Moor

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de medische wetenschappen

door

Peter VAN LOO

Mei 2008

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT GENEESKUNDE
DEPARTEMENT MENSELIJKE ERFELIJKHEID
Herestraat 49, bus 602, B–3000 Leuven

SYSTEMS BIOLOGY: IDENTIFICATION OF REGULATORY REGIONS AND DISEASE CAUSING GENES AND MECHANISMS

Jury:

Prof. E. Legius, voorzitter
Prof. P. Marynen, promotor
Prof. C. De Wolf-Peeters, copromotor
Prof. B. De Moor, copromotor
Prof. J. Vermeesch
Prof. H. Ceulemans
Prof. V. Timmerman, Universiteit Antwerpen
Prof. L. Wessels, Nederlands Kanker Instituut
Prof. E. Barillot, Institut Curie

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de medische wetenschappen

door

Peter VAN LOO

Mei 2008

© Katholieke Universiteit Leuven
Faculteit Geneeskunde
Departement Menselijke Erfelijkheid
Herestraat 49, bus 602
B–3000 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of this publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

The figure on the cover represents different aspects of this work. At the left, eight graphs are shown from which the difference in location of *cis*-regulatory modules prediction made by ModuleMiner in adult tissues and in embryonic development can be derived (see also Figure 5 in Study 2). At the top right, a netwerk is shown, generated by STRING, from all the components we highlighted in Table 2 in Study 3, for their potential importance in the specific immune response occurring in THRLBCL. At the bottom right, a schematic is shown that represents the principle of prioritization by Endeavour (for more information, see Figure 1 in Study 1).

Voorwoord

Als je zes jaar geleden zou gevraagd hebben of ik zou willen doctoreren, dan zou ik resoluut “nee” gezegd hebben. Ik dacht dat ik min of meer voorbestemd was een eigen zaak te beginnen. “De goden hebben anders beslist.” Nu ja, eigenlijk heb ik anders beslist. En dat is één van de twee beste beslissingen van mijn leven gebleken. Maar misschien zijn het wel goden die een beetje geholpen hebben bij die beslissing. Twee goden om precies te zijn.

De eerste heb ik al in een eerder voorwoord “de god van het enthousiasme” genoemd. Toen ik mijn thesis moest kiezen, wou ik in de richting van de bioinformatica gaan, omdat dat goed aansloot bij mijn curriculum, en omdat ik daar wel toekomst in zag. Zo kwam ik al snel bij Stein Aerts terecht, want zijn enthousiasme werkte heel aanstekelijk. Het is Stein die mijn eerste interesse in wetenschappelijk onderzoek heeft opgewekt. Het was een heel toffe samenwerking die uiteindelijk veel langer is blijven lopen dan die ene thesis.

Het volgende jaar moest ik nog een thesis maken, één die iets meer biologisch gericht was. Dus ik besloot om de methoden die we ontwikkeld hadden te gaan testen in het labo. Maar ik was niet zo geïnteresseerd in plantjes, bacteriën of gisten. En zo kwam ik dan terecht op de torenhoge Olympusberg (hét excuus om niet met de fiets te komen!). En de lezer met veel fantasie glimlacht een beetje, want hij heeft nu een levendig beeld van de tweede god.

Peter Marynen. Iedereen verwees me naar hem door voor een dergelijk thesisonderwerp, en hij was ook meteen enthousiast, ook al was het niet helemaal bekend terrein voor hem. Het klikte meteen, en ook aam die thesis heb ik hele goede herinneringen. Peter wist mijn wetenschappelijke interesses helemaal los te krijgen. Dus ik besloot om te doctoreren. En ik zag dat het goed was.

Mijn doctoraat deed ik opnieuw bij Peter Marynen. “En ik zal u zeggen waarom.” Peter wist altijd tot de kern van de zaak door te dringen (ook als je dat soms liever niet had). Ook kon ik Peter’s rechttoe-rechtaan houding heel erg appreciëren. Het was altijd duidelijk waaraan je nog moest werken en wat hij goed vond. Verder was Peter de hoofddocent van de sleutelcursus “kritisch wetenschappelijk denken”, ongetwijfeld het grote hoofdvak in de opleiding tot wetenschapper. Een heel gepaste term om de promotor-doctoraatsstudent relatie te beschrijven is “symbiose”: een samenwerking waar beide organismen voordeel uit halen. Wat die term niet zegt is dat beide organismen evenveel voordeel halen. En ik ben ervan overtuigd dat de balans hier naar mijn kant overhelde. Ik ben Peter ook ontzettend dankbaar voor alle kansen die hij mij gaf en om me actief op te leiden als wetenschapper door zijn voorbeeld, zijn kennis, zijn commentaren en de vele kritische discussies. Tenslotte ben ik Peter heel dankbaar voor de vrijheid die hij me gaf in mijn onderzoek - een vrijheid ongezien voor een doctoraatsstudent, die soms toch wel onverwachte wegen uitging (vergelijk dit boekje maar met het oorspronkelijke doctoraatsproject), maar waar ik me heel goed in kon vinden.

Naast Zeus wil ik nog een andere godheid nadrukkelijk vernoemen die het licht in de duisternis heeft gebracht. Chris Peeters. Misschien is Athene hier wel een goede vergelijking. Ik ben pas redelijk laat in mijn "carrière" als doctoraatsstudent bij Chris terechtgekomen, maar ik denk dat ik meer tijd van haar gekregen heb dan de gemiddelde doctoraatsstudent van zijn promotor krijgt in vier jaar tijd. Chris stond altijd klaar met wijze raad, over wetenschap of over andere zaken. Ook haar doorzicht in de pathologie en de klinische kant van de zaak hebben een wezenlijk verschil gemaakt. Chris, ik ben dikwijls met veel plezier vroeger voor u opgestaan - en sommigen weten dat ik het daar normaal gezien heel moeilijk mee heb. Ik hoop dat je als emeritus nog vele jaren met veel plezier kan bezig zijn met je hobby: de tuin. En ik hoop ook dat je daarbuiten nog precies zoveel als je wil met de kliniek en de wetenschap mag bezig zijn. En ik moet toegeven: eigenlijk hoop ik stiekem ook een beetje dat ik nog met je mag samenwerken.

En dan komen we terug bij de oorspronkelijke roots: ESAT en bioinformatica. Geografisch gezien is dit niet op de Olympusberg gelegen, dus laat me dit dan plaatsen in het oude Griekenland, meer bepaald Ithaka. Hier begin ik mijn dankwoord bij Agamemnon (Bart De Moor). Hem ben ik zeer dankbaar voor de kansen die hij me gegeven heeft voor en tijdens mijn doctoraat, en ook voor de financiering van een aantal reizen ver buiten het Griekse rijk. Ook Odysseus (Yves Moreau) ben ik heel dankbaar voor de wetenschappelijke discussies, voor de nauwe en vruchtbare samenwerking en voor vele praktische zaken. Verder bedank ik ook alle inwoners van Ithaka en het oude Griekenland. Ik was misschien maar een halfbloed, maar ik heb me bij jullie toch altijd thuis gevoeld.

Ook heb ik het voorrecht gehad om met vele andere goede wetenschappers samen te werken, waaruit ik veel heb kunnen leren en waarvoor ik oprecht dankbaar ben. Ik wil er enkele bij naam noemen. Diether Lambrechts speelde een zeer belangrijke rol bij de ontwikkeling en de validatie van Endeavour. Patrick Matthys zorgde in onze lymfoma-studie voor de uiteindelijke mechanistische interpretatie van de expressieprofielen. In addition, I would like to thank Iwona Wlodarska, Frans Schuit, Gregor Verhoef, Daan Dierickx, Jan Delabie, Agnieszka Malecka, Bernard Thienpont, Leo Tranchevent, Bert Coessens, Joos Vandewalle, Stefan Lehnert and Vera Vanhentenrijk for the pleasant and interesting collaboration within or outside my PhD project. Tenslotte wil ik nog expliciet Isabelle Vanden Bempt bedanken: onze samenwerking is ook een heel mooi voorbeeld van symbiose geweest. Bedankt Isabelle, voor de interessante, leerrijke en geanimeerde discussies over lopende en nog-niet-helemaal-lopende projecten! En weet ook dat je hulp bij de vele praktische zaken een groot verschil hebben gemaakt. Ik hoop alvast dat we in de toekomst de symbiose nog mogen verderzetten!

Als je de Griekse mythologie mag geloven, werd er op de Olympusberg ook wel eens flink gefeest. Many thanks, CB3, for the nice atmosphere, the unforgettable lab weekends and many other activities, and for all the friendship!

I would like to thank the members of the jury, professors E. Legius, J. Vermeesch, H. Ceulemans, V. Timmerman, L. Wessels and E. Barillot, for providing me with valuable comments and suggestions which improved this PhD text.

Vrienden, jullie ben ik ontzettend dankbaar voor de toffe sfeer die we altijd hadden bij onze spelletjesavonden, etentjes, cantussen, weekendjes, . . . Door die ontspanning kon ik altijd met vernieuwde energie weer aan het werk gaan.

Mijn familie en schoonfamilie wil ik bedanken voor alle hulp en steun. Een extra woord van dank ook aan onze moe's en va's, Cynthia en mijn broer Johan, om altijd klaar te staan voor mij. En natuurlijk mama en papa: dank je wel voor alle kansen die jullie me gaven, voor jullie vertrouwen in al wat ik deed, voor de interesse en voor de steun die jullie me gegeven hebben, niet alleen tijdens mijn doctoraatsjaren maar in mijn hele leven. Zonder jullie zou ik hier niet geraakt zijn!

Tenslotte mijn - goddelijke - vrouw Tina. Hier weet ik echt niet waar te beginnen. Bedankt, Tina, voor al je goede zorgen, voor je steun en voor je begrip als ik weer maar eens wat minder tijd had dan verwacht. En bedankt ook voor wie je bent. Jij bent de beste beslissing uit mijn leven geweest!

Waar goden zijn, is er ook een schepping.
Hier is ze dan, mijn schepping!
Maar is dat wel de schepping, dit doctoraatsboekje? Is het niet iets dieper, waar dit boekje symbool voor staat?
Ja, ik weet het:
Hier ben ik dan: een wetenschapper. Jullie schepping! Bedankt.

Peter, mei 2008

Contents

Voorwoord **i**

Contents **v**

List of abbreviations **ix**

Summary **xi**

Samenvatting **xiii**

General introduction **1**
- 1 Systems biology . . . 1
- 2 Key examples of systems biology approaches . . . 2
  - 2.1 Microarray gene expression profiling . . . 2
  - 2.2 Cross-context gene expression profile mapping . . . 2
  - 2.3 Network based systems biology methods . . . 3
- 3 Gene prioritization . . . 5
- 4 Computational *cis*-regulatory module detection . . . 6
  - 4.1 Type I CRM detection methods . . . 7
  - 4.2 Type II CRM detection methods . . . 11
    - Type IIa methods . . . 11
    - Type IIb methods . . . 13
  - 4.3 Type III CRM detection methods . . . 15
- 5 Conclusions . . . 17

Rationale and aims **19**
- 1 Development of novel systems biology methods . . . 19
  - Study 1: Gene prioritization through genomic data fusion . . . 19
  - Study 2: ModuleMiner: improved computational detection of *cis*-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues? . . . 20
- 2 Applications to cancer stratification and understanding . . . 21
  - Study 3: T cell/histiocyte rich large B cell lymphoma shows a tolerogenic host immune response: the lymphoma microenvironment as a target for therapy . . . 21

Study 4: Polysomy 17 in breast cancer: clinicopathological significance and impact on HER2 testing 22

1 Gene prioritization through genomic data fusion 25
Abstract 26
Introduction 26
Results 26
  Principles of prioritization used by Endeavour 26
  Validation of Endeavour when accessing individual data sources 27
  Overall prioritization by fusing multiple data sources 27
  Endeavour does not rely on literature-derived data alone 29
  Use of disease-specific data sources 30
  Prioritization of genes causing monogenic diseases 30
  Prioritization of genes underlying polygenic diseases 30
  Prioritization of regulatory pathway genes 30
  Functional validation of Endeavour in zebrafish 31
Discussion 31
Methods 32
References 33

2 ModuleMiner: computational CRM detection 35
Abstract 37
List of abbreviations 37
Background 38
Results 41
  ModuleMiner: detection of similar *cis*-regulatory modules in a set of co-expressed genes 41
  Sensitivity to noise 45
  Comparison to other CRM detection algorithms 47
  Detection of *cis*-regulatory modules in microarray clusters . . . 50
  Detection of *cis*-regulatory modules in embryonic development gene sets 54
  Characterization of the *cis*-regulatory modules 55
  Where are the *cis*-regulatory module predictions located? . . . 57
Discussion 62
Conclusions 70
Materials and methods 71
References 77

3 A tumour tolerogenic microenvironment in THRLBCL 81
Abstract 83
Introduction 84
Materials and methods 85
Results and discussion 87
  Clinical data 87
  Expression profiling of THRLBCL vs. NLPHL 88

    The gene expression signature of NLPHL 90
    The gene expression signature of THRLBCL 93
    Absence of T cell genes in the NLPHL and THRLBCL gene expression signatures 96
    THRLBCL and NLPHL gene expression signatures in additional cases 97
  Conclusion 97
  References 98

4 Significance of polysomy 17 in breast cancer 101
  Abstract 103
  Introduction 104
  Materials and methods 106
  Results 110
    Comparison between IHC and FISH for HER2 testing 110
    Impact of polysomy 17 on HER2 testing results 110
    Stratification 111
    HER2 mRNA expression by quantitative RT-PCR 112
    Clinicopathological characteristics of polysomy 17 tumors . . . 113
  Discussion 115
  References 119

General discussion and perspectives 123
  1 Computational methods for gene prioritization and *cis*-regulatory module detection 123
    1.1 ModuleMiner: *cis*-regulatory module detection 123
    1.2 Endeavour: gene prioritization by genomic data fusion . 125
    1.3 Integrated case-study: macrophage differentiation 127
  2 Towards a better stratification and understanding of cancer . . 129
    2.1 Microarray expression profiling to gain insight into the lymphoma microenvironment 129
    2.2 Stratification of breast cancer by HER2 amplification status and presence of polysomy 17 130
    2.3 Clinicopathological implications 131
  3 Long term perspectives 131
    3.1 Systems biology methods for the identification of regulatory regions 131
    3.2 Systems biology methods for disease gene identification 132
    3.3 Systems biology methods for disease mechanism elucidation 132
    3.4 Data-driven identification of disease genes and mechanisms 133

Bibliography 135

Curriculum Vitae 147

List of abbreviations

ASCO	American Society of Clinical Oncology
AUC	area under the (ROC) curve
BIND	Biomolecular Interaction Network Database
BLAST	Basic Local Alignment Search Tool
CAP	College of American Pathologists
CGH	comparative genomic hybridisation
ChIP	chromatin immunoprecipitation
CNS	conserved non-coding sequence
CRM	*cis*-regulatory module
DGS	DiGeorge syndrome
DLBCL	diffuse large B cell lymphoma
DNA	deoxy-ribonucleic acid
EEL	Enhancer Element Locator
ER	estrogen receptor
EST	expressed sequence tag
FISH	fluorencence *in situ* hybridisation
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HUGO	Human Genome Organisation
IHC	immunohistochemistry
IPI	International Prognostic Index
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOOCV	leave-one-out cross-validation
LRA	logistic regression analysis
LRCHL	lymphocyte rich classical Hodgkins lymphoma
mRNA	messenger ribunucleic acid
NLPHL	nodular lymphocyte predominant Hodgkin's lymphoma
NPI	Nottingham Prognostic Index
OMIM	Online Mendelian Inheritance in Man
PCR	polymerase chain reaction
PFR	phylogenetically footprinted non-coding region
PMA	phorbol 12-myristate 13-acetate
PR	progesterone receptor
PWM	position weight matrix
RNA	ribonucleic acid

ROC	receiver operating characteristic
RT-PCR	reverse transcriptase PCR
SNP	single nucleotide polymorphism
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TFBS	transcription factor binding site
THRLBCL	T cell/histiocyte rich large B cell lymphoma
TLR	Toll-like receptor
TRGM	transcriptional regulatory global model
TRM	transcriptional regulatory model
TSS	transcription start site
WHO	World Health Organization

Summary

Systems biology is a relatively novel field of science aiming to study biological systems in an integrated fashion. Systems biology entails an elaborate spectrum of methods including *in vitro* techniques (such as microarray gene expression profiling) as well as many bioinformatics approaches.

In this work, we aim to develop novel bioinformatics and systems biology approaches to two important biological problems: the identification of genes involved in diseases and the annotation of regulatory regions in the human genome. In a second independent part, we apply bioinformatics, systems biology and statistics to gain a better understanding of two cancer types: lymphoma and breast cancer.

We developed Endeavour, a novel systems biology method to prioritize candidate genes for a given disease or biological process, based on similarities to a set of user-defined “training genes”. Contrary to existing methods, Endeavour is able to integrate information from multiple heterogeneous data sources. We validated Endeavour extensively by a large-scale leave-one-out cross-validation, and by specific case-studies on recently identified disease genes.

We developed ModuleMiner, a computational method to detect similar *cis*-regulatory modules (CRMs) in a set of co-regulated genes. In contrast to existing methods, ModuleMiner is parameterless and performs a whole-genome optimization maximizing specificity of the CRMs for the given co-regulated genes, compared to all other genes in the genome. By direct comparison, we demonstrated that ModuleMiner outperforms other existing computational methods to detect CRMs. Additional validation experiments showed that ModuleMiner is insensitive to noise in the given set of co-regulated genes, provided a critical mass of genes with similar CRMs is present. We applied ModuleMiner on a larger scale, showing that CRMs can be detected in microarray clusters containing genes expressed in different adult tissues, as well as in custom-build gene sets related to embryonic development processes. Comparing both sets of CRM predictions leads us to hypothesize that CRMs driving expression in adult tissues are mostly proximal promotors, while CRMs involved in embryonic development are more distal enhancers.

We performed an integrated validation, combining CRM detection with Endeavour gene prioritization. Using CRM detection methods, we predicted 100 genes differentially regulated during HL-60 differentiation (as a proxy for the final stages of differentiation of hematopoietic stem cells to macrophages), and validated this by real-time quantitative RT-PCR for 20 of these genes. Although our results indicate that a significant signal is picked up, the amount of differential regulation was limited. We used Endeavour to prioritize the 100 genes predicted to be differentially regulated, and we again validated the 20 highest ranking genes by real-time quantitative RT-PCR. We obtained a significant difference between the genes after prioritization and the genes before prioritization. These results indicate that predicting new target genes based on a model of similar CRMs in a set of co-regulated genes in the human genome is possible, and that Endeavour can aid in the selection of true target genes.

Using microarray gene expression profiling, we studied the microenvironment in two related lymphoma entities with a markedly different prognosis: the indolent nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL) and the aggressive T cell/histiocyte rich large B cell lymphoma (THRLBCL). While the expression signature of NLPHL mainly consists of B cell genes, in line with expectations, the expression signature of THRLBCL is characterized up-regulation of CCL8, IFN-γ, IDO, VSIG4 and Toll-like receptors, and reflects the recruitment and activation of histiocytes/macrophages, a tumour tolerogenic microenvironment, and innate immune responses. Several of these characteristics offer potential targets for a directed therapy in THRLBCL.

Amplification of the gene HER2, located on chromosome 17, is a frequent event in breast cancer and defines a distinct subgroup associated with a bad prognosis. The relatively frequent occurrence of polysomy 17 may complicate the interpretation of tests for HER2 amplification. We investigated the impact of polysomy 17 on HER2 testing and studied its clinicopathological significance in relation to HER2 gene amplification. We observed that all cases with an inconclusive or "equivocal" HER2 result by fluorescence *in situ* hybridisation (FISH) are polysomic for chromosome 17. Polysomy 17 without HER2 amplification was not associated with HER2 overexpression. In addition, in contrast to HER2 amplification, polysomy 17 was not associated with high tumour grade, hormone receptor negativity or reduced disease-free survival. These results indicate that polysomy 17 is clinicopathologically distinct from true HER2 gene amplification, suggesting that HER2-targeted therapies are unlikely to be successful in tumours with polysomy 17 but no HER2 gene amplification.

Samenvatting

Systeembiologie is een relatief nieuwe tak van de wetenschap die biologische systemen op een geïntegreerde wijze bestudeert. Systeembiologie omvat een breed gamma aan methoden, inclusief *in vitro* technieken (zoals microrooster expressieprofilering) en vele bioinformatica methoden.

In dit werk ontwikkelen we nieuwe bioinformatica en systeembiologie methoden voor twee belangrijke biologische problemen: de zoektocht naar genen betrokken bij erfelijke aandoeningen en de annotatie van regulatorische gebieden in het menselijk genoom. In een tweede deel passen we bioinformatica, systeembiologie en statistiek toe om twee types van kanker beter te begrijpen: lymfomen en borstkankers.

We ontwikkelden Endeavour, een nieuwe systeembiologie methode om kandidaatgenen voor een erfelijke aandoening of een biologisch proces te prioritizeren, gebaseerd op gelijkenissen met een verzameling van "trainingsgenen", door de gebruiker opgegeven. In tegenstelling tot bestaande methoden kan Endeavour gegevens integreren van verscheidene heterogene gegevensbronnen. We voerden een uitgebreide validatie uit van Endeavour, door "leave-one-out cross-validation", en we deden een aantal gevallenstudies van recent geïdentificeerde ziektegenen.

We ontwikkelden ModuleMiner, een computationele methode om gelijkaardige *cis*-regulatorische modules (CRMs) te detecteren in een verzameling van co-gereguleerde genen. In tegenstelling tot bestaande methoden is ModuleMiner parameterloos en voert het algoritme een optimalisatie uit over het volledige genoom, waarbij de specificiteit voor de opgegeven verzameling van co-gereguleerde genen geoptimaliseerd wordt, in vergelijking met alle andere genen in het genoom. Door directe vergelijking met andere methoden konden we aantonen dat ModuleMiner een hogere performantie heeft. Verdere validatie-experimenten toonden aan dat ModuleMiner niet gevoelig is aan ruis in de opgegeven set van co-gereguleerde genen, op voorwaarde dat een kritische massa van genen met gelijkaardige CRMs aanwezig is. We pasten ModuleMiner toe op grotere schaal, waarbij we aantoonden dat het algoritme CRMs kan vinden in microrooster clusters die genen bevatten die tot expressie komen in verschillende volwassen weefsels, alsook in verzamelingen van genen betrokken bij embryonale ontwikkelingsprocessen. Als we beide groepen van CRM predicties vergeleken, kwamen we tot de hypothese dat CRMs betrokken bij expressie in volwassen weefsels vooral te vinden waren in proximale promotoren, terwijl

CRMs betrokken bij embryonale ontwikkeling eerder meer distale enhancers zijn.

We voerden een geïntegreerde validatie uit, waarbij we CRM detectie combineerden met Endeavour genprioritizatie. Gebruik makend van CRM detectiemethoden, voorspelden we 100 genen die differentieel gereguleerd zijn bij de differentiatie van de hematopoietische cellijn HL-60. Dit valideerden we vervolgens met kwantitatieve PCR voor 20 genen. Hoewel onze resultaten erop wijzen dat we een significant signaal kunnen oppikken, toch was de sterkte van de differentiële regulatie beperkt. We gebruikten Endeavour om de lijst van 100 voorspelde genen te prioritizeren, en we valideerden de 20 hoogst gerangschikte genen. We observeerden een significant verschil tussen de genen na prioritizatie en de genen voor prioritizatie. Deze resultaten wijzen erop dat nieuwe doelgenen, gebaseerd op een model van gelijkaardige CRMs in co-gereguleerde genen kunnen voorspeld worden, en dat Endeavour de correcte doelgenen kan selecteren.

Met microrooster expressieprofilering bestudeerden we de micro-omgeving in twee verwante lymfoom-entiteiten met een sterk verschillende prognose: de indolente nodulair lymfocyt predominante Hodgkin lymfomen (NLPHL) en de agressieve T cel/histiocyt rijke grootcellige B cel lymfomen (THRLBCL). In de expressiesignatuur van NLPHL zagen we voornamelijk B cel genen, in lijn met onze verwachtingen. De expressiesignatuur van THRLBCL is gekarakteriseerd door op-regulatie van CCL8, IFN-γ, IDO, VSIG4 en toll-like receptoren, en reflecteert de recrutering en activatie van macrofagen en dendritische cellen, een tumor tolerogenische micro-omgeving, en reacties van het aangeboren immuunsysteem. Deze karakteristieken bieden mogelijke doelwitten voor een gerichte therapie in THRLBCL.

Amplificatie van het gen HER2, gelegen op chromosoom 17, is een frequente genomische aberratie in borstkanker, en definieert een aparte subgroep met een slechte prognose. Het relatief frequente voorkomen van polysomie 17 zou de interpretatie van testen voor HER2 amplificatie kunnen bemoeilijken. We onderzochten de impact van polysomie 17 op testen voor HER2 amplificatie en we bestudeerden de clinicopathologische relevantie van polysomie 17 in verhouding tot HER2 genamplificatie. We zagen dat alle gevallen met een onbeslist of “equivocaal” HER2 resultaat polysomie 17 vertoonden. Polysomie 17 zonder HER2 amplificatie was niet geassocieerd met HER2 overexpressie. In tegenstelling tot HER2 amplificatie, was polysomie 17 niet geassocieerd met hoge tumorgraad, negativiteit voor hormoonreceptoren en lagere ziekte-vrije overleving. Deze resultaten wijzen erop dat polysomie 17 clinicopathologisch verschillend is van amplificatie van het HER2 gen. Dit lijdt ons tot de hypothese dat therapieën gericht tegen HER2 geen effect hebben in borstkankers met polysomie 17 maar zonder HER2 amplificatie.

General introduction

1 Systems biology

Biologists have historically applied a rigorous reductionist approach to their field of study. In this approach, a complex system is broken down into its individual components and these components are analyzed as much as possible in isolation. This approach has been highly successful. However, it is clear that this reductionist approach alone will not provide a full understanding of the complex and interwoven biological processes that take place in the cell and in the organism. Indeed, evolution optimizes the properties of the system as a whole instead of optimizing all individual parts. Therefore, in addition to this in depth reductionist approach, a complementary integrated approach is required. The advent of (near) whole genome sequences (Fleischmann *et al.*, 1995; *C. elegans* Sequencing Consortium, 1998; Adams *et al.*, 2000; Lander *et al.*, 2001; International Human Genome Sequencing Consortium, 2004; Waterston *et al.*, 2002; Gibbs *et al.*, 2004) for the first time provided a really integrated glimpse of biological systems, and the term "systems biology" was coined for this integrated approach. The birth of numerous genome-wide technologies (such as microarray expression profiling and derived techniques) and the resulting availability of many genome-wide data sets (Birney *et al.*, 2004; Harris *et al.*, 2004; Bader *et al.*, 2001; Su *et al.*, 2004; Son *et al.*, 2005; Rhodes *et al.*, 2004) has cultivated very high expectations for this approach in recent years.

Systems biology is a relatively novel field of science that is only broadly defined. Many avenues have been explored and perhaps even more are at this moment unexplored. We do not aim to give a full overview of systems biology approaches here, but instead we focus on a few key examples of systems biology methods, including the microarray gene expression profiling technique and its associated systems biology analysis methods, and network based systems biology approaches. Furthermore, we briefly describe existing systems biology methods for the prioritization of candidate genes. Finally, we will give a detailed overview of the existing computational approaches aiming to identify *cis*-regulatory modules, a specific field of bioinformatics that is gradually entering the realm of systems biology.

2 Key examples of systems biology approaches

2.1 Microarray gene expression profiling

Microarrays have become an established approach in biological research. Variants of this technology have been developed for many uses: gene expression profiling (Schena *et al.*, 1995; Lockhart *et al.*, 1996), DNA copy number detection (array-CGH), chromatin immunoprecipitation on a chip (ChIP-chip) (Ren *et al.*, 2000), SNP genotyping, ... In addition, protein-binding microarrays (Mukherjee *et al.*, 2004) can measure the affinity of transcription factors to the DNA. In the further text, we will focus on gene expression microarrays.

Microarray gene expression profiling aims to measure the expression of thousands of genes simultaneously by hybridization of labeled RNA (or cDNA) to complementary sequences that are "arrayed" on a chip. The intensity of each arrayed spot can then be linked to the RNA abundance of the corresponding gene.

Although not all applications of this technique can unquestionably be called systems biology, we do state that gene expression microarrays have potentiated a key change in the classically reductionist approach of the biologist. Indeed, in contrast to more advanced "real" systems biology approaches (we will discuss examples below), virtually all biologist are familiar with microarrays.

Even though microarrays provide large amounts of data, it has also become clear that data does not equal knowledge, resulting in the emergence of a new field of gene expression data analysis aiming to extract this knowledge from the microarray data (reviewed in Allison *et al.* (2006)). Most widely used are methods for making inference from microarray data (e.g. obtaining differentially expressed genes), and methods for classification (e.g. supervised or unsupervised clustering).

Microarray expression profiling and its associated analysis methods have provided some key biological results. Firstly, these techniques have led to a better stratification of cancer (Alizadeh *et al.*, 2000; Perou *et al.*, 2000; Sørlie *et al.*, 2001; Lee *et al.*, 2006; Yeoh *et al.*, 2002; Carrasco *et al.*, 2006). In addition, microarrays expression profiling has also been useful for outcome prediction (van 't Veer *et al.*, 2002; Rosenwald *et al.*, 2002; Shipp *et al.*, 2002; Bullinger *et al.*, 2004). Finally, these techniques have lead to a better understanding of cancer and disease (Dave *et al.*, 2004; Lamb *et al.*, 2003; Ramaswamy *et al.*, 2003), although perhaps in this last regard microarrays may not have entirely lived up to our expectations.

2.2 Cross-context gene expression profile mapping

Comparing results from different gene expression profiling experiments performed on different platforms or in different labs has proven difficult (Tan *et al.*, 2003; Fortunel *et al.*, 2003). One advanced method developed to overcome these difficulties is Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005). The idea behind this approach is to rank the genes in a microarray experiment (e.g. by correlation with phenotype) and to evaluate the positions of a given set of genes (e.g. from another microarray experiment or from literature) in this ranked gene list. If many of the given genes occur either high or low in the ranking, this points to similarities between the gene set and the results of the experiment. Finally, GSEA assigns a probability to this combination of rankings.

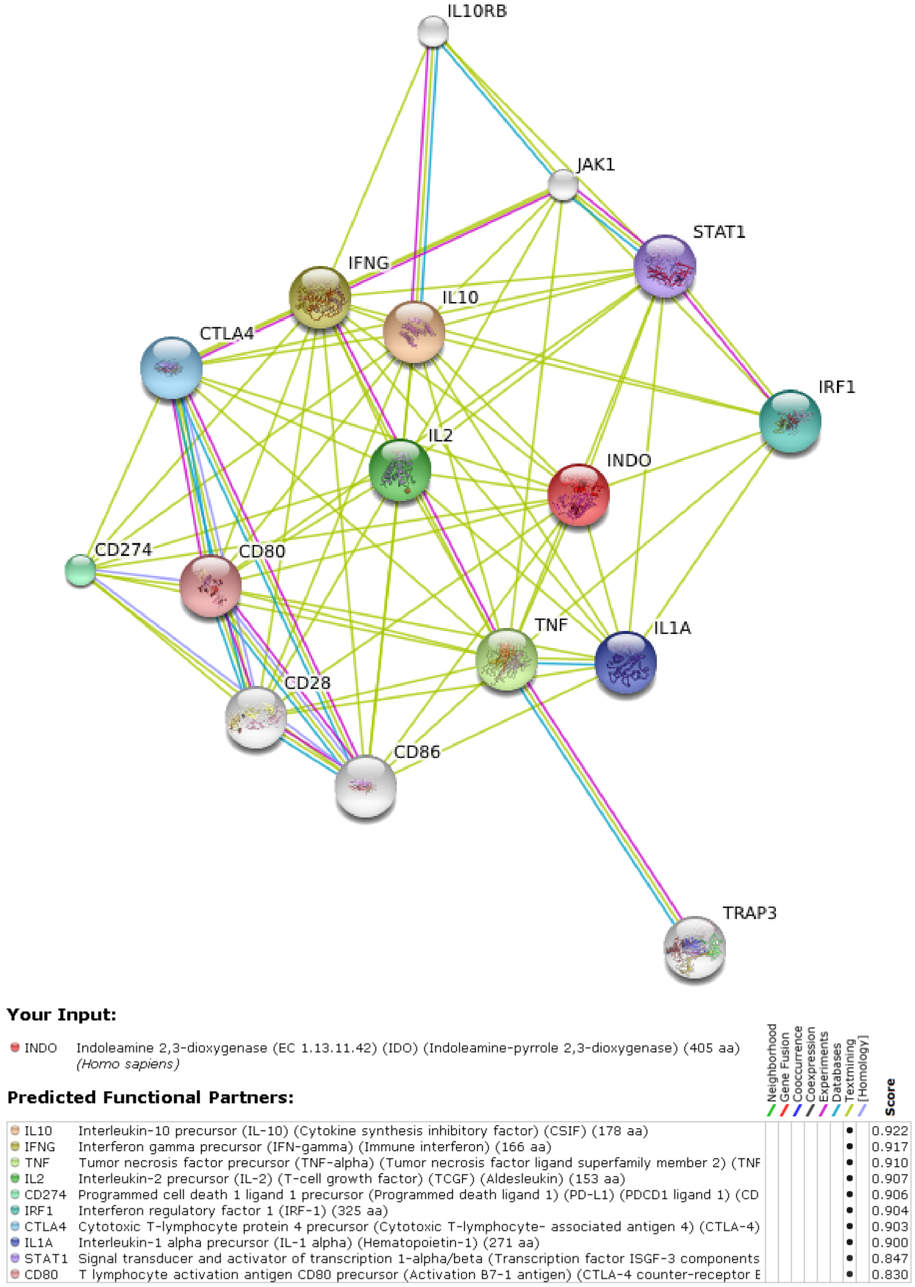
Pushing this idea one step further, gene expression data of experiments in a totally different context can also be correlated, opening up new avenues. In one study, Lamb *et al.* (2006) created a reference microarray data set of cultured human cells treated with a library of bioactive small molecules. They reasoned that if the genes found upregulated in a disease state versus a normal control are downregulated in a cell line treated with a certain compound, this suggests that that compound might counteract the effect of that disease. They used an approach similar to GSEA to link the “expression profiles” of bioactive chemical compounds to the expression profile of the disease state. In the resulting framework, termed the “Connectivity Map”, small molecule compounds are ranked, given a second focussed microarray experiment.

2.3 Network based systems biology methods

A network is an interconnected group of entities. Because a natural way to represent biological data is under the form of interconnected proteins or genes, networks are ubiquitous in biology. Much information can be obtained from these biological network: e.g. the closer two genes/proteins are within the network, the more tightly they are related. As many properties of networks may not be apparent when studying the individual components of the network, many system biology methods have been developed that focus on networks. The most well known of these is STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (von Mering *et al.*, 2007). STRING builds a protein-protein interaction network by integrating known and predicted interactions from many different data sources (including text mining) and many different organisms. Figure 1 shows the network STRING constructs around the gene IDO/INDO.

Yıldırım *et al.* (2007) constructed a bipartite graph of proteins and drugs, linked by drug-target associations. They found that drugs of similar types clustered tightly together, and that etiological drugs show a closer relationship with disease-gene products than palliative drugs. In addition, they identified properties of drug design, such as an overabundance of “follow-on” drugs (drug that target already targeted proteins) and increased focus in recent drug development.

In another study of the same group (Pujana *et al.*, 2007), a network was constructed around 4 known breast cancer genes, combining gene expression profiling and integration of functional genomic and proteomic data. This network allowed them to predict that the gene HMMR is associated with higher breast cancer susceptibility, which they subsequently validated (Pujana *et al.*, 2007). In two other recent studies (Franke *et al.*, 2006; Lage *et al.*, 2007) (discussed in detail below), this principle of using networks to identify disease genes



Your Input:

● INDO Indoleamine 2,3-dioxygenase (EC 1.13.11.42) (IDO) (Indoleamine-pyrrole 2,3-dioxygenase) (405 aa) *(Homo sapiens)*

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	**Score**
● IL10	Interleukin-10 precursor (IL-10) (Cytokine synthesis inhibitory factor) (CSIF) (178 aa)							●		0.922
● IFNG	Interferon gamma precursor (IFN-gamma) (Immune interferon) (166 aa)							●		0.917
● TNF	Tumor necrosis factor precursor (TNF-alpha) (Tumor necrosis factor ligand superfamily member 2) (TNF							●		0.910
● IL2	Interleukin-2 precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin) (153 aa)							●		0.907
● CD274	Programmed cell death 1 ligand 1 precursor (Programmed death ligand 1) (PD-L1) (PDCD1 ligand 1) (CD							●		0.906
● IRF1	Interferon regulatory factor 1 (IRF-1) (325 aa)							●		0.904
● CTLA4	Cytotoxic T-lymphocyte protein 4 precursor (Cytotoxic T-lymphocyte- associated antigen 4) (CTLA-4)							●		0.903
● IL1A	Interleukin-1 alpha precursor (IL-1 alpha) (Hematopoietin-1) (271 aa)							●		0.900
● STAT1	Signal transducer and activator of transcription 1-alpha/beta (Transcription factor ISGF-3 components							●		0.847
● CD80	T lymphocyte activation antigen CD80 precursor (Activation B7-1 antigen) (CTLA-4 counter-receptor E							●		0.830

Figure 1: The STRING network around the human gene indoleamine $2,3$-dioxygenase (IDO/INDO).

is applied in a more general setting.

3 Gene prioritization

The identification of disease genes is a key goal of biomedicine. Through cytogenetics, linkage mapping, association studies, or high-throughput techniques (e.g. expression microarrays, array-CGH), a set of candidate genes for that disease can often be defined. The next step, a careful analysis of these candidate genes in order to select a few interesting candidates or rank the candidates for further validation studies, is often not straightforward. The availability of multiple types of full-genome data (e.g. gene expression data, sequence information, Gene Ontology annotations, protein domain databases and the published literature) has made this into an interesting domain for automated approaches able to integrate large amounts of data. Multiple computational methods have been developed to tackle this "gene prioritization" problem. Most early approaches use data from one or two of these data sources (Freudenberg and Propping, 2002; Perez-Iratxeta *et al.*, 2002; Turner *et al.*, 2003; Tiffin *et al.*, 2005; Adie *et al.*, 2005; Lopez-Bigas and Ouzounis, 2004; Kent *et al.*, 2005), providing proof-of-principle that from each data source information useful for prioritizing candidate genes can be extracted. Recently, a few systems biology methods have been developed that tackle the gene prioritization problem by integrating data from multiple different data sources in a systematic way. Two of those are network-based methods (Franke *et al.*, 2006; Lage *et al.*, 2007), and are discussed below. A third method, Endeavour (discussed in study 1 in this work), uses each data source separately to prioritize candidate genes, and in a second step integrates these different prioritizations into one overall prioritization.

Franke *et al.* (2006) constructed a network based on data from Gene Ontology, microarray gene expression data (from 4 different datasets) and protein-protein interactions (predicted interactions (e.g. by yeast-two-hybrid), as well as confirmed interactions from existing databases such as Reactome and the Biomolecular Interaction Network Database (BIND)). They next hypothesize (and confirm) that genes involved in a similar inheritable disease are closer to each other in this network than randomly selected genes. Based on this principle and the constructed network, they develop a gene prioritization tool, Prioritizer. In a large-scale validation of this tool, they select 96 inheritable diseases from the Online Mendelian Inheritance in Man database (OMIM) for which between 3 and 10 disease genes have been identified. For each disease gene, they construct an artificial locus containing 100 candidate genes. Using Prioritizer to rank these candidate genes, they observed that for 54 % of the diseases, at least one candidate gene was ranked within the top five genes (representing a 2.8 fold increase over random selection).

Lage *et al.* (2007) construct a scored protein-protein interaction network by pooling interactions from many different protein-protein interaction databases, and by transferring data from 17 different eukaryotic organisms, using rigorous

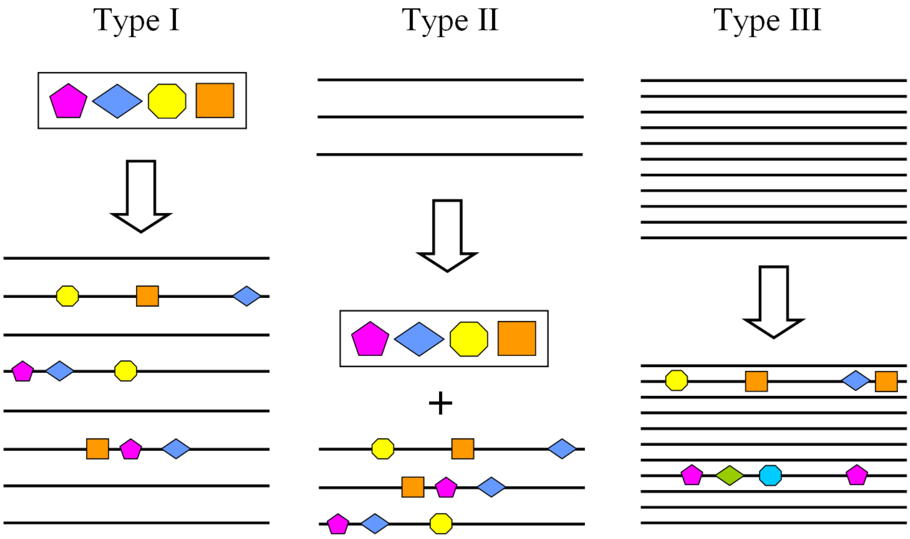


Figure 2: Three different types of *cis*-regulatory module detection algorithms.

quality control. In addition, they developed a text-mining based phenotypic similarity score that could be integrated with the protein-protein interaction network. The hypothesis underlying their gene prioritization approach is that mutations in different members of a protein complex lead to comparable phenotypes. Their approach extract a protein complex around each candidate gene from the constructed phenome-interactome network, and uses a Bayesian predictor to score the phenotype of interest in this phenotype annotated complex. In a large-scale five-fold cross-validation, they used 1404 artificial linkage intervals containing on average 109 genes, including one gene known to be involved in a particular disease. Their approach made 669 predictions, of which 298 were correct, indicating a precision of 45 % and a recall of 21 %.

4 Computational *cis*-regulatory module detection

In contrast to the available knowledge about genes, the annotation of regulatory regions in the human genome is far from complete. In the field of regulatory bioinformatics, many computational methods have been developed that aim to detect regulatory sequences. Here, we discuss these methods, focussing specifically on *in silico* approaches to detect *cis*-regulatory modules (CRMs).

The existing CRM detection approaches can be classified in three conceptually different classes, based on the specific aims of the methods (Figure 2):

1. Methods that screen sequences or complete genomes for CRMs based on a pre-defined model. These approaches aim to identify CRMs that contain binding sites for a specific combination of position weight matrices (PWMs). We call these "Type I CRM detection methods".

2. Methods that look for similar CRMs in a set of co-regulated or co-expressed genes. These approaches construct or select a combination of PWMs for which binding sites can be found in the putative regulatory regions of some or all of the given co-regulated genes. We call these "Type II CRM detection methods".

3. Methods that screen sequences or complete genomes for CRMs as homotypic or heterotypic clusters of binding sites for any combination of transcription factors. These methods do not require a predefined model or a predefined set of PWMs, but instead they look for clusters of binding sites for any combination of PWMs. We call these "Type III CRM detection methods".

4.1 Type I CRM detection methods

The properties of the different Type I CRM detection methods are outlined in tables 1 and 2. These methods commonly require a combination of PWMs and a genomic sequence as input, as well as a variable number of parameters. A number of different principles are used to incorporate homotypic and heterotypic clustering of transcription factor binding sites (PWM hits): e.g. counting of occurrences in a specific window, logistic regression analysis to predict the probability of a CRM hit and hidden Markov models (table 1).

Table 1: The different Type I CRM detection algorithms: working principles, inputs and parameters

Algorithm	Input	Parameters	Principle
LRA (Wasserman and Fickett, 1998; Krivan and Wasserman, 2001)	(i) training set of known CRMs (with identical length); (ii) negative training set; (iii) set of PWMs; (iv) genomic sequence	(i) score threshold	logistic regression analysis: model the (probability of) occurrence of CRMs as a function of the transcription factor binding site scores using multivariate logistic regression
Cister (Frith *et al.*, 2001)	(i) set of PWMs; (ii) a sequence	(i) binding site detection threshold; (ii) average distance between transcription factor binding sites; (iii) average number of transcription factor binding sites; (iv) average distance between CRMs; (v) window size for local background model	hidden Markov model
Ahab (Rajewsky *et al.*, 2002)	(i) putative regulatory sequences of the whole genome; (ii) set of PWMs	(i) window size; (ii) window step size	computes via maximum likelihood the probability that the window sequence is made up by sampling from the known PWMs or background (for each window); overlap is allowed and multiple weak instances are taken into account (since all possible segmentations in binding sites are considered)

Table 1: continuation

Algorithm	Input	Parameters	Principle
(e)CIS-ANALYST (Berman *et al.*, 2002, 2004)	(i) DNA database; (ii) set of binding sites of cooperatively working transcription factors	(i) window size; (ii) cut-off score per site (p-value); (iii) minimum number of sites	counts number of sites scoring above threshold; if this number is higher than the minimum number of sites asked, a CRM prediction is returned
Halfon *et al.* (2002)	-	-	looks for combinations of 2 Mad, 2 Tin, 2 Twi, 2 Pnt and 1 dTCF binding site, derived from *Drosophila* eve dorsal mesodermal enhancer, within a 500 base pair window
COMET (Frith *et al.*, 2002)	(i) set of sequences to search; (ii) set of PWMs	(i) gap penalty (expected average distance between motifs); (ii) window width for local nucleotide frequency background model	add PWM scores, use gap penalty for spacer sequences (in fact: hidden Markov model); statistics: log likelihood ratio of observing the data given a model of cis-element clusters versus a model of background DNA
SCORE (Rebeiz *et al.*, 2002)	(i) a (whole genome) sequence to scan; (ii) a consensus sequence	none	detect overrepresentation of binding sites of one particular transcription factor in differently sized windows
Cluster-Buster (Frith *et al.*, 2003)	(i) set of sequences to search; (ii) set of PWMs	(i) gap penalty (in fact: expected average distance between motifs); (ii) window width for (local nucleotide frequency) background model	similar to COMET
MCAST (Bailey and Noble, 2003)	(i) DNA database; (ii) set of PWMs	(i) p-value cutoff (for a single transcription factor binding site); (ii) maximum gap length; (iii) gap penalty	hidden Markov model
Module-Scanner (Aerts *et al.*, 2003b)	(i) a set of genomic sequences or conserved non-coding sequences; (ii) a set of PWMs	(i) max CRM size; (ii) overlap; (iii) penalization	looks for combination of PWMs gives the highest score (sum of binding energies)
MSCAN (Johansson *et al.*, 2003)	(i) a set of transcription factor binding profiles (PWMs) and (ii) a sequence	(i) significance threshold (for single PWM hits); (ii) window size; (iii) maximum number of motifs in a CRM	looks for significant PWM hit combinations; a p-value is assigned to each binding site, and these are later combined to a CRM score (two options: minimum p-value or product of p-values); this CRM score is then fitted to a statistical distribution to derive a p-value
Stubb (Sinha *et al.*, 2003, 2004)	(i) set of sequences (or a full genome) of one or more species; (ii) set of PWMs	(i) window length; (ii) window step size; (iii) background model	based on Ahab, with two modifications: (i) correlation between factors is modeled (e.g. factor A preferentially follows factor B) and (ii) comparative genomics: sequence conservation in multiple species is incorporated (by counting scores in aligned blocks in both species)

Table 1: continuation

Algorithm	Input	Parameters	Principle
PFR-Searcher (Grad *et al.*, 2004)	(i) set of similar PFRs (output of PFR-Sampler, tables 5 and 6); (ii) a (usually full genome) PFR database	none	a set of PFRs (phylogenetically footprinted non-coding regions) is collected by aligning two genomes (*Drosophila melanogaster* and *Drosophila pseudoobscura*) and next selecting regions of sufficient sequence conservation (60 % in 100 base pairs) and sufficient local sequence conservation (5th order hidden Markov model); Markov chain discrimination (i.e. log-likelihood of PFR generated by a "CRM" hidden Markov model compared to a "background" hidden Markov model)
ModuleFinder (Philippakis *et al.*, 2005)	(i) set of transcription factor binding profiles (PWMs); (ii) sequence	(i) window width; (ii) window step size; (iii) threshold score	for each window, the number of transcription factor binding sites is considered (along with their evolutionary conservation), and the likelihood of observing this is calculated
EEL (Hallikas *et al.*, 2006)	(i) 2 homologous DNA sequences; (ii) set of PWMs	4 parameters that weigh different aspects of the alignment score (can be calculated based on the full genome)	aligns sequences in the transcription factor binding site domain

Validation procedures used range from purely *in silico* to extensive *in vivo* studies, although the latter have been performed mostly in *Drosophila* (table 2). These validations indicate that the methods are useful in practice to detect CRMs in the complete *Drosophila* genome, although it should be kept in mind that only for a very limited number of processes sufficient data is available to construct a combination of PWMs. Although considerable progress has been made (table 2), most notably by the incorporation of comparative genomics in multiple methods, detecting CRMs by a genome-wide scan in the larger human genome remains a challenge.

Table 2: The different Type I CRM detection algorithms: validation, comments and availability.

Algorithm	Validation	Comments	Availability
LRA (Wasserman and Fickett, 1998; Krivan and Wasserman, 2001)	(i) skeletal muscle: 66 % sensitivity on training set, 60 % sensitivity on test set, one prediction every 32 kb; (ii) liver: 62 % sensitivity on training set, 50 % in complete jackknife analysis, one prediction per 35 kb	(i) first to show the principle; (ii) no direct comparative genomics, but they do use it as a second step screening strategy	available in any statistical package
Cister (Frith *et al.*, 2001)	(i) regulatory targets of LSF (human): sensitivity: 67 %, one prediction every 33 kb; (ii) skeletal muscle: performance comparable to LRA	output is difficult to interpret	available as an online tool
Ahab (Rajewsky *et al.*, 2002)	(i) body patterning of the *Drosophila* embryo (8 PWMs): 146 CRMs are found in the genome, including 17 of 27 known CRMs, estimated false positive rate is estimated to be about 50 %. (ii) Schroeder *et al.* (2004): Ahab predictions on the *Drosophila* segmentation network were experimentally validated by reporter constructs: 13 of 16 novel predictions drove expression in a correct pattern	(i) more or less the algorithm of choice (if comparative genomics not available); (ii) claimed to work also when PWMs are defined by Gibbs sampling; (iii) no comparative genomics	code available upon request

Table 2: continuation

Algorithm	Validation	Comments	Availability
(e)CIS-ANALYST (Berman *et al.*, 2002, 2004)	CIS-ANALYST: Bcd, Cad, Hb, Kr and Kni in *Drosophila* finds 9 known CRMs and 22 novel predictions (augmented to 28 by also looking for Bcd, Hb, Kr and Kni), 6 of those were positive; eCIS-ANALYST was constructed based on the results	(i) very simple, but very good performance (on the *Drosophila* segmentation network, although the choice of parameters was dictated by sensitivity/specificity for finding known CRMs)	available as an online tool
Halfon *et al.* (2002)	1 of 33 predicted enhancers was experimentally validated (and fully characterized) and shown to function in a similar way as the eve enhancer; some others were tested by reporter assays but shown to be non-functional	not a general approach, but applied to a specific example	a perl scripts is available online
COMET (Frith *et al.*, 2002)	(i) promoters regulated by LSF (in combination with Sp1, Ets-1 and the TATA box), muscle (Mef2, Myf, SRF, Tef, Sp1) (human); (ii) Comparison with Cister and LRA: performance is comparable	(i) E-value per CRM (and first to do this); (ii) construction of a model of background DNA is not straightforward; (iii) no comparative genomics	(i) downloadable executable; (ii) online tool
SCORE (Rebeiz *et al.*, 2002)	applied to Su(H) sites in *Drosophila*: one prediction was successfully validated in the lab	only homotypic cluster	none stated
Cluster-Buster (Frith *et al.*, 2003)	validated using Gene Ontology term enrichment in (i) muscle and (ii) LPS stimulation Bluthgen *et al.* (2005)	-	(i) downloadable executable and (ii) online tool
MCAST (Bailey and Noble, 2003)	(i) simulated data; (ii) real data in human and *Drosophila*; compared with COMET: similar but slightly better performance	(i) E-value per module; (ii) sensitive to the setting of its parameters; (iii) no comparative genomics	website is given, but no longer online
Module-Scanner (Aerts *et al.*, 2003b)	(i) *in silico*: human cell cycle PWM set predicted by ModuleSearcher validated by Gene Ontology; (ii) *in vitro* (human, study 1 in this work): CRMs in upregulated HL-60 cells	-	(i) available on request; (ii) integrated in Toucan (Aerts *et al.*, 2003a, 2005)
MSCAN (Johansson *et al.*, 2003)	(i) liver (66 % sensitivity, 1 putative CRM detected every 23 kb); (ii) skeletal muscle (66 % sensitivity; 1 putative CRM detected every 15 kb); comparison to LRA, Cister and COMET (slightly better performance)	-	available as an online tool
Stubb (Sinha *et al.*, 2003, 2004)	(i) synthetic sequences for multiple species; (ii) yeast toy example (sequences selected for having binding sites for 2 factors with correlations); (iii) gap gene system of *Drosophila*: all 16 known CRMs are recovered, together with only 2 novel predictions; (iv) the *Drosophila melanogaster* segmentation network, including *Drosophila pseudoobscura* sequences (*in silico*, using annotated anterior/posterior-segmentation genes)	(i) the multi-species version significantly outperformed the single-species version; (ii) more or less the standard algorithm when an extra species is available	you can request a copy online
PFR-Searcher (Grad *et al.*, 2004)	set of co-regulated genes centered around 10 *Drosophila* blastoderm genes that are known to share transcription factor binding sites, leave-one-out cross-validation	-	C-code available for download (after licence agreement)

Table 2: continuation

Algorithm	Validation	Comments	Availability
ModuleFinder (Philippakis *et al.*, 2005)	(i) skeletal muscle: sensitivity: 96.3 %, specificity: 94.4 % (although the threshold score is chosen as to maximize these values); (ii) compared to LRA, Cister, Comet and MSCAN: performance is better, but none of the other algorithms use comparative genomics; (iii) in Philippakis *et al.* (2006): *Drosophila* muscle founder cells	-	stated to be available for download, but website does not exist
EEL (Hallikas *et al.*, 2006)	*in silico* and *in vivo*: (i) detects all known *Drosophila* eve enhancers; (ii) expression in transgenic mice embryo's (successrate: $\geq$ 30 %)	novel idea and high performance	tool available for download

We would like to highlight one recent novel approach (Enhancer Element Locator (EEL), Hallikas *et al.* (2006)) that uses alignment of predicted transcription factor binding sites in two species to make CRM predictions. In this method, first the sequences of both species are used to predict binding sites using the given PWMs. In the second step, the sequences themselves are not used anymore, and a Smith-Waterman alignment (Smith and Waterman, 1981) of the predicted binding sites is performed. The validation of this methods predictions in the human-mouse system by expression constructs in transgenic mice embryo's showed a success-rate of over 30 %, indicating that this method may achieve sufficient sensitivity and specificity levels to annotate CRMs in the human genome.

4.2 Type II CRM detection methods

In general, these methods take as input a set of co-regulated or co-expressed genes (or their putative regulatory sequences), and they predict (i) the transcription factors (or PWMs) working cooperatively in regulating these genes and (ii) the *cis*-regulatory modules regulating these genes, as combinations of binding sites for these PWMs. We subclassify these Type II methods in two parts: (a) methods that select PWMs from a PWM library and (b) methods that construct their own PWMs.

Type IIa methods

The different Type IIa CRM detection methods are outlined in tables 3 and 4. The number of available algorithms is relatively limited: only two early approaches fall strictly into this category: ModuleSearcher (Aerts *et al.*, 2003b, 2004) and CREME (Sharan *et al.*, 2003, 2004), as well as our ModuleMiner algorithm, discussed in this work (study 2). The MARSMOTIF algorithm has a slightly different focus: it models microarray gene expression as a function of motif content (similar to REDUCE, Bussemaker *et al.* (2001)).

Table 3: The different Type IIa CRM detection algorithms: working principles, inputs and parameters

Algorithm	Input	Parameters	Principle
ModuleSearcher (Aerts *et al.*, 2003b, 2004)	(i) database of PWMs; (ii) sequences of co-regulated genes	(i) maximum CRM length; (ii) maximum number of PWMs; (iii) penalization	identifies PWM combinations with maximum sum of scores in the given set of genes; comparative genomics: looks in conserved non-coding regions
CREME (Sharan *et al.*, 2003, 2004)	(i) database of PWMs; (ii) promoter sequences of a (large) set of (loosely) co-regulated human genes (and orthologous sequences in the mouse and rat genome)	(i) maximum CRM length; (ii) maximum number of PWMs; (iii) threshold for individual motif hits	multistep algorithm: (i) select only single motifs that are overrepresented (compared to a background set of sequences); (ii) filter similar PWMs (by overlap in predicted binding sites); (iii) hashing algorithm to go through all combinations of PWMs and calculate their combined significance (compared to the expected frequency based on the occurrences of their component motifs); (iv) filter similar CRMs
MARS-MOTIF (Das *et al.*, 2004)	(i) microarray expression data; (ii) set of candidate motifs	number of maximum interactions allowed (corresponds to the size of CRMs)	model microarray gene expression as a function of motif content (PWM score), including combinations of motifs using multivariate adaptive regression splines
ModuleMiner	(i) database of PWMs; (ii) set of co-regulated genes	none	similar to ModuleSearcher, but identifies the CRMs that are the most discriminative for the given set of co-regulated genes, compared to the rest of the genome; see study 2 in this work

Except for the *in vitro* validation of ModuleSearcher (in combination with ModuleScanner, study 1 in this work) and the extensive *in silico* validation of ModuleMiner (study 2 in this work), these methods have only been validated to a limited extent (table 4).

Table 4: The different Type IIa CRM detection algorithms: validation, comments and availability.

Algorithm	Validation	Comments	Availability
ModuleSearcher (Aerts *et al.*, 2003b, 2004)	(i) *in silico*: human cell cycle, validated using Gene Ontology; (ii) *in vitro*: differential regulation in HL-60 differentiation; see study 1 in this work	-	(i) available on request; (ii) integrated in Toucan (Aerts *et al.*, 2003a, 2005)
CREME (Sharan *et al.*, 2003, 2004)	(i) cell cycle data, validated using correlation in microarray data; (ii) stress response data, validated using Gene Ontology	(i) only 1.5 kb 5' of TSS is tested; (ii) starts with about 500 loosely co-regulated genes	available as an online tool
MARS-MOTIF (Das *et al.*, 2004)	(i) simulated data; (ii) yeast cell-cycle, compared to REDUCE (Bussemaker *et al.*, 2001); (iii) application to tissue-specific expression modeling (Smith *et al.*, 2006): for 56 tissues (GNF Symatlas), 500 positive (tissue-specific) and negative genes were selected and MARSMOTIF was used to build a classifier; significant performance was observed for 45 tissues, although the errors were still large	(i) mostly finds binary interactions; (ii) more focussed on finding transcription factors working cooperatively than on the detection of CRMs	(i) software available after licence agreement

Table 4: continuation

Algorithm	Validation	Comments	Availability
ModuleMiner	(i) smooth muscle genes; (ii) compared to other algorithms; (iii) application to microarray clusters (tissue-specific expression) and developmental gene sets; see study 2 in this work	-	(i) online tool; (ii) stand-alone version available upon request

Type IIb methods

The different Type IIb CRM detection methods are summarized in tables 5 and 6. These methods can be viewed as extensions of approaches detecting single motifs (see Tompa *et al.* (2005) and references therein for an overview of motif detection methods), aiming to overcome the limitations of these methods by incorporating cooperativity. Often these methods are based on multiple component models, where the singular motifs and their combination are optimized simultaneously or iteratively (table 5). Two of these methods incorporate comparative genomics: PRF-sampler (Grad *et al.*, 2004) (*Drosophila melanogaster* and *Drosophila pseudoobscura*) and the Gibbs Module Sampler (Thompson *et al.*, 2004). This last algorithm extends the Gibbs sampling approach for single motif detection to combinations of motifs and to motifs conserved in two species.

Table 5: The different Type IIb CRM detection algorithms: working principles, inputs and parameters

Algorithm	Input	Parameters	Principle
CO-Bind (GuhaThakurta and Stormo, 2001)	(i) set of co-regulated genes and their sequences; (ii) set of background sequences	(i) maximum distance between binding sites; (ii) algorithm parameters: step-size and decay factor	Gibbs sampling strategy and neural network (perceptron) to find PWMs for two sets of similar binding sites close together
PFR-Sampler (Grad *et al.*, 2004)	set of co-regulated genes	(i) number of initial hits (default equal to the number of co-regulated genes); (ii) algorithm parameters	a set of PFRs (phylogenetically footprinted non-coding regions) is collected by aligning two genomes (*Drosophila melanogaster* and *Drosophila pseudoobscura*) and next selecting regions of sufficient sequence conservation (60 % in 100 base pairs) and sufficient local sequence conservation (5th order hidden Markov model); algorithm: simulated annealing, using sum of PFR-Searcher scores
Kreiman (2004)	(i) set of co-regulated genes (and their putative regulatory sequences); (ii) putative regulatory sequences of all genes in the genome	(i) maximum distance between adjacent motif occurrences; (ii) maximum overlap between binding sites; (iii) minimum number of genes with a CRM; (iv) p-value cutoff	exhaustively tries all combinations of up to 4 PWMs; ensures that the co-occurring motifs are sparsely distributed throughout the genome

Table 5: continuation

Algorithm	Input	Parameters	Principle
CisModule Zhou and Wong (2004)	sequences of a set of co-regulated genes	(i) number of PWMs and (ii) module length	(generative) hierarchical mixture model, consisting of 2 levels: (i) CRM vs. background and (ii) (within CRM) transcription factor binding sites vs. background; iterative algorithm consisting of 2 steps (similar to Gibbs sampling): (i) given CRM and transcription factor binding site positions, estimate parameters, and (ii) given parameters, estimate (sample) CRM and transcription factor binding site positions
Gibbs Module Sampler Thompson *et al.* (2004)	set of orthologous sequences of a set of co-regulated genes	(i) maximum number of motifs; (ii) maximum distance between motifs (hard-coded to 100 base pairs)	Gibbs motif sampler extended to (i) find conserved motifs (sampling over aligned pairs of sites) and (ii) find CRMs as combinations of motifs (including neighbor interactions)
EMCMODULE (Gupta and Liu, 2005)	(i) putative regulatory sequences of a set of co-regulated genes; (ii) starting set of PWMs (optional)	(i) prior probability distribution of binding site occurrence; (ii) distribution of distance between motif sites (truncated geometric distribution)	statistical model to describe CRM structure (hidden Markov model) and an evolutionary Monte Carlo motif screening strategy (similar to a genetic algorithm)
Segal and Sharan (2005)	sequence data of a set of co-regulated genes (and a negative set)	(i) window size; (ii) window step size	three-component model: (i) motif model (PWMs), (ii) module model (CRMs as weighed PWM combinations; models the probability that a sequence window contains a CRM given the binding site occurrences of the motifs in it) and (iii) regulation model (probabilities that the given positive and negative genes are regulated by the CRM); all three models are logistic functions, optimized by an expectation-maximization algorithm

The performance of these methods is relatively limited (table 6), in part because motif detection is a complex and unresolved problem (Tompa *et al.*, 2005). In this work (study 2), we compare the performance of our novel Type IIa algorithm ModuleMiner to several other Type IIa and Type IIb CRM detection algorithms, confirming this limited performance of the Type IIb algorithms. Therefore, and because we believe the emergence of the protein-binding microarray technology (Mukherjee *et al.*, 2004) will make high quality PWMs available for most transcription factors in the near future, we believe that Type IIa algorithms will prove to be more useful for the annotation of regulatory regions in the human genome.

Table 6: The different Type IIb CRM detection algorithms: validation, comments and availability.

Algorithm	Validation	Comments	Availability
CO-Bind (GuhaThakurta and Stormo, 2001)	(i) synthetic data and (ii) 4 yeast sets (promoters selected for sharing known binding sites): in three cases, both patterns could be identified	only combinations of up to two factors	downloadable executable

Table 6: continuation

Algorithm	Validation	Comments	Availability
PFR-Sampler (Grad *et al.*, 2004)	set of co-regulated genes centred around 10 *Drosophila* blastoderm genes that are known to share transcription factor binding sites, leave-one-out cross-validation	-	C-code is downloadable (after licence agreement)
Kreiman (2004)	(i) random sets of genes (negative control); (ii) yeast cell cycle; (iii) *Drosophila* pattern development: finds 3 motifs (overlap with true motifs is not discussed); predicted CRMs (regions) correspond to known CRMS; (iv) human muscle regulatory regions: finds the three correct PWMs; predicted CRMs (regions) correspond quite well with known regulatory regions	-	(i) source code is available upon request
CisModule Zhou and Wong (2004)	(i) artificial sequences, (ii) 3 cases of homotypic clustering in *Drosophila*: sensitivity (based on transcription factor binding sites discovered) was 56 %, but number of sites was not that small; and in all three cases, the correct PWM was found; (iii) (human) muscle-specific regulatory regions: 4 PWMs were correctly identified, sensitivity (based on transcription factor binding site discovery) was 88 %; an analysis was added checking sensitivity to added negative sequences: in 29 positive and 40 negative sequences, 54 % of detected CRMs were in the positive sequences	-	available online
Gibbs Module Sampler Thompson *et al.* (2004)	(i) skeletal muscle: 4 of 5 motifs were correctly identified; 17 of 20 CRMs were correctly located in 3 kb sequences (50 % overlap); transcription factor binding sites: sensitivity: 69 %, false positive rate: about 35 %; (ii) smaller liver case-study: only HNF1 was detected; (iii) comparison to COMET: roughly similar performance, showing that the addition of comparative genomics can offset the extra difficulty of modeling the PWMs	-	none stated
EMCMODULE (Gupta and Liu, 2005)	(i) *bacillus subtilis*; (ii) *Drosophila* early development: PWMs were correctly identified for 4 of the 5 factors; compared to Gibbs module sampler and CisModule: recovered none of the known motifs; (iii) human skeletal muscle: recovered 3 of the 5 motifs; when using JASPAR (Sandelin *et al.*, 2004) as starting motifs: recovered 4 of the 5 motifs	can work with a starting set of PWMs (although these need to be selected carefully)	available online
Segal and Sharan (2005)	(i) simulated data; (ii) yeast data (ChIP-chip, genes sets selected for sharing binding sites for 2 factors): in 11 of 25 sets, CRMs were identified; 7 of 11 PWMs were correct; (iii) human: CRM predictions done on all 381 Gene Ontology categories: 83 CRMs were identified in 71 Gene Ontology categories; of 203 motifs, 54 correspond to known motifs	assigns a weight to PWMs	none stated

4.3 Type III CRM detection methods

The properties of the different Type III CRM detection methods are summarized in tables 7 and 8. In general, these methods require a database of PWMs and a genomic sequence as input. We can subdivide these methods into early approaches (Argos and TraFac) that delivered proof-of-principle but are not aimed at detecting CRMs genome wide, and late approaches (PreMod and Enhancer Element Locator) that aim to make genome-wide predictions. The Regulatory Potential method does not strictly aim to detect CRMs, but calculates the regulatory potential as a function of genomic position, based on two- or three-way alignments.

Table 7: The different Type III CRM detection algorithms: working principles, inputs and parameters

Algorithm	Input	Parameters	Principle
Argos (Rajewsky *et al.*, 2002)	(i) genomic sequence; (ii) database of PWMs	(i) window size; (ii) window step size	looks for overrepresented motifs is a sequence, and combines 5 different non-overlapping motifs into one score
TraFaC (Jegga *et al.*, 2002)	(i) two orthologous sequences; (ii) PWM library	none	looks for clusters of conserved binding sites in one sequence
Regulatory Potential (Elnitski *et al.*, 2003; Kolbe *et al.*, 2004)	human-mouse(-rat) alignments	none, although many hard coded choices are made	collapses alignment alphabet to fewer symbols and uses a higher order hidden Markov model (trained on positive vs. negative sequences)
PreMod (Blanchette *et al.*, 2006)	(i) human-mouse-rat whole genome alignments; (ii) database of PWMs	(i) maximum length of CRMs; (ii) score/p-value thresholds for transcription factor binding sites and CRM detection	search for statistically significant clusters of (phylogenetically conserved) binding sites for 1-5 transcription factors (PWMs); homotypic clustering is extensively used
EEL (Hallikas *et al.*, 2006)	(i) 2 homologous DNA sequences; (ii) database of PWMs	4 parameters that weigh different aspects of the alignment score (can be calculated based on the full genome)	aligns sequences in the transcription factor binding site domain

These methods are more general than the Type I or Type II CRM detection methods: the latter aim to detect CRMs with a specific function, while Type III methods focus on the detection of CRMs as homotypic and/or heterotypic clusters of binding sites for any combination of PWMs. Hence, these Type III methods require no prior knowledge. However, as a consequence, the performance is lower and no inference can be made about the function of the predicted CRMs.

Table 8: The different Type III CRM detection algorithms: validation, comments and availability.

Algorithm	Validation	Comments	Availability
Argos (Rajewsky *et al.*, 2002)	predictions over the full genome: false negative rate estimated to be 50 %; one prediction per 5 kb	first of its kind	none stated
TraFaC (Jegga *et al.*, 2002)	very specific case-studies	leaves the detection of the CRMs to the interpretation of the user	available as an online tool
Regulatory Potential (Elnitski *et al.*, 2003; Kolbe *et al.*, 2004)	In King *et al.* (2005): applied to a class of erythroid specific genes, including β-globin: sensitivity: 60 %, specificity: 60 %	plots regulatory potential as a function of sequence position, hence does not really detect CRMs	available as a UCSC genome browser track

Table 8: continuation

Algorithm	Validation	Comments	Availability
PreMod (Blanchette *et al.*, 2006)	(i) overlap with known CRMs; (ii) ChIP-chip validation for ER and E2F4 binding sites: low performance	-	full genome predictions are available
EEL (Hallikas *et al.*, 2006)	validated as a Type I algorithm	designed as a Type I method, with Type III potential	tool and predictions available for download

5 Conclusions

Systems biology is a promising emerging field expected to complement the classical reductionist approach of the biologist. Microarrays expression profiling, together with its analysis methods, are widely-used techniques that fit well in this systems biology philosophy. Network-based methods can provide novel insights into biological processes and diseases. Bioinformatics and systems biology methods have been developed to assist in the hunt for disease genes. Although many of these systems biology methods have delivered solid proof-of-principle, this field is still embryonal and at this moment the reductionist approach maintains its prominent role in biological research.

Computational methods to detect *cis*-regulatory modules can be classified into three main classes. Type I methods aim to detect CRMs based on known examples. Type II methods aim to find similar CRMs in co-regulated genes. Type III methods aim to detect CRMs as clusters of binding sites for any combination of transcription factors. Type I methods have reached an advanced stage and predictions can be made with reasonable sensitivity and specificity, provided sufficient prior knowledge about the system under study is available. Type II and Type III methods have shown proof-of-principle, but their performance is still limited. Given the limited high-throughput possibilities to experimentally annotate regulatory regions in the human genome, computation CRM detection remains an important area of research.

Rationale and aims

In this work, we aim to (i) develop novel bioinformatics and systems biology methods for *cis*-regulatory module detection and gene prioritization, and (ii) apply bioinformatics, systems biology and statistics to tackle two biomedical problems.

1 Development of novel systems biology methods

Study 1: Gene prioritization through genomic data fusion

The identification of key disease and pathway genes is an important endeavour in current biomedical research. These genes are often selected by a two-step process: in the first step, a set of candidate genes is defined by classical genetics approaches or high-throughput systems biology techniques, while in the second step, the key disease or pathway gene is selected from these candidate genes. Computationally, this second step can be tackled by prioritizing (or ranking) a set of candidate genes, using genome-wide data sources. As explained in section 3 in the introduction to this work, multiple methods have been developed for computational gene prioritization. However, although a plethora of quasi genome-wide databases are available, each of these methods uses only one or two of these data sources.

Here, we aimed to develop a novel gene prioritization method able to integrate data from multiple heterogeneous data sources. In addition, we aimed to incorporate expert knowledge into the system and to make it highly modular. Finally, we aimed to make the system publicly available and user-friendly.

We reasoned that the integration of data from multiple data sources might significantly increase the performance of computational gene prioritization. In addition, the ability to handle heterogeneous data (e.g. literature, microarray gene expression data, structured annotation such as Gene Ontology) drastically increases the data sources available. The modularity of the system would further add to this, making it easily extendible when novel data becomes available. The incorporation of expert knowledge by the user decreases the sensitivity to noise in high-throughput databases and allow approaching the problem from specific (and possibly multiple) angles. Finally, a publicly available user-friendly tool minimizes any thresholds for biomedical researchers in using the method.

We incorporate expert knowledge specifically under the form of a set of training genes. These training genes are typically genes known to be involved in the process or disease under study. The user can influence this set by maintaining a specific confidence threshold in the selection of training genes, or by focussing on specific subphenotypes or processes. In addition, for diseases where no known genes are available, genes involved in related diseases or phenotypes can be selected. Our gene prioritization framework, Endeavour, is based on similarities between the candidate genes and the training genes. In this regard, our method can be considered a thorough extension of the BLAST algorithm, where we not only consider sequence similarities, but also e.g. expression similarities, interaction similarities and similarities in literature about the genes.

We validated Endeavour extensively *in silico*, *in vitro* and *in vivo*. The *in silico* validation consisted of a large-scale leave-cross-validation as well as specific case-studies prioritizing recently identified monogenic and complex disease genes. For the *in vitro* validation, we aimed to predict genes differentially regulated in macrophage differentiation, combining a *cis*-regulatory module detection method with Endeavour. Finally, an extensive *in vivo* validation of a novel predicted DiGeorge syndrome gene was also an important aim of this collaborative study, although this was not part of this doctoral thesis work.

Study 2: ModuleMiner: improved computational detection of *cis*-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues?

Since the sequencing of the human genome, the annotation of functional elements has taken a decisive leap, most notably for protein-coding genes. However, transcriptional regulatory regions are lagging far behind in these large-scale annotation efforts. Therefore, computational detection of regulatory regions is an important area of research. The general aim of this work is to improve computational methods to detect *cis*-regulatory modules.

Our classification of *cis*-regulatory module detection methods (section 4 in the introduction) showed that: (i) methods aiming to detect CRMs based on a specific collection of position weight matrices for transcription factors working cooperatively (Type I CRM detection methods) perform quite well, yet for few systems these data are available; (ii) methods that look for similar CRMs in co-regulated genes (Type IIa and IIb CRM detection methods) and non-parametric methods to scan the complete genome for regulatory regions (Type III CRM detection methods) have shown proof-of-principle, but may not yet be usable for reliable large-scale regulatory region annotation; (iii) Type III and Type IIb CRM detection methods aim to solve a problem that is considerably more complex than that of Type IIa CRM detection methods. For these reasons, we believe Type IIa CRM detection algorithms are the most interesting area of research. Hence, the specific aim of this work is to develop a novel Type IIa CRM detection method with an increased performance, and make this publicly available.

All existing CRM detection algorithms require *a priori* parameters about the CRMs, such as the length (in base pairs) of the CRMs and the number of PWMs involved (Tables 3 and 5 in the General introduction). We aimed to develop a CRM detection algorithm that does not require these user parameters. In addition, we aimed to maximize specificity of the algorithm for the given set of co-regulated genes.

We approach this maximization of specificity by basing our algorithm, ModuleMiner, on a whole-genome optimization, effectively optimizing the "signal" in the given co-regulated genes compared to all other genes in the genome. This whole-genome optimization strategy also allows optimization over various parameters, allowing us to eliminate all *a priori* parameters about the CRMs.

We assess the performance of ModuleMiner by direct comparison with other state-of-the-art Type IIa and Type IIb CRM detection algorithms on benchmark data. In addition, the sensitivity of ModuleMiner to false positive genes is assessed, as a low noise sensitivity is a prerequisite for application of the algorithm to gene sets obtained from clustering of microarray data.

Finally, we use ModuleMiner on a large scale to make predictions regarding CRMs directing expression in adult tissues and CRMs involved in embryonic development. These two groups of CRM predictions are subsequently compared, most notably regarding location preference.

2 Applications to cancer stratification and understanding

In the second part of this work, we apply bioinformatics, systems biology and statistical methods aiming to obtain a better understanding of two types of cancer. Although the entities investigated and the methods applied are unrelated, the general aims are somewhat similar. Indeed, in both cases, we aim to link properties of the cancer entities to clinicopathological parameters and to gain a better understanding of the precise similarities and differences of related cancer entities.

Study 3: T cell/histiocyte rich large B cell lymphoma shows a tolerogenic host immune response: the lymphoma microenvironment as a target for therapy

Many tumours contain additional cells apart from the clonal tumour cells that are not outgrown by the malignant cells. Specifically in lymphoma, the importance of this tumour microenviroment has been underlined by two microarray expression profiling studies showing that the microenvironment plays a role in the profile of the lymphoma and in predicting the prognosis (Dave *et al.*, 2004; Monti *et al.*, 2005). Aiming to elucidate the mechanisms by which this microenvironment can contribute to the prognosis, we used systems biology techniques to study two lymphoma entities with many similarities, but a prominently different microenvironment and a clear difference in prognosis.

T cell/histiocyte rich B cell lymphoma (THRLBCL) is a rare variant of diffuse large B cell lymphoma sharing many characteristics with a specific subtype of Hodgkin's disease: nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL). In both cases, the malignant cells themselves constitute only a minority of the tumour cell mass, while the majority of tumour cell mass is taken up by the microenvironment. In addition, the malignant cells of both lymphomas express pan B cell markers, show strong similarities to germinal center B cells and share a number of chromosomal aberrations. In sharp contrast, the prognosis of both lymphoma entities is clearly different: NLPHL is a very indolent disorder, while THRLBCL is very aggressive.

We aim to gain more insight into the microenvironment of both lymphomas and its link with clinicopathological parameters, and specifically into the possible involvement of the microenvironment in explaining the bad prognosis of THRLBCL. In addition, we wish to study to what extent the differences between the NLPHL and THRLBCL microenvironment can be captured by a very limited number of genes.

By microarray expression profiling, we study to what extent and in what way the differences in cellular composition of the microenvironment in THRLBCL and NLPHL translate to differences in expression profiles. Secondly, by careful study of the NLPHL and THRLBCL (microenvironment) expression profiles, we hope to attain mechanistic hypotheses regarding the (putative) involvement of the microenvironment in the bad prognosis of THRLBCL. In addition, we correlate our expression profiles with other microarray experiments related to the microenvironment in lymphoma, most particularly Monti *et al.* (2005) and Dave *et al.* (2004). Finally, we investigate to what extent three genes (selected from our microarray expression profiling experiment) can suffice to classify additional in-house and external cases using real-time quantitative RT-PCR.

Study 4: Polysomy 17 in breast cancer: clinicopathological significance and impact on HER2 testing

Amplification of the gene HER2 (also called ERBB2 or neu), located on chromosome 17 and encoding the human epidermal growth factor receptor 2 protein, defines an important distinct subgroup of breast cancers associated with a bad prognosis. However, focused HER2-targeted therapies such as trastuzumab (Herceptin®, Genentech), a monoclonal antibody targeting the extracellular domain of the HER2 protein, have been developed that improve the prognosis of these HER2 positive breast cancers, either alone or in combination therapy (Vogel *et al.*, 2002; Slamon *et al.*, 2001). As a consequence, trastuzumab is now a widely used agent in HER2 positive breast cancer.

For the determination of the HER2 status in breast cancer, a large variety of techniques are available, measuring HER2 amplification status at the protein, mRNA or DNA level. Immunohistochemistry (IHC) to measure HER2 protein levels is the most widely used technique. In this technique, HER2 protein expression is scored on a 0 to 3+ scale. A score 0 or 1+ is interpreted as HER2 negative and a score 3+ as HER2 positive. A score 2+ should be regarded as inconclusive or equivocal, indicating further molecular analysis is required to determine the HER2 status in such cases. HER2 amplification status (positive, negative or equivocal) can also be determined by fluorescence *in situ* hybridisation (FISH), a highly sensitive technique to measure DNA copies of the HER2 gene. However, because of its relatively high cost and requirement for specialized equipment, FISH analysis is not preferred for primary HER2 status screening. It is at the moment unclear whether or not patients with an equivocal HER2 status (either by IHC or by FISH) would benefit from trastuzumab therapy.

Chromosome aneuploidy occurs often in cancer and reflects genetic instability. In relation to HER2 amplification in breast cancer, chromosome 17 aneuploidy plays an important yet unknown role. Indeed, tumour with an increased chromosome 17 copy number will also have an increased number of HER2 gene copies which might result in increased HER2 protein expression levels. In addition, the impact of polysomy 17 on different HER2 testing methods is currently unclear.

We aimed to elucidate the effect of polysomy 17 on HER2 testing methods, and to investigate to what extent polysomy 17 breast cancers share biological characteristics with breast cancers showing a true HER2 amplification.

For a series of 226 primary invasive breast carcinomas, we correlate results of HER2 IHC with that of two different FISH assays: (i) one-probe FISH to measure the absolute HER2 copy number and (ii) two-probe FISH to measure the relative HER2 copy number (compared to the number of copies of chromosome 17). In addition, we investigate the impact of polysomy 17 to the results of these three HER2 status assays. We stratify the patient population by HER2 status and polysomy 17 (HER2 amplified, polysomy 17 and HER2 normal), and we measure HER2 mRNA levels by real-time quantitative RT-PCR in the different groups, aiming to gain quantitative insight into the effect of HER2 gene amplification and polysomy 17 on the transcriptional levels of the HER2 gene. Finally, a set of clinicopathological parameters, including survival, is correlated with this population stratification, with the aim to elucidate the clinicopathological significance of polysomy 17 in relation to that of HER2 gene amplification.

Study 1

Gene prioritization through genomic data fusion

Contribution of the doctorandus: co-development of the data fusion method, application to pathway data (prioritization of genes predicted to be differentially regulated in HL-60 differentiation) and experimental validation of the results.

_computational BIOLOGY

ANALYSIS

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

Gene prioritization through genomic data fusion

Stein Aerts[1,4,5], Diether Lambrechts[2,5], Sunit Maity[2,5], Peter Van Loo[3–5], Bert Coessens[4,5], Frederik De Smet[2], Leon-Charles Tranchevent[4], Bart De Moor[4], Peter Marynen[3], Bassem Hassan[1], Peter Carmeliet[2] & Yves Moreau[4]

The identification of genes involved in health and disease remains a challenge. We describe a bioinformatics approach, together with a freely accessible, interactive and flexible software termed Endeavour, to prioritize candidate genes underlying biological processes or diseases, based on their similarity to known genes involved in these phenomena. Unlike previous approaches, ours generates distinct prioritizations for multiple heterogeneous data sources, which are then integrated, or fused, into a global ranking using order statistics. In addition, it offers the flexibility of including additional data sources. Validation of our approach revealed it was able to efficiently prioritize 627 genes in disease data sets and 76 genes in biological pathway sets, identify candidates of 16 mono- or polygenic diseases, and discover regulatory genes of myeloid differentiation. Furthermore, the approach identified a novel gene involved in craniofacial development from a 2-Mb chromosomal region, deleted in some patients with DiGeorge-like birth defects. The approach described here offers an alternative integrative method for gene discovery.

With the advent of 'omics, identifying key candidates among the thousands of genes in a genome that play a role in a disease phenotype or a complex biological process has paradoxically become one of the main hurdles in the field. Indeed, contrary to some early concerns in the community that a lack of sufficient global data would still be a limiting factor[1], it is precisely the opposite, a bounty of information that now poses a challenge to scientists. This has translated into a need for sophisticated tools to mine, integrate and prioritize massive amounts of information[2,3].

Several gene prioritization methods have been developed[4–10]. Most of them determine, either directly or indirectly, the similarity between candidate genes and genes known to play a role in defined biological processes or diseases. These methods offer several advantages but also pose a number of challenges. Indeed, even though multiple data sources are available, such as Gene Ontology (GO) annotations[4–6,10], protein domain databases[6,10], the published literature[5,7], gene expression data[5,7,10] and sequence information[8–10], most of the available programs access only one or two of these databases, which each have their limitations. For instance, functional data sources (GO and literature) are incompletely annotated and biased toward better-studied genes[8], whereas sequence databases have thus far been used only to produce general disease probabilities[8,9]. Some of the existing tools access more than two databases, but do not provide an overall ranking that integrates the separate searches[5,10]. Several tools rank disease genes but only one of them prioritizes genes involved in biological pathways[10], and none offers the combination of both. Thus far, only two prioritization tools[5,10] are publicly available. Thus, there is still a need for improvement of gene prioritization.

Here, we report the development and characterization of a new gene prioritization method, and offer its freely accessible, interactive and flexible software[1]. Compared to existing methods, ours provides additional opportunities for candidate gene prioritization: it accesses substantially more data sources and offers the flexibility to include new databases; it provides the user control over the selection of training genes and thereby takes advantage of the expertise of the user; it prioritizes both known and unknown genes, ranks genes involved in human diseases and biological processes, and it uses rigorous statistical methods to fuse all the individual rankings into an overall rank and probability.

RESULTS

Principles of prioritization used by Endeavour

Genes involved in the same disease or pathway often share annotations and other characteristics in multiple databases. Indeed, genes involved in the same disease share up to 80% of their annotations in the GO and InterPro databases[6], whereas genes involved in a similar biological pathway often share a high degree of sequence similarity with other pathway members[11]. It is therefore reasonable to assume that this similarity among genes is not restricted to their annotation or sequence alone, but is also true for their regulation and expression. We reasoned that a bioinformatics framework capable of comparing and integrating all available gene characteristics might be a powerful tool to rank unknown candidate 'test' genes according to their similarity with known 'training' genes, and based on this notion, we developed Endeavour. Prioritization of genes using this algorithm involves three steps (**Fig. 1**). To validate its performance, we used several complementary strategies discussed below.

---

[1]Laboratory of Neurogenetics, Department of Human Genetics, [2]The Center for Transgene Technology and Gene Therapy, [3]Human Genome Laboratory, Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology (VIB), University of Leuven, Herestraat 49, bus 602, 3000 Leuven, Belgium. [4]Bioinformatics Group, Department of Electrical Engineering (ESAT-SCD), University of Leuven, Belgium. [5]These authors contributed equally to this work. Correspondence should be addressed to S.A. (stein.aerts@med.kuleuven.be).

Published online 5 May 2006; doi:10.1038/nbt1203

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

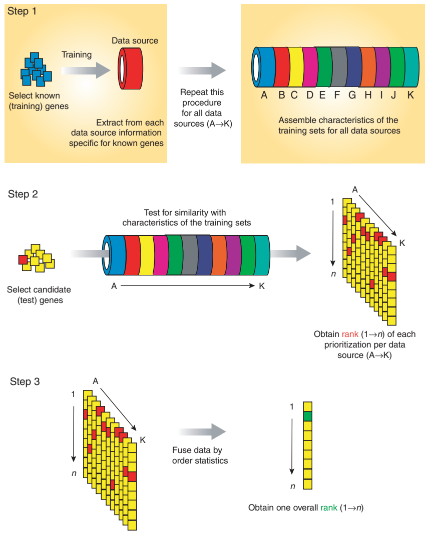

Step 1
Data source
Training
Select known (training) genes
Extract from each data source information specific for known genes
Repeat this procedure for all data sources (A→K)
A B C D E F G H I J K
Assemble characteristics of the training sets for all data sources
Step 2
Test for similarity with characteristics of the training sets
Select candidate (test) genes
A
K
1
n
Obtain rank (1→n) of each prioritization per data source (A→K)
Step 3
Fuse data by order statistics
Obtain one overall rank (1→n)

Figure 1 Concept of prioritization by Endeavour. Candidate test genes are ranked with Endeavour based on their similarity with a set of known training genes in a three-step analysis. In the first step (upper panel), information about a disease or pathway is gathered from a set of known training genes by consulting various data sources. Training genes can be loaded automatically (based on a Gene Ontology term, a KEGG pathway ID or an OMIM disease name) or manually. The latter allows the incorporation of expert knowledge. The following data sources are used: A, literature (abstracts in EntrezGene); B, functional annotation (Gene Ontology); C, microarray expression (Atlas gene expression); D, EST expression (EST data from Ensembl); E, protein domains (InterPro); F, protein-protein interactions (Biomolecular Interaction Network Database or BIND); G, pathway membership (Kyoto Encyclopedia of Genes and Genomes or KEGG); H, *cis*-regulatory modules (TOUCAN); I, transcriptional motifs (TRANSFAC); J, sequence similarity (BLAST); K, additional data sources, which can be added (e.g., disease probabilities). In the second step (middle panel), a set of test genes is loaded (again, either manually or automatically by querying for a chromosomal region or for markers). These test genes are then ranked based on their similarity with the training properties obtained in the first step, which results in one prioritized list for each data source. Vector-based data are scored by the Pearson correlation between a test profile and the training average, whereas attribute-based data are scored by a Fisher's omnibus analysis on statistically overrepresented training attributes. Finally, in the third step (lower panel), Endeavour fuses each of these rankings from the separate data sources into a single ranking and provides an overall prioritization for each test gene. As such, Endeavour prioritizes genes through genomic data fusion—a term, borrowed from engineering to reflect the merging of distinct heterogeneous data sources, even when they differ in their conceptual, contextual and typographical representations.

Validation of Endeavour when accessing individual data sources

For each individual data source, we assessed whether our approach is capable of prioritizing genes known to be involved in specific diseases or receptor signaling pathways. To this end, we performed a large-scale leave-one-out cross-validation. In each validation run, one gene, termed the 'defector' gene, was deleted from a set of training genes and added to 99 randomly selected test genes. The software then determined the ranking of this defector gene for every data source separately. We used 627 training genes, ordered in 29 training sets of particular diseases automatically selected from the Online Mendelian Inheritance In Man (OMIM) database (see **Supplementary Notes** online for selection procedure). For pathway genes, we compiled three sets of training genes involved in the WNT (43 genes), NOTCH (18 genes) and epidermal growth factor (15 genes) pathways. As a negative control for training genes, we assembled 10 sets of 20 randomly selected genes.

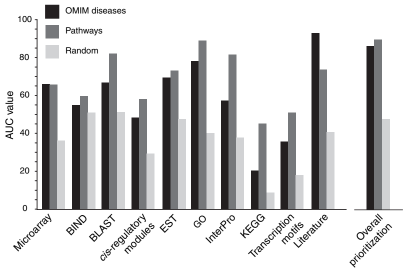


Figure 2 Cross-validation results. The AUC values obtained for all individual data sources are shown for disease prioritizations (black), pathway prioritizations (dark gray) and random prioritizations (light gray). The AUC values from the overall prioritization obtained after fusing all individual prioritizations are also shown.

Thus, a total of 903 prioritizations (627 for the disease genes, 76 for the pathway genes and 200 for the random sets) were performed for each data source. From these, we calculated sensitivity and specificity values. Sensitivity refers to the frequency (% of all prioritizations) of defector genes that are ranked above a particular threshold position. Specificity refers to the percentage of genes ranked below this threshold. For instance, a sensitivity/specificity value of 70/90 would indicate that the correct disease gene was ranked among the best-scoring 10% of genes in 70% of the prioritizations. To allow comparison between data sources we plotted rank receiver operating characteristic (ROC) curves, from which sensitivity/specificity values can be easily deduced. The area under this curve (AUC) is a standard measure of the performance of this algorithm. For instance, an AUC-value of 100% indicates that every defector gene ranked first, whereas a value of 50% means that the defector genes ranked randomly.

For every single data source, Endeavour reached a higher AUC score for disease and pathway genes than for randomly selected genes, indicating that it was sensitive and specific in ranking the defector gene, regardless of the type of data source consulted (**Fig. 2**). Not surprisingly, the data sources differed in their usefulness and suitability to rank genes (**Supplementary Notes**).

Overall prioritization by fusing multiple data sources

Although in most cases the defector gene ranked high in the prioritization list, this was not always the case (**Supplementary Fig. 1** online).

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

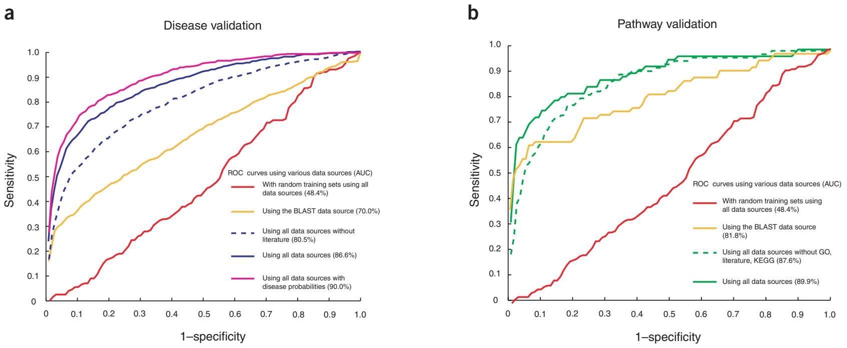


Figure 3 Cross-validation results. (**a**) Rank ROC curves obtained for the disease validation. (**b**) Rank ROC curves obtained for the pathway validation. In both figures, the control ROC curve (red line) was obtained after prioritization with randomly constructed training sets and by using all data sources. For all other ROC curves, disease or pathway-specific training sets were generated. The data sources used to construct every ROC curve are indicated on the figure.

To minimize this variability and to increase the performance of ranking, we integrated all individual prioritizations into a single overall rank by implementing an algorithm based on order statistics. With this algorithm, the probability of finding a gene at all the observed positions is calculated and a single overall rank is obtained by ranking genes according to these probabilities. To evaluate the performance of this overall ranking, we calculated its AUC values, as described above for the individual data sources. The AUC scores were 86.6% and 89.9% for disease and pathway genes compared to 48.4% for randomly selected genes (**Fig. 3a**,**b**). The correct pathway gene ranked among the top 50% of test genes in 95% of the cases, or among the top 10% in 74% of the cases. The variability of the overall prioritization was substantially smaller than that of individual data sources (**Supplementary Fig. 1**), and each of the data sources contributed to the overall ranking (**Supplementary Fig. 2** online). Our validation experiment thus results in biologically meaningful prioritizations.

Almost every data source but especially functionally annotated databases are incompletely annotated. For instance, only 63% of the genes are currently annotated in the GO database. Consequently, existing methods using these data sources introduce an undesired bias toward better-studied genes. Our approach should suffer less from these shortcomings as it also uses sequence-based sources containing information about known and unknown genes. In support of this, we found that the overall ranking of defector genes was not substantially influenced by the number of data sources if at least three sources with data annotations were available (**Supplementary Fig. 3a** online). In fact, even unknown genes lacking a

Table 1 Prioritizations of recently identified monogenic disease genes

Disease	Gene	Ensembl ID	Publication date	Rank position using the indicated data sources: All	Rank position using the indicated data sources: Literature
Arrhythmia	*CACNA1C*	ENSG00000151067	October 2004 (ref. 34)	4	3
Congenital heart disease	*CRELD1*	ENSG00000163703	April 2003 (ref. 35)	3	1
Cardiomyopathy 1	*CAV3*	ENSG00000182533	January 2004 (ref. 36)	2	1
Parkinson disease	*LRRK2*	ENSG00000188906	November 2004 (ref. 37)	50	*
Charcot-Marie-Tooth disease	*DNM2*	ENSG00000079805	March 2005 (ref. 38)	14	100
Amyotrophic lateral sclerosis	*DCTN1*	ENSG00000135406	August 2004 (ref. 39)	27	97
Klippel-Trenaunay disease	*AGGF1* (also known as *VG5Q*)	ENSG00000164252	February 2004 (ref. 40)	3	39
Cardiomyopathy 2	*ABCC9*	ENSG00000069431	April 2004 (ref. 41)	1	51
Distal hereditary motor neuropathy	*BSCL2*	ENSG00000168000	March 2004 (ref. 42)	15	62
Cornelia de Lange syndrome	*NIPBL*	ENSG00000164190	June 2004 (refs. 43,44)	9	75
			Average rank	13 ± 5	48 ± 13

For all genes, a mutation was inherited in a mendelian fashion (or was shown to cause the disease phenotype). The name of the disease and disease-causing gene, the Ensembl ID and the publication date of the article reporting the gene mutation (month-year) are shown, together with the rank (out of 200 test genes) at which they were prioritized by Endeavour, using all data sources or using the pre-publication date literature source alone. The average rank (mean ± s.e.m.) for each prioritization is indicated. For *LRRK2*, no literature information was available. This has been indicated in the table by an asterisk (*).

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

Table 2 Prioritizations of recently identified polygenic disease genes

Disease	Gene	Ensembl ID	Publication date	Rank
Atherosclerosis 1	*TNFSF4*	ENSG00000117586	April 2005 (ref. 45)	54
Crohn disease	*SLC22A4, SLC22A5*	ENSG00000197208	May 2004 (ref. 46)	71
Parkinson disease	*GBA*	ENSG00000188906	November 2004 (47)	23
Rheumatoid arthritis	*PTPN22*	ENSG00000134242	August 2004 (ref. 48)	11
Atherosclerosis 2	*ALOX5AP*	ENSG00000132965	February 2004 (ref. 49)	29
Alzheimer disease	*UBQLN1*	ENSG00000135018	March 2005 (ref. 50)	54
			Average rank	40 ± 10

The nature of the genetic variation in these genes was in each case a polymorphism, which typically was inherited as a risk factor for the respective disease. The name of the complex disease in which these genes were identified, their gene name, Ensembl ID and the publication date when the disease gene was reported as a susceptibility gene are given, together with the rank (out of 200 test genes) at which they have been prioritized by all data sources with rolled-back literature. The relative contribution of these genetic variations as risk factors for disease susceptibility will become clearer once replication studies are performed. The average rank (mean ± s.e.m.) for each prioritization is indicated.

HUGO name and with very little information available could be ranked highly (**Supplementary Fig. 3b**). Thus, our method takes into account data sources with relevant information, while disregarding noninformative ones. This may be particularly advantageous for the prioritization of disease genes, as unknown genes are not readily considered as disease candidates when selected manually.

Endeavour does not rely on literature-derived data alone

For each OMIM gene used in the disease validation, a mutation causing the disease had previously been reported in a landmark study. Because the inclusion of these publications may artificially increase the relative contribution of the literature data source in the overall performance of this algorithm, we excluded, as a test, the entire literature database from the disease validation protocol. For the same reason, the GO, KEGG and literature data sources were excluded from the pathway validation. Even under such unrealistic conditions where entire data sources were not used, the overall performance of the algorithm was only negligibly affected: the performance dropped by only 6.1% for disease genes (from 86.6% to 80.5%; **Fig. 3a**) and by only 2.3% for pathway genes (from 89.9% to 87.6%; **Fig. 3b**). Thus, the diversity of data sources used in our approach enables meaningful prioritizations, even without the use of literature information.

Clearly, this caution is only of importance in the context of a validation. In a more realistic situation, when the precise function of a disease gene is not known yet, the literature could still provide valuable indirect information about other properties of a gene. In a study of ten monogenic diseases (see below), we mimicked this situation by using only 'rolled-back' literature information, available one year before the landmark publication. Even then Endeavour provided a high rank for three genes (position 1, 1 and 3 out of 200 test genes, **Table 1**), illustrating that the literature contributes to the prioritization of yet undiscovered disease genes. For the seven other genes, use of the literature as the only data source was not very efficient, but inclusion of all the other data sources yielded a high rank (**Table 1**). Overall, even though the literature may provide valuable information, our method does not rely on literature as the only critical data source. But also, its performance is not restricted by the lack of available literature data, because of its ability to access and integrate multiple other data sources.

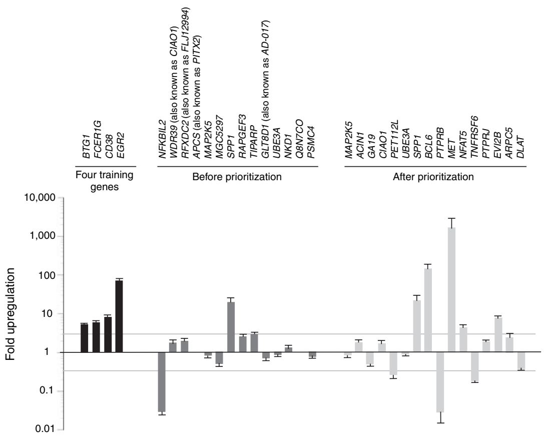


Figure 4 *In vitro* functional validation of Endeavour. Results of real-time quantitative PCR measurements in differentiated versus undifferentiated HL-60 cells. Expression profiles of 4 out of 18 training genes (left), which were tested as a positive control, and 20 target genes predicted by the *cis*-regulatory module model (center) are shown. Expression levels of *SPP1* and *NGKBIL2* differed more than threefold between differentiated and undifferentiated cells; expression levels for six genes could not be measured. The expression profiles of the 20 highest-ranking target genes after prioritization by Endeavour (right) are also shown. Expression levels of eight genes (*SPP1, BCL6, PTPRB, MET, TNFRSF6, NFAT5, PET112L* and *EVI2B*) differed more than threefold between differentiated and undifferentiated cells; four genes could not be measured. The fold difference is depicted on a logarithmic scale; error bars represent the s.e.m. The line indicates the threshold (threefold up- or downregulation).

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

Use of disease-specific data sources

An important asset of Endeavour is that its framework was designed to allow the inclusion of additional data sources, such as disease-related features, in the prioritization strategy. We illustrate this for the prioritization of disease genes. On the basis of a number of selection criteria (e.g., protein length, phylogenetic conservation), Lopez-Bigas and Adie determined for every gene a 'general' disease probability, or its probability as a disease candidate gene[8,9]. When integrating the Lopez-Bigas or Adie criteria in Endeavour as an additional data source, we found that its performance improved further (AUC scores increased by up to 5% regardless of the inclusion of literature sources). Likewise, microarray data specific for the process or disease under study can be included. Our approach thus allows the user to add, in a flexible and modular manner, additional data sources, such as appropriate disease-specific data sources, to enhance its overall performance.

Prioritization of genes causing monogenic diseases

In the large-scale validation, 627 genes were automatically selected from the OMIM database, without taking their mono- or polygenic nature into account. We therefore assessed whether our approach could be used to prioritize genes that cause monogenic diseases. As experimentalists often prefer to select their own sets of training genes, instead of relying on automatically derived genes or characteristics, we selected ten monogenic diseases and constructed sets of training genes together with a biological expert (**Table 1** and **Supplementary Table 1**). To simulate the real life situation, we deliberately chose recently identified disease-causing genes, and used rolled-back literature together with all other data sources. The set of test genes included the gene causing the monogenic disease, and 199 genes flanking its immediate chromosomal surroundings. The algorithm gave the ten monogenic disease–causing genes an average rank of 13 ± 5 out of 200 test genes (**Table 1**). When using a training set not related to the disease under study to prioritize the test sets as a negative control, the disease genes ranked randomly (position 96 on average). As a further validation the algorithm was applied to a very large set of test genes (that is, all 1,048 genes from chromosome 3; **Supplementary Notes** and **Supplementary Table 2** online).

This pseudo-prospective analysis, using rolled-back literature, reveals that expert-based construction of training sets may lead to high discovery rates when hunting for monogenic disease genes in both small and large test sets.

Prioritization of genes underlying polygenic diseases

In many cases, human disease is not monogenic, but polygenic in nature. We therefore prioritized six genes, recently identified as polygenic disease genes, together with 199 chromosomal flanking genes (**Table 2**). The sets of training genes used for these prioritizations are explained in **Supplementary Table 1**. On average, the susceptibility genes ranked at position 40 ± 10, when using the rolled-back literature together with all the other data sources. As expected, the prioritization of polygenic disease candidate genes is a greater challenge than ranking monogenic disease genes. Nonetheless, the ranking was still specific, as the susceptibility genes ranked at position 96 ± 10, when training sets for these disorders were randomly assigned to other test sets as a negative control. Thus, although the performance is lower than for monogenic diseases (as anticipated), susceptibility genes to polygenic diseases can be enriched by Endeavour's prioritization.

Prioritization of regulatory pathway genes

To analyze whether Endeavour could also rank genes involved in a particular biological process, we combined computation with functional validation *in vitro*. First, using the previously characterized ModuleSearcher

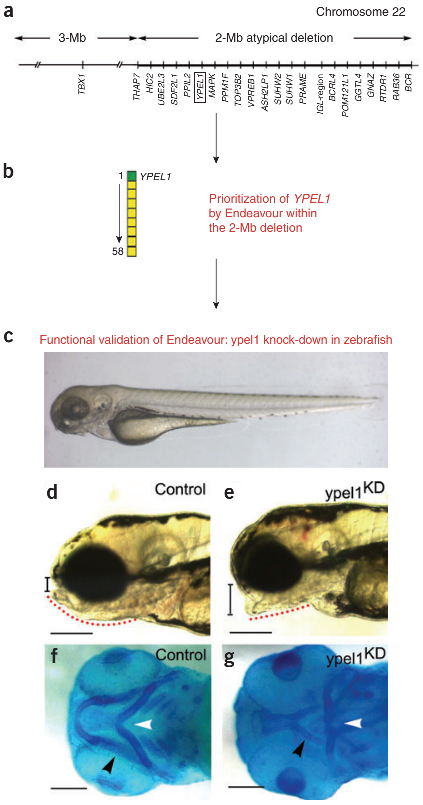
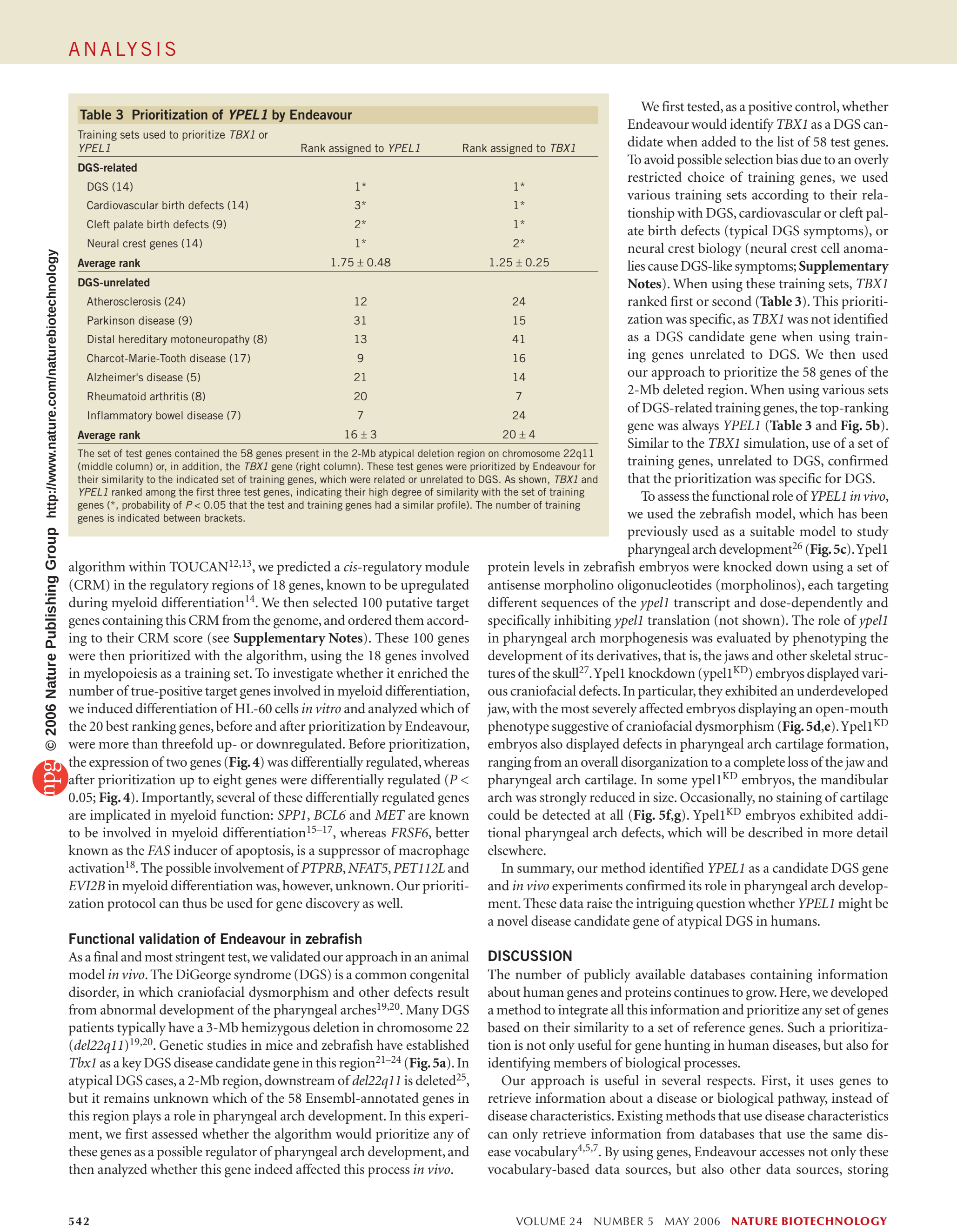


Figure 5 Functional validation of Endeavour in zebrafish. (**a**) Part of chromosome 22, illustrating the hemizygous 3-Mb region deleted in many DGS patients and the atypical 2-Mb region, which is deleted in some (atypical) DGS patients. For clarity, only some of the 58 Ensembl-annotated genes within the 2-Mb region, and only *TBX1* in the 3-Mb deleted region, are shown. It remains unknown whether any of the genes in the 2-Mb region play a role in pharyngeal arch development defects seen in DGS. (**b**) *YPEL1* was prioritized among the 58 genes of the 2-Mb deleted region by Endeavour as the most likely candidate involved in pharyngeal arch development. (**c**) Photo of a zebrafish, which has been used as a suitable model to study the role of *YPEL1* in pharyngeal arch development. (**d**,**e**) Lateral view of the head in live embryos at 4 d after fertilization. The lower jaw is clearly visible in the control, whereas ypel1^{KD} embryos show an underdeveloped lower jaw (mandibular arch; indicated by the red dotted line) and open mouth (indicated by the vertical line). (**f**,**g**) Ventral view of the pharyngeal arch cartilage using alcian blue stain at 3 d after fertilization. Black arrow depicts the mandibular arch; white arrow depicts hyoid arch. In ypel1^{KD} embryos, the jaw arches were severely malformed with the mandibular arch often reduced in size. The pharyngeal arch cartilage also showed reduced or no staining.



© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

Table 3 Prioritization of *YPEL1* by Endeavour

Training sets used to prioritize *TBX1* or *YPEL1*	Rank assigned to *YPEL1*	Rank assigned to *TBX1*
DGS-related		
DGS (14)	1*	1*
Cardiovascular birth defects (14)	3*	1*
Cleft palate birth defects (9)	2*	1*
Neural crest genes (14)	1*	2*
Average rank	1.75 ± 0.48	1.25 ± 0.25
DGS-unrelated		
Atherosclerosis (24)	12	24
Parkinson disease (9)	31	15
Distal hereditary motoneuropathy (8)	13	41
Charcot-Marie-Tooth disease (17)	9	16
Alzheimer's disease (5)	21	14
Rheumatoid arthritis (8)	20	7
Inflammatory bowel disease (7)	7	24
Average rank	16 ± 3	20 ± 4

The set of test genes contained the 58 genes present in the 2-Mb atypical deletion region on chromosome 22q11 (middle column) or, in addition, the *TBX1* gene (right column). These test genes were prioritized by Endeavour for their similarity to the indicated set of training genes, which were related or unrelated to DGS. As shown, *TBX1* and *YPEL1* ranked among the first three test genes, indicating their high degree of similarity with the set of training genes (*, probability of $P < 0.05$ that the test and training genes had a similar profile). The number of training genes is indicated between brackets.

algorithm within TOUCAN[12,13], we predicted a *cis*-regulatory module (CRM) in the regulatory regions of 18 genes, known to be upregulated during myeloid differentiation[14]. We then selected 100 putative target genes containing this CRM from the genome, and ordered them according to their CRM score (see **Supplementary Notes**). These 100 genes were then prioritized with the algorithm, using the 18 genes involved in myelopoiesis as a training set. To investigate whether it enriched the number of true-positive target genes involved in myeloid differentiation, we induced differentiation of HL-60 cells *in vitro* and analyzed which of the 20 best ranking genes, before and after prioritization by Endeavour, were more than threefold up- or downregulated. Before prioritization, the expression of two genes (**Fig. 4**) was differentially regulated, whereas after prioritization up to eight genes were differentially regulated ($P <$ 0.05; **Fig. 4**). Importantly, several of these differentially regulated genes are implicated in myeloid function: *SPP1*, *BCL6* and *MET* are known to be involved in myeloid differentiation[15–17], whereas *FRSF6*, better known as the *FAS* inducer of apoptosis, is a suppressor of macrophage activation[18]. The possible involvement of *PTPRB*, *NFAT5*, *PET112L* and *EVI2B* in myeloid differentiation was, however, unknown. Our prioritization protocol can thus be used for gene discovery as well.

Functional validation of Endeavour in zebrafish

As a final and most stringent test, we validated our approach in an animal model *in vivo*. The DiGeorge syndrome (DGS) is a common congenital disorder, in which craniofacial dysmorphism and other defects result from abnormal development of the pharyngeal arches[19,20]. Many DGS patients typically have a 3-Mb hemizygous deletion in chromosome 22 (*del22q11*)[19,20]. Genetic studies in mice and zebrafish have established *Tbx1* as a key DGS disease candidate gene in this region[21–24] (**Fig. 5a**). In atypical DGS cases, a 2-Mb region, downstream of *del22q11* is deleted[25], but it remains unknown which of the 58 Ensembl-annotated genes in this region plays a role in pharyngeal arch development. In this experiment, we first assessed whether the algorithm would prioritize any of these genes as a possible regulator of pharyngeal arch development, and then analyzed whether this gene indeed affected this process *in vivo*.

We first tested, as a positive control, whether Endeavour would identify *TBX1* as a DGS candidate when added to the list of 58 test genes. To avoid possible selection bias due to an overly restricted choice of training genes, we used various training sets according to their relationship with DGS, cardiovascular or cleft palate birth defects (typical DGS symptoms), or neural crest biology (neural crest cell anomalies cause DGS-like symptoms; **Supplementary Notes**). When using these training sets, *TBX1* ranked first or second (**Table 3**). This prioritization was specific, as *TBX1* was not identified as a DGS candidate gene when using training genes unrelated to DGS. We then used our approach to prioritize the 58 genes of the 2-Mb deleted region. When using various sets of DGS-related training genes, the top-ranking gene was always *YPEL1* (**Table 3** and **Fig. 5b**). Similar to the *TBX1* simulation, use of a set of training genes, unrelated to DGS, confirmed that the prioritization was specific for DGS.

To assess the functional role of *YPEL1 in vivo*, we used the zebrafish model, which has been previously used as a suitable model to study pharyngeal arch development[26] (**Fig. 5c**). Ypel1 protein levels in zebrafish embryos were knocked down using a set of antisense morpholino oligonucleotides (morpholinos), each targeting different sequences of the *ypel1* transcript and dose-dependently and specifically inhibiting *ypel1* translation (not shown). The role of *ypel1* in pharyngeal arch morphogenesis was evaluated by phenotyping the development of its derivatives, that is, the jaws and other skeletal structures of the skull[27]. Ypel1 knockdown (ypel1$^{\mathrm{KD}}$) embryos displayed various craniofacial defects. In particular, they exhibited an underdeveloped jaw, with the most severely affected embryos displaying an open-mouth phenotype suggestive of craniofacial dysmorphism (**Fig. 5d,e**). Ypel1$^{\mathrm{KD}}$ embryos also displayed defects in pharyngeal arch cartilage formation, ranging from an overall disorganization to a complete loss of the jaw and pharyngeal arch cartilage. In some ypel1$^{\mathrm{KD}}$ embryos, the mandibular arch was strongly reduced in size. Occasionally, no staining of cartilage could be detected at all (**Fig. 5f,g**). Ypel1$^{\mathrm{KD}}$ embryos exhibited additional pharyngeal arch defects, which will be described in more detail elsewhere.

In summary, our method identified *YPEL1* as a candidate DGS gene and *in vivo* experiments confirmed its role in pharyngeal arch development. These data raise the intriguing question whether *YPEL1* might be a novel disease candidate gene of atypical DGS in humans.

DISCUSSION

The number of publicly available databases containing information about human genes and proteins continues to grow. Here, we developed a method to integrate all this information and prioritize any set of genes based on their similarity to a set of reference genes. Such a prioritization is not only useful for gene hunting in human diseases, but also for identifying members of biological processes.

Our approach is useful in several respects. First, it uses genes to retrieve information about a disease or biological pathway, instead of disease characteristics. Existing methods that use disease characteristics can only retrieve information from databases that use the same disease vocabulary[4,5,7]. By using genes, Endeavour accesses not only these vocabulary-based data sources, but also other data sources, storing

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

information about a gene (e.g., derived from a microarray experiment) or a gene sequence (e.g., BLAST sequence similarity). Moreover, by using genes, the method is also suitable for gene prioritization in biological processes as well.

Second, compared to existing methods, which access only one or two data sources[4–7], our method accesses many more data sources (currently up to 12). Importantly, consultation of each of the individual sources by Endeavour generates biologically relevant prioritizations. We developed an algorithm based on order statistics to fuse all these separate prioritizations into a single overall rank. This algorithm is able to handle genes with missing values, thereby minimizing the bias for known or well-characterized genes. This bias will decrease even further in the future, when new and better high-throughput data become available, and when the genome annotation and curation processes reach their finalization.

Third, the algorithm is publicly available as a software tool, built by bioinformaticians, but designed for experimentalists, helping them to focus readily on key biological questions. The only other available prioritization tool for diseases, G2D, uses GO and literature data sources and is therefore restricted in making predictions about annotated or known genes[5].

Fourth, the approach gives the user maximal control over the set of training and test genes. Biologists prefer the flexibility of interactively selecting their own set of genes over an automatic and noninteractive data-mining selection procedure.

We validated the method extensively, in a large-scale validation study of 703 disease and pathway genes, and in a number of case-specific analyses. The validation results were remarkably good: on average, the correct gene was ranked 10th out of 100 test genes—for monogenic diseases, the performance was even better. The algorithm was capable of prioritizing large test sets (up to 1,000 genes)—the upgrade of Endeavour into a package capable of prioritizing the entire genome would be an interesting perspective for the future. Functional validation studies *in vitro* further demonstrated that the method worked equally well for prioritization of pathway genes. Furthermore, *in vivo* studies in zebrafish revealed that *YPEL1*, a gene prioritized by Endeavour in a 2-Mb chromosomal region deleted in patients with craniofacial defects, indeed regulates morphogenesis of the pharyngeal arches and their craniofacial-derivative structures.

Lastly, the Endeavour software design is modular and allows the inclusion of publicly available or proprietary data sources (e.g., disease-specific microarray experiments). We have illustrated and validated this possibility by including the general disease probability criteria of Lopez-Bigas[9] and Adie[8].

In summary, we present a computational method for fast and interactive gene prioritization that fuses genomic data regardless of its origin.

METHODS

Data sources. A more detailed description of the data sources is available as **Supplementary Methods** online. Briefly, for information retrieved from attribute-based data sources (that is, Gene Ontology, EST expression, InterPro and KEGG), the algorithm uses a binomial statistic to select those attributes that are statistically overrepresented among the training genes, relative to their genome-wide occurrence. Each overrepresented attribute receives a P-value p_i that is corrected for multiple testing. For information retrieved from vector-based data sources (that is, literature, microarray expression data or *cis*-regulatory motif predictions), the algorithm constructs an average vector profile of the training set. The literature profile is based on indexed abstracts and contains inverse document frequencies for each term of a GO-based vocabulary[28]; the expression profile contains expression ratios; the motif profile contains scores of TRANSFAC position weight matrices, obtained by scanning promoter sequences of the training genes that are conserved with their respective mouse orthologous sequences. To rank a set of test genes, attribute-based data are scored by Fisher's omnibus meta-analysis (Σ-2logp_i), generating a new P-value from a χ^2-distribution. Vector-based data are scored by Pearson correlation between the test vector and the training average. The data in the BLAST, BIND and *cis*-regulatory module (CRM) databases are neither vector- nor attribute-based. For BLAST, the similarity score between a test gene and the training set is the lowest *e*-value obtained from a BLAST against an *ad hoc* indexed database consisting of the protein sequences of the training genes. For BIND (Biomolecular Interaction Network Database)[29], the similarity score is calculated as the overlap between all protein-protein interaction partners of the training set and those of the test gene. Lastly, for CRM data, the best combination of five clustered transcription factor binding sites—in all human-mouse conserved noncoding sequences (up to 10 kb upstream of transcription start site) of the training genes—is determined using a genetic algorithm[12,30]. The similarity of this trained model to a test gene is determined by scoring this motif combination on the conserved noncoding sequences of the test gene.

Order statistics. The rankings from the separate data sources are combined using order statistics. A Q statistic is calculated from all rank ratios using the joint cumulative distribution of an N-dimensional order statistic as previously done by Stuart *et al.*[31]

$$Q(r_1,r_2,\ldots,r_N) = N!\int_0^{r_1}\int_{s_1}^{r_2}\ldots\int_{s_{N-1}}^{r_N} ds_N ds_{N-1}\ldots ds_1 \qquad (1)$$

They propose the following recursive formula to compute the above integral:

$$Q(r_1,r_2,\ldots,r_N) = N!\sum_{i=1}^{N}(r_{N-i+1}-r_{N-i})Q(r_1,r_2,\ldots,r_{N-i},r_{N-i+2},\ldots,r_N) \qquad (2)$$

where r_i is the rank ratio for data source i, N is the number of data sources used, and $r_0 = 0$. However, two problems arose when we tried to use this formula. First, we noticed that this formula is highly inefficient for moderate values of N, and even intractable for $N > 12$ because its complexity is $O(N!)$. We therefore implemented a much faster alternative formula with complexity $O(N^2)$:

$$V_k = \sum_{i=1}^{k}(-1)^{i-1}\frac{V_{k-i}}{i!}r_{N-k+1}^{i} \qquad (3)$$

with $Q(r_1,r_2,\ldots,r_N) = N!V_N$, $V_0 = 1$, and r_i is the rank ratio for data source i.

Second, we noticed that the Q statistics calculated by (1) are not uniformly distributed under the null hypothesis and can thus not directly be used as P-values. Therefore, we fitted a distribution for every possible number of ranks and used this distribution to calculate an approximate P-value. We found that the Q statistics for $N \le 5$ randomly and uniformly drawn rank-ratios are approximately distributed according to a beta distribution. For $N > 5$ the distributions can be modeled by a gamma distribution. The cumulative distribution function of these distributions provides us with a P-value for every Q statistic from the order statistics. Next to the original N rankings, we now have an $(N + 1)^{th}$ that is the combined rank of all separate ranks.

Cell culture, RNA isolation and RT-PCR. HL-60 cells were grown in RPMI 1640, supplemented with 10% FCS. Differentiation was induced by 10 nM phorbol 12-myristate 13-acetate (PMA), when cells were grown to a density of 7×10^5/ml. Before induction and 24 h after induction, cells were harvested by centrifugation and RNA was isolated using the trizol reagent (Invitrogen), and subsequently treated with Turbo DNA-free DNase (Ambion). First-strand cDNA was synthesized using Superscript II reverse transcriptase (Invitrogen). Real-time quantitative PCR was performed using the qPCR core kit for SYBR green (Eurogentec), on an ABI PRISM 7700 SDS (Applied BioSystems). The mRNA levels were normalized to the geometric average of four different housekeeping genes: *ACTB*, *GAPDH*, *UBC* and *HPRT1*. Numbers of differentially expressed genes before and after prioritization were compared with a chi-square test.

© 2006 Nature Publishing Group http://www.nature.com/naturebiotechnology

Zebrafish care and embryo manipulations. Wild-type zebrafish (*Danio rerio*) of the AB strain were maintained under standard laboratory conditions[32]. Morpholino oligonucleotides (Gene Tools) were injected into one- to four-cell-stage embryos[27]. Alcian blue cartilage staining was carried out as previously described[33]. All animal studies were reviewed and approved by the institutional animal care and use committee for Medical Ethics and Clinical Research of the University of Leuven.

Software availability. Endeavour is freely available for academic use as a Java application at http://www.esat.kuleuven.be/endeavour.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We wish to thank all groups and consortia that made their data freely available: Ensembl, NCBI (EntrezGene and Medline), Gene Ontology, BIND, KEGG, Atlas, InterPro, BioBase, the Disease Probabilities from Lopez-Bigas and Ouzounis[9] and the Prospectr scores from Euan Adie[8]. Ouzounis[8] and the Prospectr scores from Euan Adie[9]. We also thank the following people for their help in particular areas: Robert Vlietinck with the manuscript, Patrick Glenisson with text mining, Joke Allemeersch and Gert Thijs with the order statistics and Camilla Esguerra with the zebrafish experiments. S.A., D.L. and P.V.L. are sponsored by the Research Foundation Flanders (FWO). This work is supported by Flanders Institute for Biotechnology (VIB), Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT) (STWW-00162), Research Council KULeuven (GOA-Ambiorics, IDO genetic networks), FWO (G.0229.03 and G.0413.03), IUAP V-22, K.U.L. Excellentiefinanciering CoE SymBioSys (EF/05/007), EU NoE Biopattern and EU EST BIOPTRAIN to Y.M., and by the FWO (G.0405.06), GOA/2006/11 and GOA/2001/09, Squibb and EULSHB-CT-2004-503573 to P.C.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

---

1. Quackenbush, J. Genomics. Microarrays—guilt by association. *Science* **302**, 240–241 (2004).

2 Kanehisa, M. & Bork, P. Bioinformatics in the post-sequence era. *Nat. Genet.* **33** Suppl. 305–310 (2003).

3. Ball, C.A., Sherlock, G. & Brazma, A. Funding high-throughput data sharing. *Nat. Biotechnol.* **22**, 1179–1183 (2004).
4. Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18** Suppl. 2, S110–S115 (2002).
5. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316–319 (2002).
6. Turner, F.S., Clutterbuck, D.R. & Semple, C.A. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* **4**, R75 (2003).
7. Tiffin, N. *et al.* Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* **33**, 1544–1552 (2005).
8. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**, 55 (2005).
9. Lopez-Bigas, N. & Ouzounis, C.A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* **32**, 3108–3114 (2004).
10. Kent, W.J. *et al.* Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737–741 (2005).
11. Altermann, E. & Klaenhammer, T.R. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics* **6**, 60 (2005).
12. Aerts, S. *et al.* TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.* **33**, W393–W396 (2005).
13. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. & De Moor, B. Computational detection of cis-regulatory modules. *Bioinformatics* **19** (Suppl 2), II5–II14 (2003).
14. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
15. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet.* **36**, 257–263 (2004).
16. Pixley, F.J. *et al.* BCL6 suppresses RhoA activity to alter macrophage morphology and motility. *J. Cell Sci.* **118**, 1873–1883 (2005).
17. Galimi, F. *et al.* Hepatocyte growth factor is a regulator of monocyte-macrophage function. *J. Immunol.* **166**, 1241–1247 (2001).
18. Brown, N.J. *et al.* Fas death receptor signaling represses monocyte numbers and macrophage activation in vivo. *J. Immunol.* **173**, 7584–7593 (2004).
19. Scambler, P.J. The 22q11 deletion syndromes. *Hum. Mol. Genet.* **9**, 2421–2426 (2000).
20. Baldini, A. Dissecting contiguous gene defects: TBX1. *Curr. Opin. Genet. Dev.* **15**, 279–284 (2005).
21. Jerome, L.A. & Papaioannou, V.E. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat. Genet.* **27**, 286–291 (2001).
22. Merscher, S. *et al.* TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell* **104**, 619–629 (2001).
23. Lindsay, E.A. *et al.* Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410**, 97–101 (2001).
24. Piotrowski, T. *et al.* The zebrafish van gogh mutation disrupts tbx1, which is involved in the DiGeorge deletion syndrome in humans. *Development* **130**, 5043–5052 (2003).
25. Rauch, A. *et al.* A novel 22q11.2 microdeletion in DiGeorge syndrome. *Am. J. Hum. Genet.* **64**, 659–666 (1999).
26. Graham, A. The development and evolution of the pharyngeal arches. *J. Anat.* **199**, 133–141 (2001).
27. Stalmans, I. *et al.* VEGF: a modifier of the del22q11 (DiGeorge) syndrome? *Nat. Med.* **9**, 173–182 (2003).
28. Glenisson, P. *et al.* TXTGate: profiling gene groups with text-based information. *Genome Biol.* **5**, R43 (2004).
29. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
30. Aerts, S., Van Loo, P., Moreau, Y. & De Moor, B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics* **20**, 1974–1976 (2004).
31. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
32. Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish*, (University of Oregon Press, Eugene, Oregon, 1994).
33. Kimmel, C.B. *et al.* The shaping of pharyngeal cartilages during early development of the zebrafish. *Dev. Biol.* **203**, 245–263 (1998).
34. Splawski, I. *et al.* Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
35. Robinson, S.W. *et al.* Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects. *Am. J. Hum. Genet.* **72**, 1047–1052 (2003).
36. Hayashi, T. *et al.* Identification and functional analysis of a caveolin-3 mutation associated with familial hypertrophic cardiomyopathy. *Biochem. Biophys. Res. Commun.* **313**, 178–184 (2004).
37. Zimprich, A. *et al.* Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
38. Zuchner, S. *et al.* Mutations in the pleckstrin homology domain of dynamin 2 cause dominant intermediate Charcot-Marie-Tooth disease. *Nat. Genet.* **37**, 289–294 (2005).
39. Munch, C. *et al.* Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* **63**, 724–726 (2004).
40. Tian, X.L. *et al.* Identification of an angiogenic factor that when mutated causes susceptibility to Klippel-Trenaunay syndrome. *Nature* **427**, 640–645 (2004).
41. Bienengraeber, M. *et al.* ABCC9 mutations identified in human dilated cardiomyopathy disrupt catalytic KATP channel gating. *Nat. Genet.* **36**, 382–387 (2004).
42. Windpassinger, C. *et al.* Heterozygous missense mutations in BSCL2 are associated with distal hereditary motor neuropathy and Silver syndrome. *Nat. Genet.* **36**, 271–276 (2004).
43. Tonkin, E.T., Wang, T.J., Lisgo, S., Bamshad, M.J. & Strachan, T. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nat. Genet.* **36**, 636–641 (2004).
44. Krantz, I.D. *et al.* Exclusion of linkage to the CDL1 gene region on chromosome 3q26.3 in some familial cases of Cornelia de Lange syndrome. *Am. J. Med. Genet.* **101**, 120–129 (2001).
45. Wang, X. *et al.* Positional identification of TNFSF4, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nat. Genet.* **37**, 365–372 (2005).
46. Peltekova, V.D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet.* **36**, 471–475 (2004).
47. Aharon-Peretz, J., Rosenbaum, H. & Gershoni-Baruch, R. Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. *N. Engl. J. Med.* **351**, 1972–1977 (2004).
48. Begovich, A.B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**, 330–337 (2004).
49. Helgadottir, A. *et al.* The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat. Genet.* **36**, 233–239 (2004).
50. Bertram, L. *et al.* Family-based association between Alzheimer's disease and variants in UBQLN1. *N. Engl. J. Med.* **352**, 884–894 (2005).

Study 2

ModuleMiner: improved computational detection of *cis*-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues?

Contribution of the doctorandus: design of the study, development of the ModuleMiner method for computational detection of *cis*-regulatory modules, validation and application of the method and interpretation of the results.
This study has been prepublished online in Genome Biology 2008 (reference: Genome Biology 2008, 9:R66).

MODULEMINER: improved computational detection of *cis*-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues?

Peter Van Loo[1,2,3,§], Stein Aerts[1,2], Bernard Thienpont[2], Bart De Moor[3], Yves Moreau[3], Peter Marynen[1,2]

[1]Department of Molecular and Developmental Genetics, VIB, Herestraat 49, box 602, B-3000 Leuven, Belgium

[2]Department of Human Genetics, University of Leuven, Herestraat 49, box 602, B-3000 Leuven, Belgium

[3]Bioinformatics group, Department of Electrical Engineering (ESAT-SCD), University of Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium

[§]Corresponding author

Email addresses:

PVL: Peter.VanLoo@med.kuleuven.be

SA: Stein.Aerts@med.kuleuven.be

BT: Bernard.Thienpont@med.kuleuven.be

BDM: Bart.DeMoor@esat.kuleuven.be

YM: Yves.Moreau@esat.kuleuven.be

PM: Peter.Marynen@med.kuleuven.be

Abstract

We present MODULEMINER, a novel algorithm for computationally detecting *cis*-regulatory modules (CRMs) in a set of co-expressed genes. MODULEMINER outperforms other methods for CRM detection on benchmark data and successfully detects CRMs in tissue-specific microarray clusters and in embryonic development gene sets. Interestingly, CRM predictions for differentiated tissues show a strong enrichment close to the transcription start site, while CRM predictions for embryonic development gene sets are depleted in this region.

List of abbreviations

AUC: area under the curve

CNS: conserved non-coding sequence

CRM: *cis*-regulatory module

GO: Gene Ontology

LOOCV: leave-one-out cross-validation

PWM: position weight matrix

ROC: receiver operator characteristic

TFBS: transcription factor binding site

TRGM: transcriptional regulatory global model

TRM: transcriptional regulatory model

TSS: transcription start site

Background

The identification and functional annotation of transcriptional regulatory sequences in the human genome is lagging far behind the rapidly increasing knowledge of protein-coding genes. These transcriptional regulatory sequences are often build up in a modular fashion and exert their function in *cis* through the concerted binding of multiple transcription factors (and co-factors), resulting in the formation of protein complexes that interact with RNA polymerase II [1,2]. These sequences are called *cis*-regulatory modules (CRMs). In theory, these CRMs can be detected by the presence of multiple transcription factor binding sites. However, in practice, the reliable detection of functional transcription factor binding sites is difficult and results in many false positives, partly because these binding sites are too short and too degenerate [3]. Hence, the computational detection of functional regulatory sequences in the human genome remains a formidable challenge.

Multiple method have been developed that aim to computationally detect regulatory sequences [4-8]. Promising and validated results have been delivered mostly in model organisms with relatively compact genomes (e.g. *Drosophila melanogaster*) [9-11]. In the larger human genome, deep sequence conservation (e.g. up to zebrafish) or extreme sequence conservation (e.g. perfect conservation in mouse over 200 base pairs), irrespective of transcription factor binding site detection, remains the method of choice for approaches validating regulatory sequences *in vitro* or *in vivo* [12-14]. While these conservation approaches are quite successful in predicting which regions have a regulatory function, they provide no information on what expression pattern these regions produce and by which transcription factors they are targeted.

When several similar CRMs have been characterized, and the regulatory factors and binding sites have been elucidated, one can use this knowledge to find new examples of similar CRMs directing the transcription of other genes involved in the same process. A number of computational methods have been described that apply this approach [15-17]. These methods have been highly successful [10,11,18], but in practice, apart from in *Drosophila* embryonic development, the lack of available data often precludes the application of these approaches.

When this knowledge is not available, the detection of tissue- or process-specific CRMs can be tackled by looking for recurring combinations of transcription factor binding sites in putative regulatory regions of a set of co-expressed genes. A few methods applying this approach have been developed [19-22]. However, in part because this is a more complex problem, these methods have only been applied on a limited scale and did not report many successful predictions. To our knowledge, only our ModuleSearcher method [20] has provided results subjected to experimental validation [23].

Here, we develop MODULEMINER, a novel algorithm to detect similar CRMs in a set of co-expressed genes, focussed on the human genome. MODULEMINER does not require prior knowledge of regulating transcription factors or annotated binding sites, but uses only a library of position weight matrices (PWMs). Contrary to existing algorithms that require *a priori* unknown CRM properties (such as the length of the CRMs or the number of binding sites) as input parameters, MODULEMINER is parameterless. In addition, MODULEMINER differs from existing similar approaches in that it implements a whole-genome optimization strategy to specifically look for signals that discriminate the given co-expressed genes from all other genes in the genome. By leave-one-out cross-validation on benchmark data, we show that

MODULEMINER outperforms other methods that computationally detect CRMs. Finally, we demonstrate that MODULEMINER can successfully detect similar CRMs in microarray clusters with a tissue specific expression profile, as well as in custom-build gene sets related to specific embryonic developmental processes. In total, MODULEMINER predicted 257 CRMs near the genes studied, as well as an additional 1400 CRM predictions resulting from full genome scans for new target genes. We further analyze these CRM predictions to elucidate differences between CRMs directing transcription in differentiated tissues and CRMs directing transcription during embryonic development.

Results

MODULEMINER: detection of similar *cis*-regulatory modules in a set of co-expressed genes

We developed MODULEMINER, a novel algorithm to detect similar CRMs in a set of co-expressed genes. MODULEMINER models similar CRMs as a combination of motifs (represented by PWMs), as in [20]. These models are called “transcriptional regulatory models” (TRMs) [24]. We postulate that a good TRM is able to retrieve targets in the genome. Therefore, we express the fitness of a TRM in terms of its target gene recovery and we select the TRM that has maximum specificity for the given set of co-expressed genes, by a whole-genome optimization strategy. To determine the fitness of a TRM, each gene’s search space is first scored with the TRM, where we define a gene’s search space as the collection of all conserved non-coding sequences within 10 kb 5’ of the transcription start site (see Materials and methods). These scores are then used to rank all genes in the genome. Finally, the ranks of the given co-expressed genes are determined, and the probability of observing this collection of ranks by chance is calculated using order statistics (see Materials and methods). If a large part of the co-expressed genes are ranked high, the order statistic is highly significant, and hence the TRM is considered to have a high fitness for modelling similar *cis*-regulatory modules regulating these genes. MODULEMINER searches the TRM with the most significant order statistic (i.e. the best fitness) using a genetic algorithm (detailed in Materials and methods).

We introduce MODULEMINER and its rigorous validation procedure by an example case study. We constructed a high quality set of 12 smooth muscle marker genes [25], and performed leave-one-out cross-validation (LOOCV). In each validation run, one gene was left out and MODULEMINER constructed a TRM using the remaining 11 genes. This TRM was then used to rank all genes in the genome and the position of the left-out gene was determined. The set of 12 ranks obtained in this way was used to calculate sensitivity/specificity pairs, which were subsequently plotted on a receiver operator characteristic (ROC) curve. We used the area under this curve (AUC) as a measure of MODULEMINER performance on this set of co-expressed genes.

We repeated the LOOCV for three sets of candidate transcription factor binding sites (TFBSs): (i) predicted binding sites in human-mouse conserved non-coding sequences (CNSs), obtained by aligning 10 kb 5' of all human-mouse orthologs and selecting regions of at least 75 % identity over a minimum of 100 base pairs; (ii) binding sites from (i), retaining only the PWMs for which in both the human and mouse CNS an instance is predicted (we follow the nomenclature in [10] and call these sites *preserved* sites); (iii) as in (ii), but here the CNSs are obtained by aligning 10 kb 5' of all human genes to 110 kb 5' + 100 kb 3' of the transcription start site of their mouse orthologs (and hence correcting for possible differences in transcription start site annotation) (Table 1). The resulting ROC curves are shown in Figure 1A. In all three cases, the AUC values are significantly above 50 % (the theoretical value obtained if the left-out genes would be ranked randomly), indicating that the TRMs obtained are sensitive and specific in predicting *cis*-regulatory modules near the left-out genes.

We observed that similar TRMs have a similar fitness and a similar order statistic. The TRM that is selected by MODULEMINER (the one that has the lowest order statistic) is surrounded by similar TRMs with order statistics that are only slightly larger. The selection of one TRM out of these similar TRMs is inherently arbitrary and depends only marginally on the true regulatory signals. To make MODULEMINER more robust to this "noise", we cluster the top-scoring TRMs and select the most prominent cluster, instead of the single optimal TRM. We call this cluster of TRMs a "transcriptional regulatory global model" (TRGM). The results of a LOOCV when using these TRGMs (Figure 1B) show that this indeed has a positive effect on MODULEMINER performance, as the AUCs increased by 6 % on average. Furthermore, these TRGMs provide additional information compared to singular TRMs, since they allow an estimate of the relative importance of each PWM involved, as discussed below.

Table 1. Genome-wide databases of candidate transcription factor binding sites

Nr	Database properties	Nr genes	Nr regions	Nr binding sites
1	human-mouse conserved regions, 10 kb 5' of TSS	8759	22582	1858800
2	(1) + limited to binding sites occurring both in the human and mouse CNS	8759	22582	878338
3	(2) + correct for possible mouse TSS differences (add 100 kb of mouse sequence 5' and 3')	11653	35021	1316927

When comparing the performance of MODULEMINER (using TRGMs) on the three sets of candidate binding sites, a large difference between selecting all detected binding sites (set 1, AUC value of 84.6 %) and restricting to preserved sites only (set 2, AUC value of 92.8 %) is apparent. Correcting for transcription start site (TSS) differences in human and mouse (set 3, AUC value of 92.5 %) did not increase this performance further. Thus, for this high quality set of co-expressed genes, the preservation of binding sites is highly beneficial for efficient detection of *cis*-regulatory modules. This strongly suggests that for this gene set, the trans-acting factors are conserved between human and mouse.

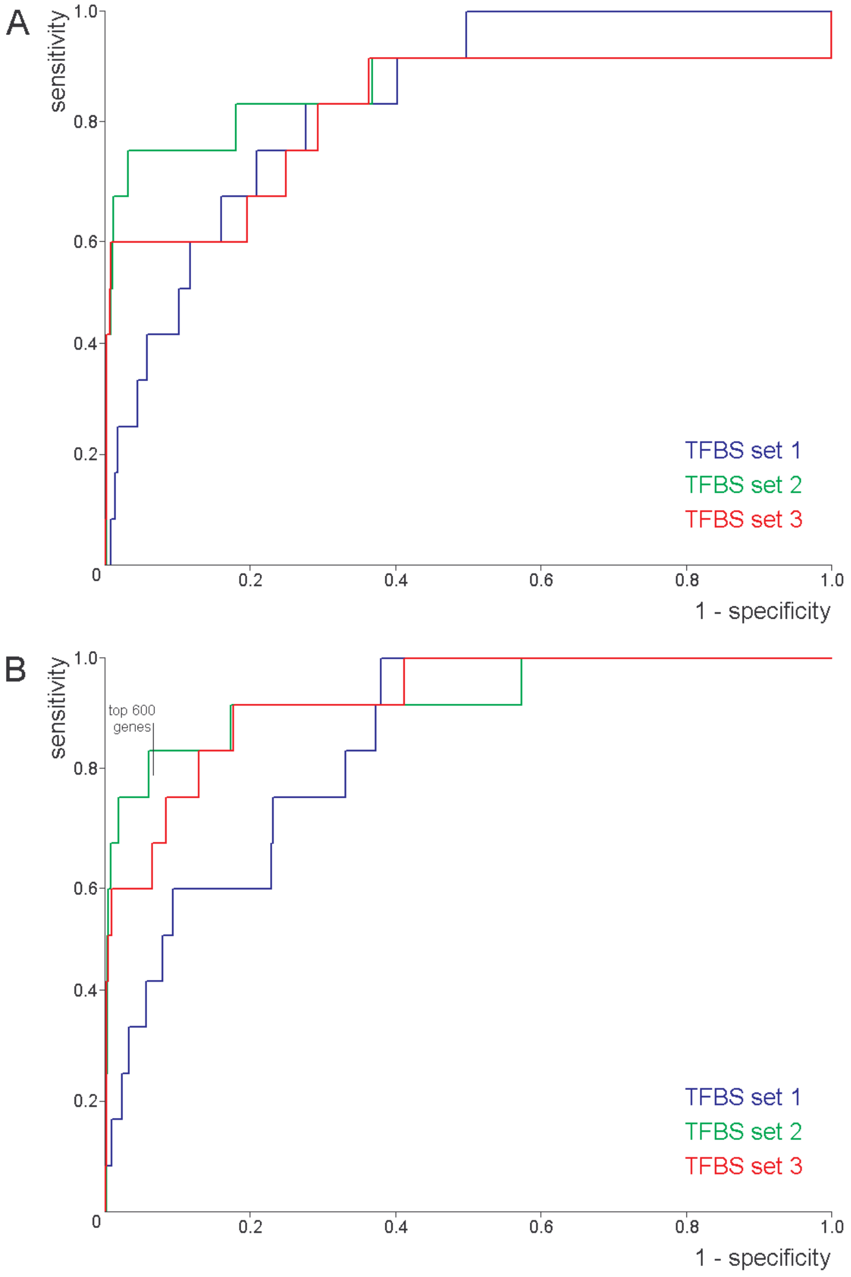


Figure 1. Performance of ModuleMiner on a set of smooth muscle marker genes, using the three different sets of candidate transcription factor binding sites. ROC curves are shown, representing results for leave-one-out cross-validations on the set of smooth muscle markers, (A) using singular transcriptional regulatory models (TRMs) and (B) using transcriptional regulatory global models (TRGMs).

We next applied the ModuleMiner algorithm to the full set of 12 smooth muscle marker genes, using the site preservation measure (set 2). The resulting TRGM identifies SRF, SMAD4, SP1 and ATF3 as the main transcription factors involved in the co-regulation of these genes (detailed ModuleMiner output is reported at our website). Importantly, ModuleMiner implicates SRF as the most important smooth muscle regulator, and suggests that smooth muscle specific regulation often entails two or more SRF binding sites, in agreement with literature [26].

To verify the added value of the resulting combination of PWMs over SRF alone, we manually generated a TRGM containing only PWMs for SRF and compared this to ModuleMiners performance. When we applied this "SRF only" TRGM to rank the genome, we obtained an AUC of 79.9 %, significantly smaller than the 92.8 % AUC of ModuleMiner (obtained in an LOOCV setting).

Sensitivity to noise

To assess the performance of ModuleMiner as a function of the composition of the input set of co-expressed genes, we performed LOOCV on input sets that contain a varying percentage of genuinely co-regulated genes ("true positives"). As true positive genes, we selected the set of 10 smooth muscle markers that share similar *cis*-regulatory modules that can be identified by ModuleMiner (these 10 genes all are ranked within the top 7 % of the genome by a LOOCV, as shown in Figure 1B). We approximated negative genes (genes that do not contain the smooth muscle *cis*-regulatory module) by random genes.

In a first analysis, we kept the number of true positive genes constant at 10, and we added a varying number of negative genes. The decrease in performance as a function of an increasing number of negative genes was surprisingly small (Figure 2). Even when only 10 of 50 genes contained the smooth muscle *cis*-regulatory module, MODULEMINER was able to pick up this signal (the AUC was 85.2 %, and SRF and SP1 were still found as key factors).

In a second analysis, we kept the total number of genes constant at 10, and we varied the percentage of negative genes. We now observed a steep decrease in MODULEMINER performance as a function of an increasing percentage of negative genes (Figure 2).

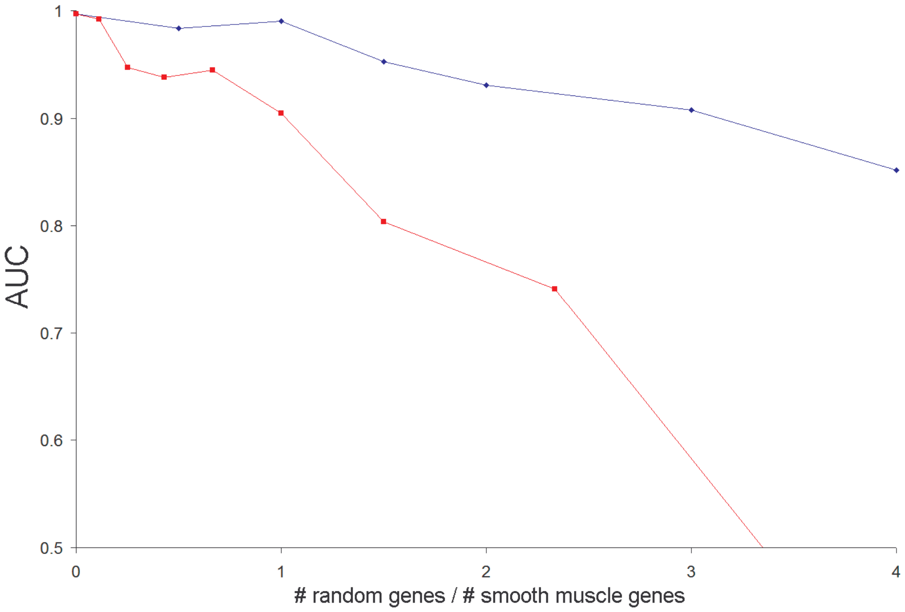


Figure 2. Sensitivity of MODULEMINER's performance to the quality of the input genes. The ratio of true positive genes (containing the smooth muscle CRM) to negative genes (approximated by random genes) was varied. Each time, a leave-one-out cross-validation was performed, an ROC curve was constructed, and the area under the curve was calculated. These AUCs were plotted as a function of the ratio negative genes/positive genes. As an AUC of 50 % signifies random ordering of the left-out genes (and hence indicates that no CRMs can be detected), this value was taken as the origin on the Y-axis. Blue: the number of positive genes was kept constant at 10, and the number of negative genes was varied. Red: the total number of genes was kept constant at 10, and the ratio negative genes/positive genes was varied.

We conclude from these experiments that MODULEMINER requires a critical mass of true positive genes for successful detection of similar *cis*-regulatory modules.

However, when this critical mass is present, MODULEMINER is highly robust to false positive genes.

Comparison to other CRM detection algorithms

We next compared MODULEMINER to other *in silico* approaches for CRM detection on benchmark data. From PAZAR [27], we selected all 'boutiques' containing annotated regulatory regions directing expression in a particular system: (i) M02, muscle; (ii) M03, liver; (iii) M08, ORegAnno Stat1 and (iv) M09, ORegAnno Erythroid. As a fifth benchmark set, we used the 12 smooth muscle genes described above. On each of these 5 sets, we compared the performance of MODULEMINER to that of 4 state-of-the-art publicly available algorithms to detect similar CRMs in co-expressed genes: ModuleSearcher [28], CREME [19], CisModule [22] and EMCMODULE [29]. We also included the Clover algorithm [30], which looks for individual overrepresented transcription factor binding sites in putative regulatory sequences of a set of co-expressed genes. We note that our analysis does not focus specifically on the known enhancers, but in contrast, we consider all CNSs in the entire 10 kb 5' of the TSS (which may or may not contain the known enhancer, as well as other sequences). This effectively mimics a real-life situation, where the exact location of the regulatory sequences is not known *a priori*.

The CREME algorithm was unable to identify similar CRMs in any of the 5 benchmark sets, most likely in part because of its focus on larger sets of more loosely co-expressed genes [19]. Using the remaining algorithms, we performed LOOCV on each of the 5 benchmark sets. For this LOOCV, we used each algorithm to train a TRM or TRGM using gene sets where one gene is left out (see Materials and methods for details). Hence, as training data, we used all CNSs in the 10 kb 5' of the TSS of the benchmark set, except for the left-out gene. For CisModule and EMCMODULE, the inputs were the sequences of the CNSs; for Clover, the inputs where the sequences of the CNSs as well as all TRANSFAC and JASPAR vertebrate PWMs; for ModuleSearcher, the inputs were the predicted binding sites within those CNSs, using all TRANSFAC and JASPAR vertebrate PWMs. The combination of PWMs that each algorithm provided as output was used to build a TRM or TRGM. We subsequently used the ModuleScanner algorithm to rank all genes in the genome based on the predicted TRM/TRGM, and we used the results to construct ROC curves. We used the site preservation measure (candidate TFBS set 2) for the MODULEMINER runs (as this was the set where we obtained the best results for the smooth muscle genes). Since the other algorithms do not use site preservation in the discovery step, we used candidate TFBS set 1 (without preservation) also in their genome ranking step. We also constructed random ROC curves based on genome ranking using random TRMs (see Materials and methods for details). On the OregAnno Erythroid benchmark set neither MODULEMINER nor any of the other algorithms seem to perform better than random (Figure 3A). As this is the smallest set, containing only 6 genes with human-mouse CNSs, this is consistent with the results we obtained in the previous section, where we concluded that a critical number of co-regulated genes is required for CRM detection. In contrast, on each of the 4 other benchmark sets, MODULEMINER performs better than random TRMs, as do some of the other algorithms (Figure 3B-E). Comparing the performance of all CRM detection algorithms, MODULEMINER seems to show the best performance in all 4 cases. Interestingly, only MODULEMINER can compete with “simple” transcription factor binding site overrepresentation in this setup, emulating a real-life situation where the regulatory sequences are not known. Indeed, only MODULEMINER outperforms Clover on four of the five benchmark sets.

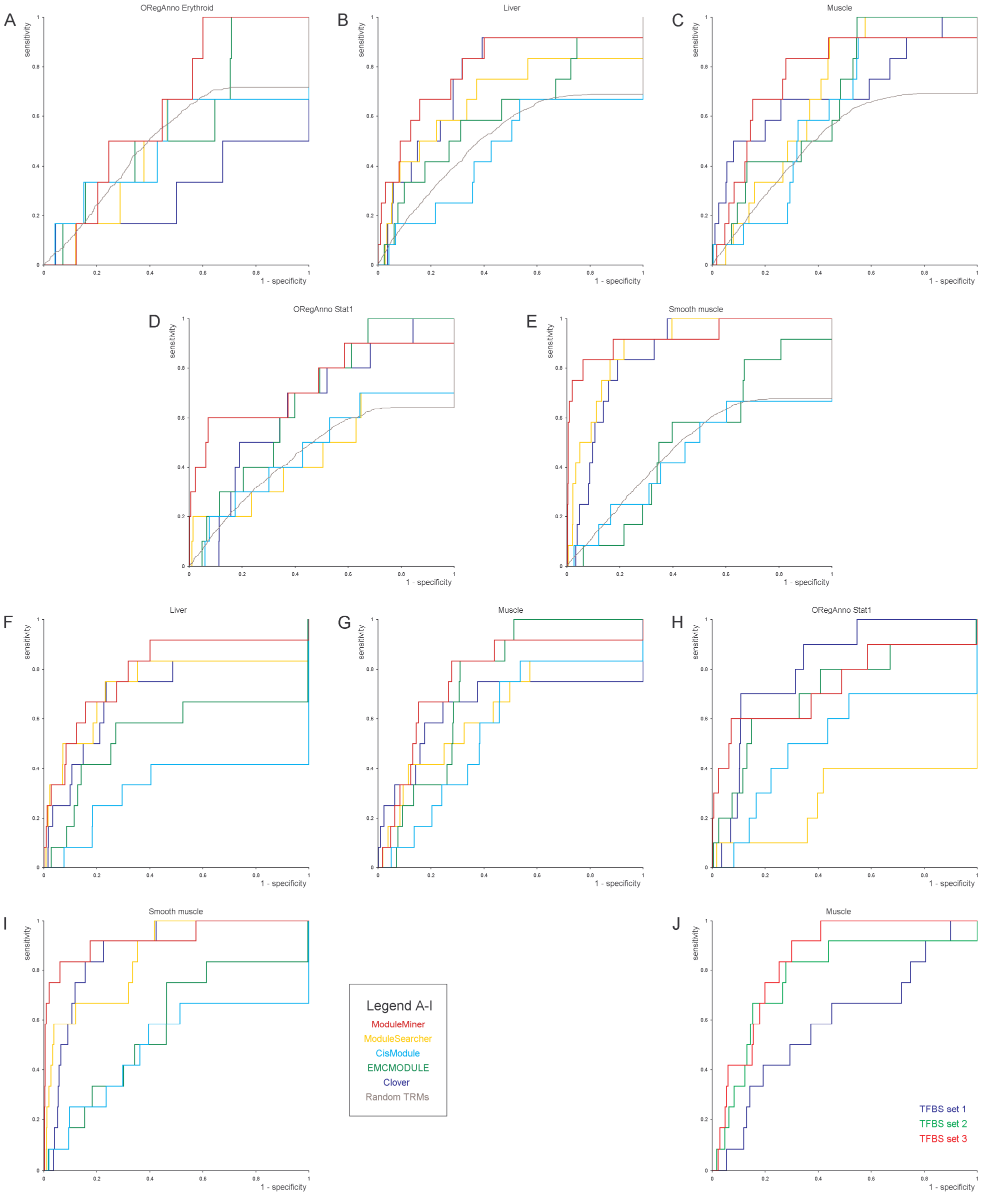


Figure 3. Comparison to other CRM detection algorithms. (A-E) ROC curves for the LOOCV using ModuleMiner, ModuleSearcher, CisModule, EMCMODULE, Clover and random TRMs for each of the 5 benchmark sets: (A) ORegAnno Erythroid, (B) liver, (C) muscle, (D) ORegAnno Stat1 and (E) smooth muscle. (F-I) ROC curves when using TFBS preservation (TFBS set 2) in the genome ranking step for all algorithms, on the 4 benchmark sets that performed above random: (F) liver, (G) muscle, (H) ORegAnno Stat1 and (I) smooth muscle. (J) ModuleMiner performance for the three TFBS sets on the muscle benchmark data.

On the fifth benchmark set (muscle), Clover and MODULEMINER seem to be closely matched, with the Clover method showing a steeper start of the ROC curve.

The performance of the other CRM detection algorithms can be improved by using site preservation (TFBS set 2) in the genome ranking step (Figure 3F-I), although here as well, MODULEMINER outperforms all other CRM detection algorithms, suggesting that the TRMs predicted by MODULEMINER are more informative or more specific than those suggested by other methods. Candidate TFBS set 2 was not in all cases the optimal choice for MODULEMINER: on the muscle benchmark set, candidate TFBS set 3 performed better (Figure 3J).

We noticed the CRM predictions MODULEMINER made on the muscle, liver and ORegAnno Stat1 sets, correspond well with the known regulatory elements. The TRGMs MODULEMINER contructed contain PWMs for SRF, MEF2, Myf and MyoD (muscle), HNF1, HNF3, HNF4 and CEBP (liver) and STAT (ORegAnno Stat1), even though we used all CNSs in the 10 kb upstream region. In addition, the CRM predictions mostly overlap the true enhancer, when the real regulatory sequence was in our CNS collection. Indeed, for the muscle set, in 9 of the 11 cases where the known enhancer was in our CNS set, MODULEMINER was ably to identify this region. For the liver set, MODULEMINER identified 7 out of 8 regulatory elements (data not shown).

Detection of *cis*-regulatory modules in microarray clusters

Realizing that clustering of microarray data provides a rich source of large co-expressed gene sets, where robustness to genes that are not co-regulated ("false positive genes") is critical, our sensitivity to noise analysis above encouraged us to apply MODULEMINER to microarray clusters on a larger scale. The GNF SymAtlas [31] contains expression profiles of 140 human and mouse tissues. Nelander *et al.*

[32] obtained gene clusters by hierarchically clustering this dataset, followed by a Pearson's correlation coefficient cut-off. From this clustering, we selected all clusters with at least 25 genes in our dataset (i.e. genes with at least one CNS within 10 kb 5' of the TSS). This results in 10 clusters with sizes ranging from 26 to 214 genes. Large clusters were randomly divided in a training set of 50 genes, and a test set containing the remaining genes.

As it was our goal here to identify similar *cis*-regulatory modules within a subset of the genes in each microarray cluster, we used a two-step procedure, first detecting which subset of genes potentially share *cis*-regulatory modules, and next detecting the actual *cis*-regulatory modules in their upstream regions (Figure 4A). The first step consisted of a five-fold cross-validation, where in each validation run, we used MODULEMINER to train a TRGM on four-fifth of the genes in a cluster, and next we determined which of the other one-fifth left-out genes were targets of the TRGM. If the total number of true target genes among left-out genes would not be significantly higher than random, we concluded that MODULEMINER is not able to detect similar CRMs within this cluster. If on the other hand there is a significant enrichment of these true target genes, we concluded that MODULEMINER is able to detect similar CRMs, and we use these high scoring genes in the second step. In this second step, MODULEMINER was applied to this focussed subcluster, identifying similar *cis*-regulatory modules regulating these genes. As an extra validation, LOOCV was used to confirm the presence of similar *cis*-regulatory modules, as done previously on the smooth muscle and other benchmark sets.

Application of this procedure to the microarray clusters described above resulted in successful *cis*-regulatory module detection in 9 of the 10 clusters (Table 2, Figure 4B). In each case, this success was confirmed by a LOOCV on the selected subcluster

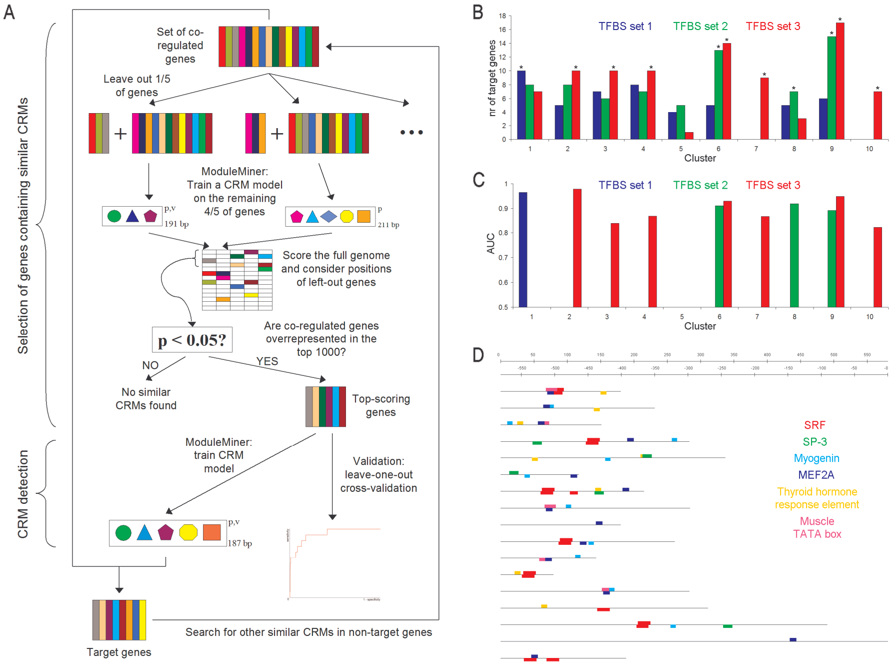


Figure 4. Application of ModuleMiner to microarray clusters. (A) The 2-step procedure used to detect similar *cis*-regulatory modules in a subset of genes within a given microarray cluster. In the first step, a five-fold cross-validation is performed, and the number of left-out genes considered as target genes is counted. If this number is significantly more than expected under a random distribution of the ranks, these genes are transferred to the second step. In this second step, ModuleMiner is used to model the similar *cis*-regulatory modules regulating the genes in this focused subcluster. (B) Results of the first step of the procedure in (A) for the 10 microarray clusters and the three different sets of candidate transcription factor binding sites. Significantly higher numbers of target genes among the left-out genes than randomly expected are depicted by an asterisk. Clusters 7 and 10 only contained sufficient genes (≥ 25) in transcription factor binding site set 3 and therefore are omitted for the other two sets. (C) Leave-one-out cross-validation results on the subclusters with a significant enrichment of target genes from (B). Each left-out gene was ranked using the TRGM obtained on the remaining genes. Next, sensitivity/specificity pairs where calculated for different detection thresholds, and these were used to construct ROC curves. The area under these ROC curves (AUC) was calculated and is depicted here. Colors: as in (B). (D) Example of a set of similar *cis*-regulatory modules identified by ModuleMiner. These results were obtained on the cardiac muscle genes by the procedure depicted in (A). Each horizontal line represents a human-mouse conserved non-coding sequence upstream of a gene within the cluster. The different colored boxes represent binding sites of different transcription factors. Detailed results, including descriptions of the genes shown, and the exact positions of the CNSs are available at [33].

(all AUCs were significantly above 50 %, with an average AUC of 90.3 %, Figure 4C). For the TRGMs obtained for clusters containing over 50 genes, the number of targets in the independent test set was determined. This was significantly higher than random in three of the five cases (Table 2). In total, we predicted 209 CRMs. These MODULEMINER predictions can be viewed in detail at our website.

Table 2. Summary of MODULEMINER's results for the 10 microarray clusters.

Cluster	Annotation	TFBS set	Nr target genes after cross-validation (p-val)	AUC on target genes	Nr target genes in independent test set (p-val)	Total nr of CRMs
1	Protein synthesis	1	10 / 50 (p = 0.025)	0.96	14 / 123 (p = 0.35)	30
2	Oocyte / fertilized egg	3	10 / 50 (p = 0.025)	0.98	30 / 164 ($p = 8.6 \times 10^{-4}$)	43
3	Neural tissues	3	10 / 50 (p = 0.025)	0.84	15 / 122 (p = 0.24)	29
4	Lymphocytes	3	10 / 50 (p = 0.025)	0.87	23 / 85 ($p = 7.0 \times 10^{-6}$)	36
5	Testis / spermatogenesis	-	-	-	-	-
6	Liver	3	14 / 50 ($p = 2.9 \times 10^{-4}$)	0.93	7 / 29 (p = 0.022)	23
7	Mitochondrion	3	9 / 31 (p = 0.0026)	0.87	-	12
8	Extracellular matrix	2	7 / 32 (p = 0.036)	0.92	-	10
9	Cardiac muscle	3	17 / 32 ($p = 6.6 \times 10^{-10}$)	0.95	-	16
10	Energy metabolism	3	7 / 26 (p = 0.012)	0.82	-	10

TFBS: transcription factor binding site. TFBS sets: set 1: human-mouse CNSs, 10 kb 5' of TSS; set 2: set 1 + binding site preservation; set 3: set 2 + correction for TSS differences. For clusters where multiple TFBS sets resulted in successful CRM detection, only the result showing the best cross-validation performance is shown. Genes (in the cluster) that by cross-validation where ranked within the top 10 % of the genome where considered target genes of the TRGM. The total number of CRMs constitutes all successful CRM predictions near genes in the cluster. CRM predictions were considered successful if the TRGM score was sufficient to rank the target gene within the top 10 % of the genome. In some cases, multiple CRMs are found controlling the same target gene.

Detection of *cis*-regulatory modules in embryonic development gene sets

In the previous section, we detected CRMs in microarray clusters expressed in different adult tissues. Next, we aimed to predict CRMs involved in embryonic development processes.

We constructed 5 gene sets involved in specific embryonic development processes, based on literature (Table 3). Contrary to the previous section, where we aimed to detect similar CRMs in a subset of the genes in the microarray clusters (using a two-step approach), here we can assume that the embryonic development gene set are more focussed, and hence we can directly apply MODULEMINER to these sets (as in our high quality smooth muscle gene set). We performed LOOCV, confirming that MODULEMINER was able to successfully detect similar CRMs in all five gene sets (Table 3).

Table 3. Summary of MODULEMINER's results for the 5 embryonic development gene sets.

Embryonic development process	TFBS set	Nr target genes after leave-one-out cross-validation (p-val)	AUC
Primary heart field [34]	1	6 / 7 ($p = 6.4 \times 10^{-6}$)	0.92
Secondary heart field [34]	1	6 / 9 ($p = 6.4 \times 10^{-5}$)	0.79
Neural crest cells [35]	2	6 / 10 ($p = 1.5 \times 10^{-4}$)	0.86
Eye development [36]	1	10 / 15 ($p = 1.9 \times 10^{-7}$)	0.79
Limb development [37]	1	10 / 24 ($p = 5.2 \times 10^{-5}$)	0.77

A key review or book used as a basis for construction of the development gene set is given in the first column. The genes in each set, as well as the detailed results can be viewed at our website [33]. TFBS: transcription factor binding site. TFBS sets: set 1: human-mouse CNSs, 10 kb 5' of TSS; set 2: set 1 + binding site preservation; set 3: set 2 + correction for TSS differences. For clusters where multiple TFBS sets resulted in successful CRM detection, only the result showing the best cross-validation performance is shown. Genes (in the cluster) that by cross-validation where ranked within the top 10 % of the genome where considered target genes of the TRGM.

Characterization of the *cis*-regulatory modules

The transcriptional regulatory global models that were predicted by ModuleMiner in each of the 10 microarray clusters and each of the 5 embryonic development gene sets are summarized in Tables 4 and 5. Apart from this TRGM, ModuleMiner also provides additional information characterizing the *cis*-regulatory modules. We will discuss here the results we obtained on cluster 9, which contains genes related to cardiac muscle function.

Table 4. Transcriptional regulatory global models constructed for the 10 microarray clusters.

Cluster	Key transcription factors and binding sites in TRGM (weight)
Protein synthesis	NF-Y (1.59), DEC (1.13), HIC1 (1.09), general initiator sequence (0.47), CCAAT box (0.44), TCF-4 (0.32)
Oocyte / fertilized egg	T3R (1.00), NF-Y (1.00), ETS/PEA3 (0.99), MAZ (0.92), AP2α (0.78), SP1 (0.30)
Neural tissues	UF1-H3β (1.13), CRE-BP/CJUN/ATF-1 (1.00), AP-2 (0.87), ETF (0.55), AP-1/NF-E2 (0.33)
Lymphocytes	STAT6 (1.00), PU.1 (0.99), ETS (0.96), STAT5/STAT (0.95), SP1 (0.89)
Testis / spermatogenesis	-
Liver	TCF1/HNF-1 (1.00), NF-1 (1.00), C/EBP (0.99), HNF-4/COUP (0.99), PPAR/HNF-4/COUP/RAR (0.66), MYC-MAX (0.58), PPAR (0.33)
Mitochondrion	c-ETS (1.35), VDR (1.00), GATA-1/GATA-2 (1.00), ZID (0.82), AR (0.43), ROAZ (0.34)
Extracellular matrix	AP-1/NF-E2/BACH1 (2.00), FOXD1 (1.00), BLIMP1 (1.00), SRF (0.70), MEF-2/RSRFC4 (0.51), STAT5/STAT6 (0.35)
Cardiac muscle	SP-3 (1.00), Myogenin (1.00), MEF2A (1.00), SRF (1.00), Tyroid hormone receptor/RAR/RXR (0.91), Muscle TATA box (0.48)
Energy metabolism	CREB/ATF/HLF (1.01), WHN (1.00), SPIB (0.71), PPARγ/RXRα (0.65), general initiator sequence (0.51), RFX (0.31)

Table 5. Transcriptional regulatory global models constructed for the 5 embryonic development sets.

Development process	Key transcription factors and binding sites in TRGM (weight)
Primary heart field	D type LTRs (1.12), HAND1/TCF3 (1.01), STAT3 (0.92), STAT5A (0.89), GATA1/GATA2 (0.63), ELK1 (0.32)
Secondary heart field	HNF3α (1.56), STAT5A/STAT5B (1.00), GATA2 (0.56), NFAT (0.56), GATA/GATA3 (0.48), WHN (0.35)
Neural crest cells	FREAC-7 (1.00), Poly A (1.00), TBX5 (1.00), HSF (0.89), FREAC-2 (0.30)
Eye development	RREB1 (1.00), IRF (0.96), POU3F2 (0.92), ZF5 (0.80), GATA/GATA1 (0.46), LMO2 (0.39), NKX6-1 (0.32)
Limb development	TEF (1.00), PLZF (1.00), PAX4 (0.96), EGR (0.87), AP-2 (0.65), PBX (0.63), Ikaros 1 (0.37)

First, ModuleMiner characterizes the given input genes, retrieving descriptions and commonly used identifiers (e.g. HGNC) from the Ensembl database. In addition, the Gene Ontology (GO) terms annotated to the input genes are retrieved, and the overrepresented GO terms are reported. For the cardiac muscle subcluster, “muscle contraction” (GO:0006936), “muscle development” (GO:0007517), “organogenesis” (GO:0009887), “contractile fibre” (GO:0043292) and “regulation of heart contraction rate” (GO:0008016) were among the overrepresented GO terms.

Next, ModuleMiner determines the weight of each PWM in the transcriptional regulatory global model (see Materials and methods). By grouping similar PWMs, the weight of each trans-factor involved is determined. The cardiac muscle TRGM contains PWMs for SRF, MEF2A, myogenin, SP3, a thyroid hormone response element (all with weights of approximately 1), and a muscle TATA box (with weight approximately 0.5). ModuleMiner also displays the *cis*-regulatory modules it identifies on the input genes. Figure 4D shows this for the heart muscle genes.

As our approach uses only human and mouse sequences to model *cis*-regulatory modules, sequenced genomes of other species can be used as validation data. MODULEMINER employs the rat and dog genomes for this purpose, by checking for *cis*-regulatory modules that fit the obtained TRGM in rat-dog conserved non-coding sequences. For the cardiac muscle genes, 11 orthologs were present in our rat-dog TFBS database, 7 of which were ranked within the top 10 % of the genome ($p = 2.28 \times 10^{-5}$).

Finally, MODULEMINER selects putative new target genes of the TRGM from the complete genome. We aim to minimize noise in these target gene predictions by using network level conservation [38], particularly through phylogenetic fusion of target gene rankings. To this end, first all genes in the human-mouse transcription factor binding site database (excluding the input genes), and all (non-input) genes in the dog-rat transcription factor binding site database are ranked separately. MODULEMINER then fuses these two rankings into one global ranking using order statistics (similar to the approach used in [23] and [39]). Among the 100 top ranking new target genes of the cardiac muscle TRGM were MYL3 (“Cardiac myosin light chain 1”), MYOD1 (“Myoblast determination protein 1”), TNNI1 (“Troponin I”) and MYH3 (“Myosin heavy chain, embryonic skeletal muscle”).

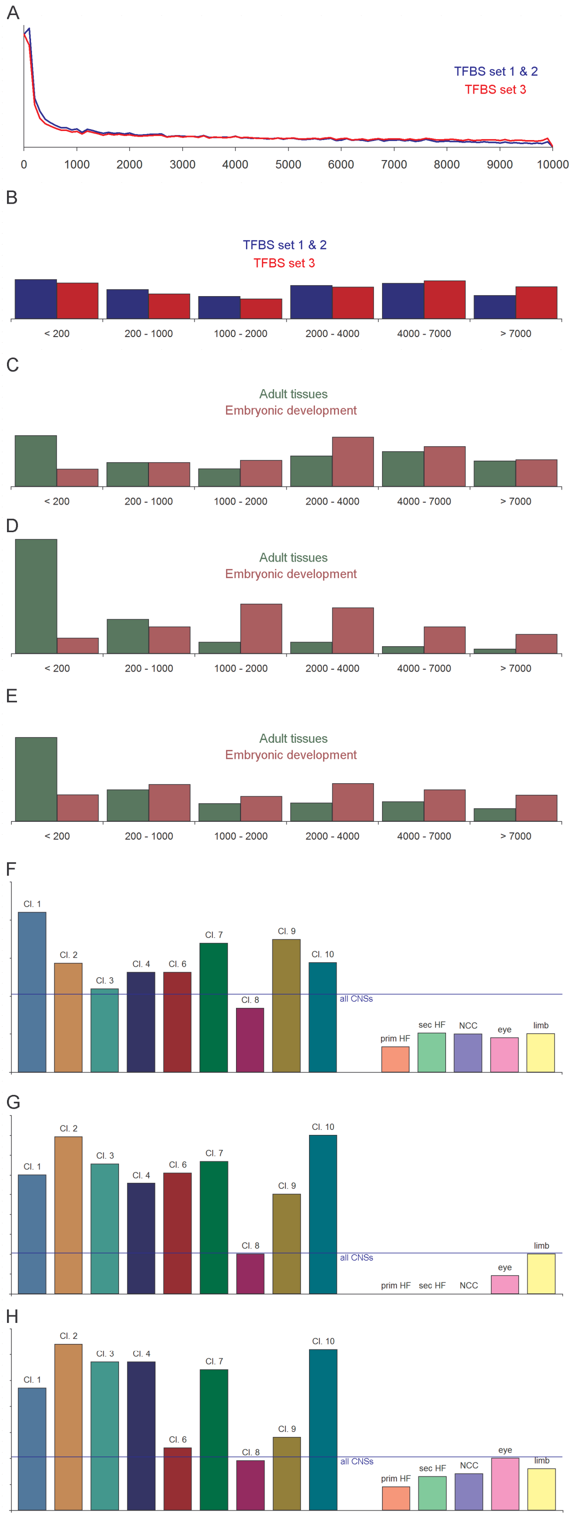
The results we obtained on all sets of co-expressed genes discussed in this work, can be viewed at [33].

Where are the *cis*-regulatory module predictions located?

MODULEMINER successfully detected 9 sets of similar CRMs in the 10 microarray clusters and 5 sets of similar CRMs in the 5 embryonic development gene sets. In total, 257 CRMs were predicted. In addition to this, MODULEMINER predicted 100 new target genes of each TRGM. We next used this compendium of 1657 CRMs to examine their positions relative to the TSS of the genes they regulate.

Since a gene's search space was defined as all CNSs within 10 kb 5' of the TSS, we first examined the distributions of CNS locations, as these represent the background distribution to which the CRM locations will be compared. A first important observation is that the CNSs are highly overrepresented close to the TSS, as shown in Figures 5A and 5B. The type of gene set, namely adult tissue versus embryonic development, introduces a second CNS location bias (Figure 5C). Indeed, the adult tissue CNS set is enriched in sequences close to the TSS (< 200 base pairs) ($p = 7.6 \times 10^{-16}$ by a Wilcoxon rank sum test), while the embryonic development CNS set is depleted in sequences close to the TSS and enriched in sequences further from the TSS (2000 – 4000 base pairs) ($p = 5.6 \times 10^{-7}$). When evaluating each of the gene sets separately (Figure 5F), 8 of the 9 adult tissue CNS sets are enriched in sequences less

Figure 5 (next page). Distribution of distance to transcription start site for CNSs and predicted *cis*-regulatory modules. (A) All human-mouse CNSs in TFBS sets 1 and 2 (both are based on the same set of CNSs), and in TFBS set 3. (B) The distribution from (A), when divided in 6 unequal bins. (C) Distribution of all conserved non-coding sequences upstream of genes within the microarray clusters (of genes expressed in different adult tissues) and the embryonic development gene sets, where CRMs could successfully be detected (Tables 2 and 3), divided in the same 6 bins as under (B). (D) Distribution of the distance to transcription start for the *cis*-regulatory modules MODULEMINER identified near the genes from (C). (E) Distribution of distance to transcription start for the *cis*-regulatory modules MODULEMINER identified in a whole genome scan (genes in (D) where removed, such that only new target genes where represented here). Note that (B), (C), (D) and (E) are drawn to the same scale. (F) Portion of CNSs near the genes in the different microarray clusters and embryonic development sets that is located within 200 bp of the transcription start site. (G) Portion of predicted CRMs near the genes in the different microarray clusters and embryonic development sets that is located within 200 bp of the transcription start site. (H) Portion of CRMs, predicted in a whole genome scan for the TRGM built for the different gene sets that is located within 200 bp of the transcription start site. The blue line in (F), (G) and (H) indicates the portion of all CNSs (within 10 kb 5' of all human genes) that is less then 200 bp of the transcription start site.


A
TFBS set 1 & 2
TFBS set 3
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
B
TFBS set 1 & 2
TFBS set 3
< 200
200 - 1000
1000 - 2000
2000 - 4000
4000 - 7000
> 7000
C
Adult tissues
Embryonic development
< 200
200 - 1000
1000 - 2000
2000 - 4000
4000 - 7000
> 7000
D
Adult tissues
Embryonic development
< 200
200 - 1000
1000 - 2000
2000 - 4000
4000 - 7000
> 7000
E
Adult tissues
Embryonic development
< 200
200 - 1000
1000 - 2000
2000 - 4000
4000 - 7000
> 7000
F
Cl. 1
Cl. 2
Cl. 3
Cl. 4
Cl. 6
Cl. 7
Cl. 8
Cl. 9
Cl. 10
all CNSs
prim HF
sec HF
NCC
eye
limb
G
Cl. 1
Cl. 2
Cl. 3
Cl. 4
Cl. 6
Cl. 7
Cl. 8
Cl. 9
Cl. 10
all CNSs
prim HF
sec HF
NCC
eye
limb
H
Cl. 1
Cl. 2
Cl. 3
Cl. 4
Cl. 6
Cl. 7
Cl. 8
Cl. 9
Cl. 10
all CNSs
prim HF
sec HF
NCC
eye
limb

than 200 base pairs from the TSS (in 6 cases, this was statistically significant by a Chi-square test), while all 5 embryonic development CNS sets are depleted in sequences less then 200 base pairs from the TSS (in 3 cases, this was statistically significant).

Next, we examine the location distribution of the CRMs that were identified by MODULEMINER. For adult tissue genes, CRMs are strongly overrepresented close to the TSS (Figure 5D). Sixty-three percent of these CRMs are within 200 base pairs of the TSS. In contrast, the CRMs MODULEMINER identified near the embryonic development genes are depleted close to the TSS and enriched further away (1000 – 2000 base pairs). These conclusions remain valid even when controlling for both biases mentioned above: comparing Figure 5D to Figure 5C (the predicted CRMs in Figure 5D can be considered as a selection from the CNS sets in Figure 5C), the enrichment of predicted CRMs directing expression in adult tissues close to the TSS persisted: $p = 2.6 \times 10^{-27}$ (calculated as follows: the distances to the TSS of (i) the predicted CRMs and (ii) all CNSs of the genes in the microarray clusters were ranked and the Wilcoxon rank sum test was applied). For the CRMs directing expression in embryonic development, no statistically significant deviation from random selection from the embryonic development CNS sets could be observed ($p = 0.18$). When considering the gene sets separately, in 8 microarray clusters expressed in adult tissues, CRMs are enriched in sequences close to the TSS (Figure 5G) (this was statistically significant when controlling for bias in 6 cases). In contrast, in 4 embryonic development gene sets, CRMs are depleted close to the TSS (markedly, for three of these sets, no CRMs were predicted within 200 base pairs of the TSS).

A similar difference in TSS distance distribution was also seen for the new target genes (Figure 5E). Here as well, the distances to the TSS of the CRMs predicted to direct expression in adult tissues were clearly non-randomly distributed compared to all CNSs ($p = 3.6 \times 10^{-74}$ by a Wilcoxon rank sum test). For the CRMs predicted to direct expression in embryonic development, no statistically significant difference was observed (by a Wilcoxon rank sum test). However, these sequences seem to be (slightly) depleted within 200 base pairs of the TSS ($p = 1.5 \times 10^{-4}$ by a Chi-square test). Considering each of the gene sets separately (Figure 5H), in 7 adult tissue microarray clusters, CRMs were significantly enriched within 200 base pairs of the TSS, while for 2 embryonic development gene sets, CRMs were significantly depleted close to the TSS. Although in six cases this effect was highly significant ($p < 10^{-9}$), it was smaller than the effect within the clusters (compare Figures 5D and 5E).

In summary, the *cis*-regulatory modules MODULEMINER detected were non-randomly positioned in the genome. CRMs predicted to direct expression in adult tissues were highly enriched very close to the transcription start site, while CRMs predicted to direct expression in embryonic development were depleted very close to the transcription start site.

Discussion

Although the sequence of the human genome has been available for a considerable time now, our ability to chart the regions controlling gene expression is still very limited. The situation seems to improve as a function of smaller genome size. Indeed, in the *Drosophila* early segmentation network, CRMs can be predicted based on known examples [10,11]. In the yeast *Saccharomyces cerevisiae*, with an even much smaller genome, it is possible to go one step further and predict the expression of genes based only on upstream sequences [40]. Here we focus on the computational detection of CRMs in the human genome, and hence this work is a contribution in bridging this gap.

MODULEMINER detects CRMs by taking as input a set of co-expressed genes, under the assumption that a subset of these are co-regulated, and looking for a recurrent pattern of (computationally predicted) transcription factor binding sites. The advantages of this approach are that it does not require known examples and that it allows prediction of a probable function for the detected CRMs.

MODULEMINER is similar in scope to ModuleSearcher [20,28] and CREME [19]. It differs from these previous approaches in that MODULEMINER maximizes specificity for the given set of co-expressed genes by performing a whole genome optimization. Indeed, MODULEMINER optimizes the combined rankings of the given gene set in a ranking of the complete genome. In addition, this approach allows comparison between TRMs with different parameters (e.g. maximum CRM length, number of PWMs in the TRM). Therefore, MODULEMINER is able to optimize over these parameters, and hence, our approach effectively eliminates the parameters required by the previous approaches.

Other algorithms have been developed that aim to detect similar CRMs in a set of co-expressed genes that (contrary to the approaches above) do not use a library of PWMs [21,22,29,41]. Instead, these algorithms optimize, besides the combination of motifs, also the motifs themselves. Hence, these methods attempt to solve a problem with considerably higher complexity, resulting in lower performance, as confirmed by our comparison on benchmark data. Given the extremely poor performance of motif detection methods in other organisms than yeast [42], we have opted to circumvent motif optimization by using experimentally determined PWMs. Note that this decision not necessarily limits the search to known PWMs, as libraries of computationally predicted PWMs are also available (e.g. the phylofacts PWM library [43]). In addition, we believe that with the emergence of the protein binding microarray technology [44], high quality PWMs will soon become available for a large fraction of the human transcription factor repertoire. Even though the currently available libraries of experimental PWMs show high redundancy and may contain low quality PWMs, our new approach of clustering similar TRMs is able to group redundant PWMs and our validations show that in many cases a combination of five experimental PWMs can capture enough information of a CRM to yield acceptable genome-wide specificity levels.

MODULEMINER outputs the predicted CRMs, and a transcriptional regulatory global model (TRGM). This TRGM can be considered as a bag of PWMs (selected from TRANSFAC and JASPAR), with a weight associated to each PWM. Therefore, this TRGM not only predicts the transcription factors functioning in the process under study, but in addition also allows an assessment of the relative importance of each of these transcription factors.

TRGMs do not contain spatial relations between transcription factor binding sites (except for the total size of the CRMs and a Boolean parameter indicating whether different binding sites can overlap or not). Although certain spatial relations between transcription factors working in concert are known to exist (e.g. [45,46]), we did not find any reports indicating that this is the rule rather then the exception. Therefore, we reasoned that any such relationships should not be hard-coded in the TRGMs, but rather would become apparent by inspection of the predicted CRMs. Upon inspection of the predicted CRMs presented above, no such spatial relationships surfaced.

Our method for scoring a sequence using a TRM or TRGM (see Materials and methods) does not take homotypic clustering of transcription factor binding sites into account (like HMM based methods do [15,17,47]). However, this cooperative binding of one transcription factor can nevertheless be modelled in our framework by the construction of a TRM or TRGM that contains multiple instances of the same PWM. Therefore, if multiple instances of a specific transcription factor are important for the regulation of a set of co-regulated genes, this is represented accordingly in the optimal model. For example, when applying MODULEMINER to the tightly co-expressed set of smooth muscle markers, the transcription factor SRF occurs 2 or 3 times in each of the TRMs in the resulting TRGM, suggesting an extensive cooperation between SRF binding sites for smooth muscle specific transcription regulation. In contrast, the SMAD4, SP1 and ATF3 PWMs occur exactly once in 97.5 % of the TRMs (SMAD4 and SP1 occur twice in 1.5 % and 1 % of the TRMs respectively).

MODULEMINER takes the genomic background sequence into account in two ways. Firstly, a third order background model is used in the process of annotating putative transcription factor binding sites. Secondly, our optimization strategy selects the TRM (or TRGM) that optimally separates the given genes (sequences) from all other genes in the genome. Hence, our system corrects both for local sequence properties (by the third order background model) as for more global sequence properties (by selecting against combinations of transcription factor binding sites that occur independently of the given sequences).

We included all CNSs up to 10 kb 5' of the transcription start site in our pipeline. Although this choice is inherently arbitrary, it is motivated by the following arguments: (i) sequences 3' of the transcription start site might harbour translational regulatory signals, which we do not want to model here; (ii) potential regulatory sequences far upstream can be difficult to assign to a target gene; (iii) selecting 10 kb 5' of the transcription start site has proven to be valuable in our previous study [20], and others have made similar choices as well [48]; (iv) in a previous study where CRMs were predicted in an unbiased way across the complete human genome [8], it was shown that CRMs are highly depleted between 10 kb and 30 kb 5' of the transcription start site.

The validation framework we use, combining genome-wide ranking with leave-one-out cross-validation, could also be useful in evaluating or comparing hypotheses regarding the working principles of transcription regulation, and in this regard can be considered similar in scope to CodeFinder [24]. In this work, two such tests are implicitly performed: (i) CRMs driving a tissue-specific expression pattern are compared to CRMs driving an embryonic development expression pattern and (ii) by the comparison of the three sets of putative transcription factor binding sites (e.g. Figure 1, Figure 3J, Figure 4B), the importance of binding site preservation is evaluated as well as the impact of a correction for differences in transcription start sites between human and mouse.

Construction of a high-quality set of co-regulated genes involved in a certain process under study is not always straightforward. In this regard, robustness to noise in a set of putative co-expressed genes is highly desirable in an algorithm to detect similar CRMs. We found MODULEMINER to be highly robust to the quality of this input gene set. Indeed, in our experiments with smooth muscle marker genes, we observed that even when only 10 of 50 given genes are really co-regulated, MODULEMINER was still able to pick up the correct signal (Figure 2). These properties of MODULEMINER prompted us to apply the algorithm to gene sets obtained from clustering microarray data. In 9 out of 10 microarray clusters, MODULEMINER succeeded in finding similar CRMs in a subset of the genes. Perhaps unsurprisingly, a critical mass of co-regulated genes is required for MODULEMINER to detect similar CRMs. However, this minimum required number of co-regulated genes is sufficiently small so as not to preclude application of the algorithm. This is illustrated both by our results obtained on the smooth muscle genes (Figure 2), and by the successful CRM detection in two small heart development gene sets (Table 3).

Application of MODULEMINER to the smooth muscle marker genes resulted in CRMs with multiple binding sites for SRF, and with single binding sites for SMAD4, SP1 and ATF3. Both SRF and SP1 have been shown to play a role in regulating smooth muscle specific expression [26]. Furthermore, SMADs are effectors of the TGF-β signalling pathway, and have been shown to work in concert with SRF to control smooth muscle cell differentiation [49]. MODULEMINER identified transcription factors known to play a key role in other co-expressed gene sets as well. Examples are GATA-factors, NFATs and HAND1 in heart development, HNF-1 and HNF-4 in liver-specific gene expression, PU.1 in lymphocyte specific gene expression, and

Myogenin, SRF, the thyroid hormone receptor, and MEF2 in heart specific gene expression.

Imposing trans-factor conservation by motif preservation between human and mouse sequences of a CNS significantly improved the performance of MODULEMINER on the set of smooth muscle marker genes. A similar approach has also been shown to improve CRM detection performance in the *Drosophila* early segmentation gene network [10]. When we applied MODULEMINER to the microarray clusters and the embryonic development gene sets, in some cases this trans-factor conservation also increased performance (microarray clusters 6, 7 and 9, and the neural crest cell gene set), but in other cases it did not.

Correcting for possible differences in transcription start site in human and mouse by a three-step alignment procedure (see Materials and methods), resulted in increased performance for most of the microarray clusters, but not for the development gene sets. This marked difference may be related to the different locations of the detected CRMs in these two different systems.

We observed a significant difference in the locations of the CRMs MODULEMINER predicted to direct expression in adult tissues and the CRMs MODULEMINER predicted to direct expression in embryonic development. CRMs driving tissue-specific expression are highly overrepresented within 200 base pairs of the TSS. In contrast, CRMs driving expression in embryonic development are more evenly distributed in the 10 kb sequences we considered, and seem to be underrepresented within 200 base pairs of the TSS. These results suggest that transcription regulation of tissue-specific expression is mainly exerted by proximal promoters, while transcription regulation of expression during embryonic development seems to be mainly exerted by more distal enhancers.

MODULEMINER can be applied to 3 conceptually different tasks: (i) prediction of transcription factors that play a role in regulating a set of co-regulated genes, (ii) prediction of regulatory regions and (iii) predictions of new target genes of a TRGM. It is important to realize that the accuracy of the predictions differs between those tasks. Although exact performance statistics can only be obtained through the careful experimental testing of our predictions, which is outside the scope of the present study, the results we obtained in this work can be used to provide rough estimates of the predictive accuracy. When we applied MODULEMINER to the two well-studied benchmark sets, we obtained HNF1, CEBP, HNF3, GATA1, PAX6 and HNF4 for the liver benchmark set, and MZF1, PPARγ, SRF, MEF2, the Epstein-Barr virus transcription factor R, MYF and MYOD for the muscle benchmark set. Comparing this to literature [4,50] and to the PWM libraries we use, we obtain a sensitivity of 70 % (7 out of 10 known PWMs are recovered), a specificity of 99.6 % (630 of 633 (liver) and 619 of 621 (muscle) likely incorrect PWMs are rejected) and a positive predictive power of 62 % (8 of 13 total predicted PWMs are correct). These values need to be regarded with some reservations when extrapolating to other cases, since both liver and muscle are well-studied systems with high quality PWMs available. Nevertheless, we can conclude that MODULEMINER is quite accurate in selecting PWMs/transcription factors that play a key role in the regulation of the genes under study. Regarding detection of regulatory sequences, MODULEMINER was able to detect 16 of 24 known muscle/liver enhancers, when a total of 24 predictions where made. This is a sensitivity of 67 % and a positive predictive power of 67 %, although we emphasize that this last value is an underestimate as some of our predictions may be yet unknown enhancers. Notwithstanding some reservations on extrapolating these data, we conclude that the predictive accuracy of MODULEMINER for detection of regulatory regions (CRMs) near a set of co-regulated genes is quite high. Regarding the predictive accuracy of MODULEMINER for the detection of new target genes given a TRGM, the results of our LOOCV procedure can provide some estimates. From the resulting ROC curves, one can see that for a sensitivity of 50 %, the specificity is about 90 %, and for a sensitivity of 80 %, specificity is about 80 %, although the differences between different gene sets can be large. However, typically only a few dozen new target genes can be tested, and thus specificity may not be high enough to select the right targets from the complete genome. In our previous study [23], we confirmed that the predictive accuracy of new target genes is quite low, although we showed it to be detectably present. We note that in that study, we used our previous ModuleSearcher algorithm which was shown here to have a lower performance than MODULEMINER. In addition, MODULEMINER's use of network level conservation between human/mouse and rat/dog predictions of new target genes might increase performance. Finally, the results we obtained in the transcription start site distribution of the CRMs predicted near the new target genes are consistent with these performance predictions: Figures 5E and 5H show a similar trend as Figures 5D and 5G, but to a lesser extend, hence pointing to a substantial amount of noise, but also indicating that a signal can be picked up, even in a whole genome scan.

Conclusions

We present MODULEMINER, the first algorithm to detect similar *cis*-regulatory modules in the human genome that is based on whole-genome optimization. MODULEMINER is generally applicable, and outperforms other similar approaches to detect CRMs on benchmark data. In addition, MODULEMINER can detect similar CRMs in noisy sets of co-expressed genes, such as microarray clusters. We successfully applied the algorithm to sets of genes expressed in adult tissues and sets of genes expressed in embryonic development processes. We show that CRMs predicted to regulate genes expressed in adult tissues are highly overrepresented within 200 base pairs of the transcription start site, while CRMs predicted to regulate genes involved in embryonic development processes are depleted within this region. These findings suggest that expression in adult tissues is mainly directed by proximal promoters, while expression in embryonic development is more often regulated by distal enhancers.

Materials and methods

Construction of 3 sets of candidate transcription factor binding sites

We constructed three sets of genome-wide candidate transcription factor binding sites in human-mouse conserved non-coding sequences (CNSs). The first set contains all predicted binding sites in all CNSs. Sequences 10 kb 5' (+ 50 bp 3') of the transcription start site of all human genes and their mouse orthologs were obtained from Ensembl (version 36). When another gene was encountered, only the sequence up to that gene was included. Conserved non-coding sequences were selected by LAGAN alignments [51]. Thresholds were set to 75 % conservation over at least 100 base pairs. Transcription factor binding site predictions were performed using MotifScanner [52], with the prior set to 0.2. Both TRANSFAC [53] (version 9.4) and JASPAR [43] were used as PWM libraries.

The second set aims to restrict the candidate binding sites by enforcing that the regulatory factors should be conserved. This is achieved by selecting only binding sites in each human region for transcription factors for which we also detect binding sites in the orthologous mouse region (preserved sites). We note that this constraint does not impose that the binding sites should be conserved, nor that they should align.

In the construction of the third set we aimed to correct for differences in human and mouse transcription start sites (TSSs), and for possible annotation errors of TSSs. To this end, we extended the mouse sequences used in the alignments by 100 kb in both directions. Alignment errors were kept in check by applying a multi-step alignment procedure. The human 10 kb sequence was aligned to (A) the 10 kb mouse sequence, (B) the mouse sequence extended by 10 kb in both directions, and (C) the mouse sequence extended by 100 kb in both directions. If in alignment (A) CNSs were predicted, we assumed that the correct orthologous region in the mouse is not off by more then 10 kb, and hence we used the CNSs from alignment (A), supplemented by all additional CNSs from alignment (B). CNSs that were truncated in alignment (A) because they extended over the sequence borders, were replaced by their counterpart from alignment (B). If in alignment (A) no CNSs were predicted, we reasoned that the correct orthologous region in the mouse might be off by more then 10 kb, and we used the CNSs from alignment (C). Here also, for each CNS (in human), we selected only preserved binding sites.

The same procedure was used with the dog and rat sequences to create sets of candidate transcription factor binding sites corresponding to the three human-mouse sets. As neither dog nor rat could serve as a reference species, we did not extend the sequences in the dog-rat candidate transcription factor binding site set that corresponds to human-mouse set 3.

Transcriptional regulatory models

We model similar *cis*-regulatory modules in a set of co-expressed genes by transcriptional regulatory models. These TRMs are parameterized as in [20]. A TRM is a combination of PWM instances (up to 6), supplemented by three parameters: (i) the maximum length of *cis*-regulatory modules, (ii) a Boolean parameter stating whether different binding sites can overlap or not, and (iii) a Boolean parameter that indicates whether incomplete modules will be penalized or not. Given a TRM and a sequence, a score S_{seq} can be calculated, as detailed in [20]. A TRM may contain multiple instances of one specific PWM: in the calculation of S_{seq}, each PWM in the TRM is matched to at most one binding site – thus if a PWM occurs twice, up to two binding sites for the corresponding transcription factor can be taken into account. We assign a score S_g to a gene by taking the maximum of S_{seq} for all CNSs of that gene.

The S_g scores for the given set of co-regulated genes are used to determine a 'fitness score' of a TRM. This fitness score of a TRM for a given set of co-expressed genes is determined by the positions of the co-expressed genes in a ranking of S_g for all genes in the genome. We use order statistics to assign a probability to the combination of ranks of the given co-expressed genes (using the numerical approach detailed in [23]). Hence, the resulting p-value represents how well that TRM models the given set of co-expressed genes, compared to all other genes in the genome. We use 1 minus that p-value as the fitness score for the TRM.

The MODULEMINER algorithm

MODULEMINER uses a genetic algorithm to find the TRM with the optimal fitness score. At the onset, a starting population of TRMs is obtained by running our ModuleSearcher algorithm [28] using many different combinations of parameters. This initial step is not absolutely required (one can start from a population of randomly generated CRMs), but it provides a speed advantage. These TRMs obtained by ModuleSearcher are assigned a fitness score, and the 200 best scoring TRMs are retained as starting population for the ModuleMiner genetic algorithm. During each 'generation' of the algorithm, 200 new individuals (TRMs) are generated (based on the TRM population at that time) and added to the population. This population of 400 TRM is then required to compete (by fitness score), and the 200 best scoring TRMs are retained. This procedure is repeated until the stop criterion is reached (at least 300 generations and at most 1000 generations). Generation of new individuals (TRMs) is done using 2 'parent' TRMs randomly selected from the population. Each of the TRM parameters (number of PWMs, length, overlap and penalization) is determined by random selection from both parents, allowing a small probability of mutation (i.e. each parameter is set to a random value with a probability of 0.1). Subsequently, PWMs are selected at random from both parents. Here as well, each PWM can be ‘mutated’ (replaced by a PWM randomly selected from TRANSFAC and JASPAR) with a probability of 0.1. As stop criterion, we use homogeneity of the population: if more than 80 % of the TRMs can be grouped into one TRGM (see below) and at least 300 generations have passed, the algorithm is stopped. If this stop criterion is not reached, the algorithm is stopped after 1000 generations. The parameters of the ModuleMiner genetic algorithm (e.g. population size, mutation probability, …) were selected by optimizing for speed. The convergence of the algorithm is highly insensitive to these parameters over a wide range, and sensitivity of speed to these parameter settings is also limited (data not shown).

Transcriptional regulatory global models

Aiming to minimize the sensitivity of our models of similar *cis*-regulatory modules to noise in transcription factor binding site predictions, we constructed composite models (TRGMs) from multiple high-scoring TRMs. To this end, similar TRMs are clustered, and the largest cluster is returned as resulting TRGM. TRMs were clustered when the *cis*-regulatory modules they predict near the high scoring genes (out of the given set of co-expressed genes) occur in the same CNS. As a cut-off for determining which genes are among the “high scoring genes”, we used the top 2.5 % in a ranking of the complete genome.

Scoring a sequence with a TRGM is performed by scoring this sequence for each TRM within the TRGM, subsequently normalizing this score (maximum CNS score = 1), and finally adding the normalized TRM scores.

As a TRGM is a collection of TRMs and TRMs each contain a collection of PWM instances, TRGMs are also collections of PWMs. In addition, a weight can be assigned to each PWM in the TRGM, quantifying the significance of the PWM for the process under study. This weight of a PWM is calculated as follows: for each TRM in the TRGM, the number of instances of that PWM is counted, and this number is averaged over all the TRMs in the TRGM.

Performance comparison on benchmark data

Four benchmark data sets containing annotated regulatory regions directing expression in a particular system were selected from PAZAR [27]. We selected all human genes (or human orthologs) from each of these 'boutiques'. The regulatory sequence search space was defined as all CNSs within 10 kb 5' of the TSS (as throughout our study). We used this search space for all algorithms, except CREME [19], where only the online version was available that by default uses one CNS within 1.5 kb of the TSS. As the other CRM detection algorithms had multiple parameters (absent in MODULEMINER), these parameters were set to default options. For the ModuleSearcher algorithm [28], we used the same parameters as in the cell cycle case study reported [20]. For CisModule [22] and EMCMODULE [29] we used the default parameter settings. We used Clover [30] as follows: for each PWM found overrepresented, we constructed a TRM (with parameters: no overlap between binding sites, no penalization and a maximum distance of 1000 bp), and this way we constructed TRGMs containing enriched PWMs reported by Clover. We also generated 100 random TRMs (combinations of 3-6 PWMs with randomly generated parameters) and we used these to rank the genes of each benchmark set, as a proxy for a method unable to detect similar CRMs.

Availability

MODULEMINER can be accessed at our website [33]. A stand-alone version is available upon request.

Acknowledgements

PVL is supported by a PhD fellowship and SA by a postdoc fellowship of the Research Foundation – Flanders (FWO). BT is supported by a PhD fellowship of the IWT. Research performed by YM and BDM is supported by: (i) Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants; (ii) Flemish Government: FWO: PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

This research was conducted utilizing high performance computational resources provided by the University of Leuven.

We are grateful to Irina Balikova, Boyan Dimitrov and Koen Devriendt for help with the construction of the development gene sets. The authors declare no competing financial interests.

References

1. Davidson EH: *Genomic Regulatory Systems: Development and Evolution*. San Diego, USA: Academic Press; 2001.

2. Balmer JE, Blomhoff R: **Anecdotes, data and regulatory modules.** *Biol Lett* 2006, **2:** 431-434.

3. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5:** 276-287.

4. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:** 167-181.

5. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Res* 2003, **13:** 64-72.

6. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:** 1034-1050.

7. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J: **Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity.** *Cell* 2006, **124:** 47-59.

8. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006, **16:** 656-668.

9. Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH: **Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin.** *Dev Biol* 2002, **246:** 148-161.

10. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.** *Genome Biol* 2004, **5:** R61.

11. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U: **Transcriptional control in the segmentation gene network of Drosophila.** *PLoS Biol* 2004, **2:** E271.

12. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers.** *Science* 2003, **302:** 413.

13. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3:** e7.

14. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De VS, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:** 499-502.

15. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.** *BMC Bioinformatics* 2002, **3:** 30.

16. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99:** 757-762.

17. Sinha S, van NE, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19 Suppl 1:** i292-i301.

18. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:** 1019-1028.

19. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics* 2003, **19 Suppl 1:** i283-i291.

20. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of *cis*-regulatory modules.** *Bioinformatics* 2003, **19 Suppl 2:** II5-II14.

21. Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE: **Decoding human regulatory circuits.** *Genome Res* 2004, **14:** 1967-1974.

22. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.** *Proc Natl Acad Sci U S A* 2004, **101:** 12114-12119.

23. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion.** *Nat Biotechnol* 2006, **24:** 537-544.

24. Philippakis AA, Busser BW, Gisselbrecht SS, He FS, Estrada B, Michelson AM, Bulyk ML: **Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells.** *PLoS Comput Biol* 2006, **2:** e53.

25. Nelander S, Mostad P, Lindahl P: **Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes.** *Genome Res* 2003, **13:** 1838-1854.

26. Kumar MS, Owens GK: **Combinatorial control of smooth muscle-specific gene expression.** *Arterioscler Thromb Vasc Biol* 2003, **23:** 737-747.

27. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW: **PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation.** *Genome Biol* 2007, **8:** R207.

28. Aerts S, Van Loo P, Moreau Y, De Moor B: **A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes.** *Bioinformatics* 2004, **20:** 1974-1976.

29. Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes.** *Proc Natl Acad Sci U S A* 2005, **102:** 7079-7084.

30. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation.** *Nucleic Acids Res* 2004, **32:** 1372-1381.

31. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:** 6062-6067.

32. Nelander S, Larsson E, Kristiansson E, Mansson R, Nerman O, Sigvardsson M, Mostad P, Lindahl P: **Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals.** *BMC Genomics* 2005, **6:** 68.

33. ModuleMiner: computational detection of *cis*-regulatory modules. http://www.esat.kuleuven.be/moduleminer

34. Buckingham M, Meilhac S, Zaffran S: **Building the mammalian heart from two sources of myocardial cells.** *Nat Rev Genet* 2005, **6:** 826-835.

35. Stoller JZ, Epstein JA: **Cardiac neural crest.** *Semin Cell Dev Biol* 2005, **16:** 704-715.

36. Graw J: **The genetic and molecular basis of congenital eye defects.** *Nat Rev Genet* 2003, **4:** 876-888.

37. Epstein CJ, Erickson BP, Wynshaw-Boris A: *Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis.* New York, USA: Oxford University Press; 2004.

38. Pritsker M, Liu YC, Beer MA, Tavazoie S: **Whole-genome discovery of transcription factor binding sites by network-level conservation.** *Genome Res* 2004, **14:** 99-108.

39. Aerts S, van HJ, Sand O, Hassan BA: **Fine-Tuning Enhancer Models to Predict Transcriptional Targets across Multiple Genomes.** *PLoS ONE* 2007, **2:** e1115.

40. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:** 185-198.

41. Grad YH, Roth FP, Halfon MS, Church GM: **Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura.** *Bioinformatics* 2004, **20:** 2738-2750.

42. Tompa M, Li N, Bailey TL, Church GM, De MB, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van HJ, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:** 137-144.

43. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van RF, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34:** D95-D97.

44. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML: **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.** *Nat Genet* 2004, **36:** 1331-1339.

45. Fickett JW: **Coordinate positioning of MEF2 and myogenin binding sites.** *Gene* 1996, **172:** GC19-GC32.

46. Papatsenko D, Levine M: **A rationale for the enhanceosome and other evolutionarily constrained enhancers.** *Curr Biol* 2007, **17:** R955-R957.

47. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31:** 3666-3668.

48. Chang LW, Nagarajan R, Magee JA, Milbrandt J, Stormo GD: **A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles.** *Genome Res* 2006, **16:** 405-413.

49. Nishimura G, Manabe I, Tsushima K, Fujiu K, Oishi Y, Imai Y, Maemura K, Miyagishi M, Higashi Y, Kondoh H, Nagai R: **DeltaEF1 mediates TGF-beta signaling in vascular smooth muscle cell differentiation.** *Dev Cell* 2006, **11:** 93-104.

50. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11:** 1559-1566.

51. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13:** 721-731.

52. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes.** *Nucleic Acids Res* 2003, **31:** 1753-1764.

53. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:** 374-378.

Study 3

T cell/histiocyte rich large B cell lymphoma shows transcriptional features suggestive of a tolerogenic host immune response

Contribution of the doctorandus: design of the study, analysis of gene expression data and development and application (including experiments) of the quantitative PCR classifier.
This study is in revision at leukemia (revised version submitted).

T cell/histiocyte rich large B cell lymphoma shows transcriptional features suggestive of a tolerogenic host immune response

Peter Van Loo[1,2,3], Vera Vanhentenrijk[4], Daan Dierickx[5], Agnieszka Malecka[6], Isabelle Vanden Bempt[4], Gregor Verhoef[5], Jan Delabie[4,6], Peter Marynen[1,2], Patrick Matthys[7,#,*], Chris De Wolf-Peeters[4,#,*]

[1]Department of Molecular and Developmental Genetics, VIB; [2]Department of Human Genetics, K.U.Leuven; [3]Bioinformatics group, Department of Electrical Engineering, K.U.Leuven; [4]Department of Pathology, University Hospitals K.U.Leuven; [5]Department of Hematology, University Hospitals K.U.Leuven; [6]Department of Pathology, The Norwegian Radium Hospital, University of Oslo; [7]Department of Microbiology and Immunology, Rega Institute for medical research, K.U.Leuven; [#]PM and CDW-P share senior authorship; [*]To whom correspondence may be addressed, at Rega Institute for medical research, K.U.Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium, Email: Patrick.Matthys@rega.kuleuven.be, Tel: +32 16 337370, Fax: +32 16 337340 (PM) and Department of Pathology, University Hospitals K.U.Leuven, Minderbroedersstraat 12, B-3000 Leuven, Belgium, Email: Christiane.Peeters@uz.kuleuven.be, Tel: +32 16 336582, Fax: +32 16 336548 (CDW-P).

Running title: A tumour tolerogenic microenvironment in THRLBCL

This work was supported by a Concerted Research Action Grant from the K.U.Leuven (PeterM) and grant G.05.28.06 from the Research Foundation – Flanders (FWO) (CDW-P).

Abstract

Gene expression profiling has successfully identified the prognostic significance of the host response in lymphomas. We endeavour to obtain a better understanding of this host response, comparing two B cell lymphomas characterized by a paucity of tumour cells embedded in an overwhelming background: the aggressive T cell/histiocyte rich large B cell lymphoma (THRLBCL) and the indolent nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL). The tumour cells of both lymphomas share several characteristics, while the cellular composition of their microenvironment is clearly different. Aiming to study this microenvironment, which constitutes the majority of the tumour cell mass in both THRLBCL and NLPHL, we performed microarray expression profiling on entire tissue sections. We observed that the NLPHL microenvironment is molecularly very similar to a lymph node characterized by follicular hyperplasia, while the THRLBCL microenvironment is clearly different. The THRLBCL signature is hallmarked by up-regulation of CCL8, IFN-γ, IDO, VSIG4 and Toll-like receptors. These features may be responsible for the recruitment and activation of T cells, macrophages and dendritic cells, characterizing the stromal component of this lymphoma, and may point towards innate immunity and a tumour tolerogenic immune response in THRLBCL.

Keywords: T cell/histiocyte rich large B cell lymphoma, Hodgkin's lymphoma, microenvironment, expression profiling

Introduction

B cell lymphomas with a high content of T cells, occasionally misinterpreted as T cell lymphomas in the past, have been recognized as a caveat for pathologists and were therefore indicated as “T cell rich B cell lymphoma” (1). Initial studies demonstrated that a particular subgroup of T cell rich B cell lymphomas may mirror nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL) and are characterized by a histiocyte-rich stroma (2, 3). These lymphomas carry a distinct clinical behaviour and a bad prognosis (4). In the WHO classification of 2001, this T cell/histiocyte rich large B cell lymphoma (THRLBCL) is listed as a variant of diffuse large B cell lymphoma (DLBCL) and is defined by the presence of scattered large B cells in a background rich in T cells, together with or without histiocytes (5). The precise relationship between THRLBCL and other lymphomas, more particularly NLPHL, remains unclear (2, 3, 6). Indeed, the atypical B cells of NLPHL and of THRLBCL share many characteristics, including expression of pan B cell markers, germinal centre B cell origin, and common chromosomal imbalances (7-10). An important difference between both lymphomas lies in their clinical presentation and prognosis. THRLBCL is a very aggressive disorder, mostly not responding to therapy (11). These patients frequently present with stage III and IV disease, splenomegaly, hepatomegaly and bone marrow involvement. In contrast, NLPHL is an indolent disorder. Most patients present at an early stage of disease and carry a good prognosis (12).

Gene expression profiling of lymphomas clearly illustrated that aside from the characteristics of the tumour cells, the microenvironment of the tumour also defines the profile of the lymphoma, and more importantly plays a role in predicting the prognosis (13, 14). Here, we aimed to gain insight into the role of the microenvironment of THRLBCL, by comparing its expression profile with that of NLPHL and a pool of reactive lymph nodes with follicular hyperplasia.

Materials and Methods

Patients

A series of 98 cases were retrieved from the files of the department of pathology of the University Hospitals of K.U.Leuven, all documented by frozen material. The series includes all cases recorded over the last 25 years (i) as NLPHL or lymphocyte rich classical Hodgkin's lymphoma (LRCHL) or (ii) as THRLBCL or DLBCL with a prominent histiocyte and T cell rich stromal component. As an additional and external control series, 26 cases recorded as THRLBCL, NLPHL or LRCHL at the department of pathology of the Rikshospitalet-Radiumhospitalet HF Oslo were added to the study material. Upon review, the diagnosis of THRLBCL and NLPHL was confirmed in 34 and 57 cases respectively. In all these cases, the atypical cells represented less than 10 % of the tumour mass. Thirty-one cases were excluded from the study because the frozen material was not representative or because material for review or additional immunostainings were not available, or because upon review they were diagnosed as DLBCL or LRCHL. Finally, 2 cases were excluded from the study because an unambiguous diagnosis of NLPHL or THRLBCL could not be made; their morphology fulfilled that of the cases described by Boudova *et al*. (15).

World Health Organization (WHO) criteria (version of 2008) were applied to assign cases to the different categories (16).

From the series diagnosed at the University Hospitals of K.U.Leuven, we randomly selected 10 typical NLPHL and 10 typical THRLBCL cases for microarray analysis. Finally, a pool of 5 reactive lymph node biopsies, characterized by follicular hyperplasia, was constructed for use as a reference tissue. Most lymphoma cases were included in one of our previous studies on NLPHL and/or THRLBCL (3, 7, 10, 11, 17).

This study was approved by the local ethical commission of the University Hospitals K.U.Leuven.

RNA extraction

Total RNA was extracted from 20 micron sections of each frozen tissue sample using the TriZol reagent (Invitrogen, Merelbeke, Belgium), followed by purification using the RNeasy mini kit (Qiagen, Venlo, The Netherlands), according to the manufacturer's recommendations. RNA quality and concentration were measured using a Nanodrop spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA).

Gene expression profiling

Five micrograms of RNA were biotin labeled and hybridized onto human oligonucleotide microarrays (Affymetrix HG-U133 Plus 2.0; Affymetrix, High Wycombe, UK). The resulting data are available online at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/projects/geo/), accession number GSE7788. These data were analyzed using Bioconductor software (18). Statistical testing for genes differentially expressed between the two lymphomas was done by a T-test. Multiple testing corrections were performed by a step-down maxT procedure (19).

The statistical significance of overlap with other expression profiling studies was calculated using hypergeometric statistics.

Immunohistochemistry

Aside from the immunohistochemical stainings used for diagnostic purposes, including CD20, CD3, CD4, CD8 and CD57 stainings, paraffin embedded sections were stained with a commercially available mouse anti-indoleamine 2,3 dioxygenase (IDO) monoclonal antibody (Chemicon international), following the manufacturer's recommendations. Figure 2 was taken with a Leica DMBL microscope, using a HCX PL Fluotar 40X/0.75 objective and a Leica DFC 300 camera. Adobe software was used for brightness/contrast correction.

Results and discussion

Clinical data

Clinical characteristics of the patients are summarized in Table 1. Both T cell/histiocyte rich large B cell lymphoma (THRLBCL) and nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL) show a clear male predominance. Ann Arbor staging, the International Prognostic Index (IPI) and the initial response to treatment confirm that THRLBCL is a very aggressive disease, while NLPHL is an indolent disorder. These results are further strengthened by the Kaplan-Meier estimates of overall survival (Supplementary Figure 1).

Table 1. Clinical data.

Disease entity		THRLBCL[1]	NLPHL[1]
Nr cases		28 [10]	47 [10]
male/female		22/6 [8/2]	39/8 [8/2]
median age in years (range)		50 (20-75) [50 (40-75)]	35 (7 - 74) [43 (22-71)]
stage (Ann Arbor)[2]	**I**	0 [0]	15 [5]
	II	3 [0]	9 [3]
	III	7 [3]	6 [1]
	IV	14 [7]	2 [0]
prognostic score (IPI)[2]	**Low**	5 [0]	not applicable
	low intermediate	8 [5]	
	high intermediate	4 [2]	
	High	6 [2]	
initial response to treatment[2]	**complete remission**	8 [4]	30 [8]
	partial remission	2 [0]	1 [0]
	progressive disease	13 [5]	0 [0]
median follow-up in years (range)[2]		2 (<1-8) [2 (<1-4)]	8 (<1 - 19) [10 (5-12)]
status at last follow-up[2]	**alive without disease**	4 [2]	24 [5]
	alive with disease	2 [1]	2 [1]
	death without disease	4 [0]	4 [2]
	death with disease	14 [7]	1 [0]

---

[1] The numbers of the 20 cases selected for microarray expression profiling are given between square parentheses.

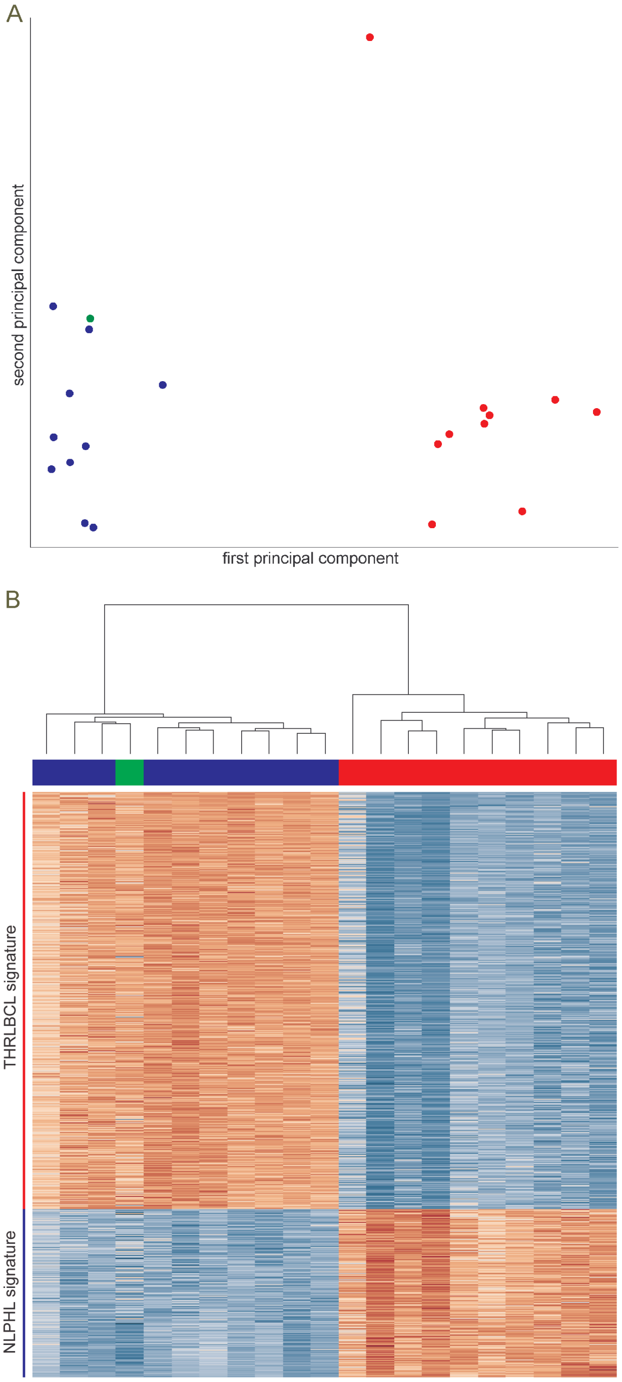
[2] Ann Arbor staging was available for 56 of 75 cases; IPI scoring for 23 of 28 THRLBCL cases; initial response to treatment for 54 of 75 cases; and follow-up for 55 of 75 cases.

Expression profiling of THRLBCL vs. NLPHL

Aiming to study the microenvironment, which constitutes the majority of the tumour cell mass in both THRLBCL and NLPHL, we performed microarray expression profiling on entire tissue sections. Principal component analysis revealed a clear distinction between these two lymphomas (Figure 1A). One THRLBCL was clearly separated from the other THRLBCL cases. Interestingly, this was the only sample taken from a spleen, while all other samples originated from lymph nodes. As the separation of this sample from the other THRLBCL tumours was in a direction perpendicular to the direction of separation of THRLBCL and NLPHL (Figure 1A), this sample was not removed in subsequent analyses. However, as a control, all subsequent analyses were repeated leaving out this aberrant sample, revealing similar results (data not shown). The reactive lymph node pool was located near the NLPHL samples, in agreement with expectations. Indeed, the microenvironment in NLPHL comprises components of the B follicle (with numerous small B cells) and adjacent T cell areas (with numerous T cells), while remnants of B follicles are mostly missing in THRLBCL.

Using highly significant differentially expressed genes, we constructed expression signatures of THRLBCL and NLPHL (Supplementary Figure 2). 392 genes were part of the THRLBCL signature, while the NLPHL signature contained 135 genes (Figure 1B, Supplementary Table 1). Consistent with the principal component analysis above, the reactive lymph node reference sample clustered together with the 10 NLPHL cases, when the 21 microarray profiles were clustered using only these two gene expression signatures (Figure 1B). Finally,

***Figure 1 (next page)*. Microarray expression profiling of 10 NLPHL cases, 10 THRLBCL cases, and a reference pool of lymph nodes with follicular hyperplasia.** A. Principal component analysis, performed on the complete microarray data (54675 probesets). Blue: NLPHL; red: THRLBCL; green: reactive lymph node pool. The first principal component (separating NLPHL from THRLBCL) captured 42 % of the total variance. The second principal component captured 11 % of the total variance. B. Heat map of the 874 differentially expressed probesets (527 unique genes, Supplementary Table 1). Top: cluster dendrogram, showing the a priori expected separation between the THRLBCL and NLPHL samples, and confirming the similarity between NLPHL and the reactive lymph node reference; middle: identity of samples (colors as in A.); bottom: graphical representation of gene expression (blue: high expression; red: low expression).


A
second principal component
first principal component
B
THRLBCL signature
NLPHL signature

the sample originating from a THRLBCL located in the spleen again clustered together with the other THRLBCL cases. This result persisted even when only the other 9 THRLBCL samples were used for building the gene expression signatures (data not shown).

To validate these gene expression data, we performed real-time quantitative RT-PCR on 10 genes (Supplementary Materials and Methods), selected for their involvement in interferon pathways, macrophage activation and innate immune responses. Five of these genes were in the THRLBCL signature (Supplementary Table 1B), while the other 5 genes were key genes in the selected pathways that were not differentially expressed according to our strict statistical criteria. With the exception of the gene CD74 (differentially expressed according to the microarray data but showing no significant expression difference by real-time quantitative RT-PCR), the obtained expression fold differences for these genes correspond well to the microarray data (Supplementary Table 2).

The gene expression signature of NLPHL: B cell genes

In comparison with the gene expression profile of THRLBCL, the expression signature of NLPHL comprises mainly genes characteristic of B cells (Table 2A, Supplementary Table 1A), in line with morphological findings. Moreover, the observed similarities between the expression profiles of NLPHL and the reactive lymph nodes, characterized by follicular hyperplasia, suggest that the components of the B follicle play a major role in both profiles. In line with these findings, the NLPHL signature shows significant overlap with the signature Monti *et al*. (14) found to be related to B cell receptor/proliferation in a subgroup of DLBCL (of the 43 genes in the B cell receptor/proliferation signature present on our microarray platform, 7 were a part of the NLPHL signature, $p = 4.4 \times 10^{-8}$, Supplementary Table 3A). In contrast, the overlap with the oxidative phosphorylation signature and host response signature of Monti *et al*. (14) was not more than randomly expected (2 genes and 0 genes respectively).

Table 2. A selection of genes differentially expressed between NLPHL and THRLBCL with $p < 0.001$.

A. Expressed at higher levels in NLPHL

HGNC	description	fold difference	p-value
FCRL1	Fc receptor-like 1	32.1	2.3E-12
CD79A	B-cell antigen receptor complex-associated protein alpha-chain precursor (Ig-alpha) (MB-1 membrane glycoprotein) (Surface IgM-associated protein) (Membrane-bound immunoglobulin-associated protein) (CD79a antigen)	12.4	2.2E-09
CD79B	B-cell antigen receptor complex-associated protein beta-chain precursor (B-cell-specific glycoprotein B29) (Immunoglobulin-associated B29 protein) (IG-beta) (CD79b antigen)	7.2	8.4E-08
CD19	B-lymphocyte antigen CD19 precursor (B-lymphocyte surface antigen B4) (Leu-12) (Differentiation antigen CD19)	18.1	4.1E-08
CD22	B-cell receptor CD22 precursor (Leu-14) (B-lymphocyte cell adhesion molecule) (BL-CAM) (Siglec-2)	15.0	2.4E-09
MS4A1	B-lymphocyte antigen CD20 (B-lymphocyte surface antigen B1) (Leu-16) (Bp35)	5.5	1.7E-09
PAX5	Paired box protein Pax-5 (B-cell-specific transcription factor) (BSAP)	8.3	4.9E-12
BCL11A	B-cell lymphoma/leukemia 11A (B-cell CLL/lymphoma 11A) (COUP-TF- interacting protein 1) (Ecotropic viral integration site 9 protein) (EVI-9)	12.0	1.1E-11
FGFR1OP	C-C chemokine receptor type 6 (C-C CKR-6) (CC-CKR-6) (CCR-6) (LARC receptor) (GPR-CY4) (GPRCY4) (Chemokine receptor-like 3) (CKR-L3) (DRY6) (G-protein coupled receptor 29) (CD196 antigen)	23.4	2.1E-10
FCER2	Low affinity immunoglobulin epsilon Fc receptor (Lymphocyte IgE receptor) (Fc-epsilon-RII) (BLAST-2) (Immunoglobulin E-binding factor) (CD23 antigen)	14.0	2.3E-10
BANK1	B-cell scaffold protein with ankyrin repeats 1	21.0	5.1E-08

B. Expressed at higher levels in THRLBCL

HGNC ID	Description	fold difference	p-value
FCER1G	High affinity immunoglobulin epsilon receptor gamma-subunit precursor (FceRI gamma) (IgE Fc receptor gamma-subunit) (Fc-epsilon RI-gamma)	9.7	5.1E-13
VSIG4	V-set and immunoglobulin domain-containing protein 4 precursor (Z39Ig protein)	569.0	6.9E-13
IDO	Indoleamine 2,3-dioxygenase (EC 1.13.11.42) (IDO) (Indoleamine-pyrrole 2,3-dioxygenase)	9.0	3.9E-08
CCL8	Small inducible cytokine A8 precursor (CCL8) (Monocyte chemotactic protein 2) (MCP-2) (Monocyte chemoattractant protein 2) (HC14)	143.5	1.1E-09
TLR1	Toll-like receptor 1 precursor (Toll/interleukin-1 receptor-like protein) (TIL) (CD281 antigen)	3.1	4.4E-11
TLR2	Toll-like receptor 2 precursor (Toll/interleukin 1 receptor-like protein 4) (CD282 antigen)	11.6	2.2E-11
TLR4	Toll-like receptor 4 precursor (hToll) (CD284 antigen)	4.0	2.5E-09

HGNC ID	Description	fold difference	p-value
TLR8	Toll-like receptor 8 precursor	11.5	1.4E-09
CD14	Monocyte differentiation antigen CD14 precursor (Myeloid cell-specific leucine-rich glycoprotein)	9.2	4.0E-10
STAT1	Signal transducer and activator of transcription 1-alpha/beta (Transcription factor ISGF-3 components p91/p84)	1.6	1.6E-10
CCR1	C-C chemokine receptor type 1 (C-C CKR-1) (CC-CKR-1) (CCR-1) (CCR1) (Macrophage inflammatory protein 1-alpha receptor) (MIP-1alpha-R) (RANTES-R) (HM145) (LD78 receptor) (CD191 antigen)	10.4	1.2E-10
CXCL10	Small inducible cytokine B10 precursor (CXCL10) (10 kDa interferon- gamma-induced protein) (Gamma-IP10) (IP-10)	7.7	6.6E-09
CXCL16	Small inducible cytokine B16 precursor (Transmembrane chemokine CXCL16) (SR-PSOX) (Scavenger receptor for phosphatidylserine and oxidized low density lipoprotein)	7.9	6.2E-10
CCRL2	C-C chemokine receptor-like 2 (Putative MCP-1 chemokine receptor) (Chemokine receptor CCR11) (Chemokine receptor X)	11.9	7.4E-08
CD80	T-lymphocyte activation antigen CD80 precursor (Activation B7-1 antigen) (CTLA-4 counter-receptor B7.1) (B7) (BB1)	3.4	1.7E-09
CD86	T-lymphocyte activation antigen CD86 precursor (Activation B7-2 antigen) (CTLA-4 counter-receptor B7.2) (B70) (FUN-1) (BU63)	3.8	2.5E-09
CD274	Programmed cell death 1 ligand 1 precursor (Programmed death ligand 1) (PD-L1) (PDCD1 ligand 1) (B7 homolog 1) (B7-H1) (CD274 antigen)	6.6	3.7E-08
CSF1R	Macrophage colony-stimulating factor 1 receptor precursor (CSF-1-R) (EC 2.7.10.1) (Fms proto-oncogene) (c-fms) (CD115 antigen)	4.2	3.0E-08
CSF3R	Granulocyte colony-stimulating factor receptor precursor (G-CSF-R) (CD114 antigen)	9.7	4.4E-11
PDCD1LG2	Programmed cell death 1 ligand 2 precursor (Programmed death ligand 2) (PD-L2) (PD-1-ligand 2) (PDCD1 ligand 2) (Butyrophilin B7-DC) (B7-DC) (CD273 antigen)	11.8	7.1E-08
FCGR3B	Low affinity immunoglobulin gamma Fc region receptor III-B precursor (IgG Fc receptor III-1) (Fc-gamma RIII-beta) (Fc-gamma RIIIb) (FcRIIIb) (Fc-gamma RIII) (FcRIII) (FcR-10) (CD16b antigen)	25.7	5.5E-10
FCGR1A	High affinity immunoglobulin gamma Fc receptor I precursor (Fc-gamma RI) (FcRI) (IgG Fc receptor I) (CD64 antigen)	35.3	3.8E-08
ICAM1	Intercellular adhesion molecule 1 precursor (ICAM-1) (Major group rhinovirus receptor) (CD54 antigen)	4.8	2.2E-10
IL1RN	Interleukin-1 receptor antagonist protein precursor (IL-1ra) (IRAP) (IL1 inhibitor) (IL-1RN) (ICIL-1RA)	67.7	3.7E-09
IL18BP	Interleukin-18-binding protein precursor (IL-18BP) (Tadekinig-alfa)	6.4	1.2E-09
IRAK3	Interleukin-1 receptor-associated kinase 3 (EC 2.7.11.1) (IRAK-3) (IL- 1 receptor-associated kinase M) (IRAK-M)	11.3	1.6E-11
CD74	HLA class II histocompatibility antigen gamma chain (HLA-DR antigens- associated invariant chain) (Ia antigen-associated invariant chain) (Ii) (p33) (CD74 antigen)	2.8	2.2E-08
S100A9	Protein S100-A9 (S100 calcium-binding protein A9) (Calgranulin-B) (Migration inhibitory factor-related protein 14) (MRP-14) (P14) (Leukocyte L1 complex heavy chain) (Calprotectin L1H subunit)	35.5	1.7E-11
CASP5	Caspase-5 precursor (EC 3.4.22.-) (CASP-5) (ICH-3 protease) (TY protease) (ICE(rel)-III)	20.0	1.7E-08

HGNC ID	Description	fold difference	p-value
MSR1	Macrophage scavenger receptor types I and II (Macrophage acetylated LDL receptor I and II) (Scavenger receptor class A member 1) (CD204 antigen)	38.3	2.9E-07
CD163	Scavenger receptor cysteine-rich type 1 protein M130 precursor (CD163 antigen) (Hemoglobin scavenger receptor)	50.5	3.1E-09
SOD2	Superoxide dismutase [Mn], mitochondrial precursor (EC 1.15.1.1)	5.7	5.4E-10
IFNAR1	Interferon-alpha/beta receptor alpha chain precursor (IFN-alpha-REC)	2.2	6.7E-09
IFNGR2	Interferon-gamma receptor beta chain precursor (Interferon-gamma receptor accessory factor 1) (AF-1) (Interferon-gamma transducer 1)	2.8	1.3E-09
IFIT3	Interferon-induced protein with tetratricopeptide repeats 3 (IFIT-3) (IFIT-4) (Interferon-induced 60 kDa protein) (IFI-60K) (ISG-60) (CIG49) (Retinoic acid-induced gene G protein) (RIG-G)	7.1	7.7E-10
IFI6	Interferon-induced protein 6-16 precursor (Ifi-6-16) (Interferon alpha-inducible protein 6)	4.7	1.7E-08
C1QA	Complement C1q subcomponent subunit A precursor	8.3	3.6E-09
C1QC	Complement C1q subcomponent subunit C precursor	6.4	4.8E-09
C2	Complement C2 precursor (EC 3.4.21.43) (C3/C5 convertase)	10.9	8.8E-10
C3AR1	C3a anaphylatoxin chemotactic receptor (C3a-R) (C3AR)	11.0	1.3E-11

The gene expression signature of THRLBCL: a tolerogenic immune response

The signature of THRLBCL underlines the crucial role of an IFN-γ regulated and tolerogenic pathway within the microenvironment. Indeed, IFN-γ is up-regulated in THRLBCL (Supplementary Table 2), as well as several genes encoding for proteins that are up-regulated in macrophages and dendritic cells upon treatment with IFN-γ (20, 21) (Table 2B, Supplementary Table 1B), indicative for a macrophage activated status. These genes include those encoding STAT1, Fc-γ receptor I (FcRI or CD64), ICAM-1, IFN-γ induced protein (IP-10/CXCL10), CXCL16, and in particular CCL8 and IDO (Table 2B, Supplementary Table 1B). CCL8, also designated as monocyte chemotactic protein 2 (MCP2), belongs to the CC chemokines. It is strongly induced by IFN-γ (22) and is one of the most potent chemoattractants for mononuclear cells, including monocytes and T cells (21). Thus, CCL8 may contribute to the histiocyte-rich (and T-cell rich) composition of the microenvironment in THRLBCL. IFN-γ also promotes, in a STAT1 dependent way, the induction of the tryptophan-degrading enzyme indoleamine 2,3-dioxygenase (IDO) in monocytes, macrophages and dendritic cells (23, 24). Interestingly, both B7-1 (CD80) and B7-2 (CD86) were part of this signature, and through a reverse interaction with CTLA4, these membrane proteins have been shown to activate IDO expression, as reviewed by Munn and Mellor (25). IDO has been described to promote tumour immune tolerance by suppressing local T cell responses and by altering the conversion of effector T cells into T regulatory cells (26, 27). Intriguingly, VSIG4 (V-set and Ig domain-containing 4, also known as Z39Ig), one of the most significant and strongly up-regulated genes of the THRLBCL signature, is a B7 family-related protein expressed by macrophages and dendritic cells that acts as a strong negative regulator of CD4 and CD8 T cell activation *in vitro* and *in vivo* (28). Thus together with IDO, VSIG4 may contribute to a state of immune suppression and tumour tolerance. Of interest, aside from its suppressive properties on T cells, VSIG4 has also been recognized as a new Complement Receptor of the Immunoglobulin superfamily (CRIg), required for phagocytosis of circulating pathogens (29). In line with this finding, the THRLBCL signature contains scavenger receptors (CXCL16, MSR1, CD163) and Toll-like receptors (TLR1, TLR2, TLR4 and TLR8).

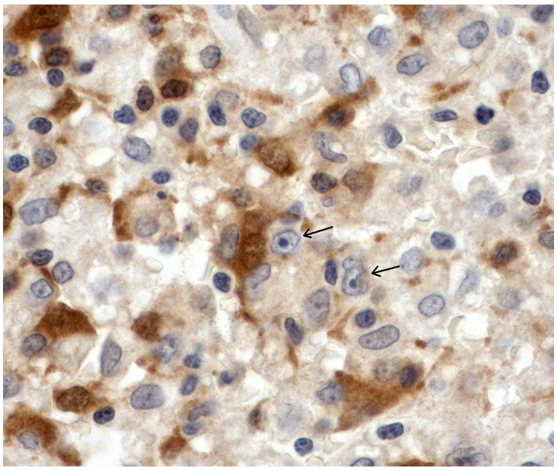


***Figure 2*. Immunohistochemical staining of IDO on a lymph node involved by THRLBCL.** Aside from small round cells expressing IDO (weak staining), several positive dendritic cells (large cells, intensely stained brown) are seen in the near vicinity of the tumour cells (indicated by arrows). These strongly stained dendritic cells were found in all assayed THRLBCL cases and in none of the NLPHL cases.

These data are indicative for innate immune responses in THRLBCL. Therefore, a possible involvement of pathogens in the initiation or propagation of the disease cannot be excluded and should be further investigated.

Altogether these findings suggest that CCL8 and IFN-γ are responsible for respectively the recruitment and the activation of monocytes, macrophages and dendritic cells and, in synergy with TLR-ligands, for the production of high levels of IDO. This key mediator is at least partly produced by dendritic cells, a subpopulation of the numerous histiocytes characterizing the THRLBCL stroma (Figure 2). These dendritic cells are present in all these THRLBCL cases and intensely stained by IDO immunohistochemistry, but absent in NLPHL. We speculate that the production of IDO and VSIG4 results in a tolerogenic microenvironment of the tumour cells, as schematically represented in Figure 3. This could explain the bad prognosis of these patients, in comparison with those affected by NLPHL.

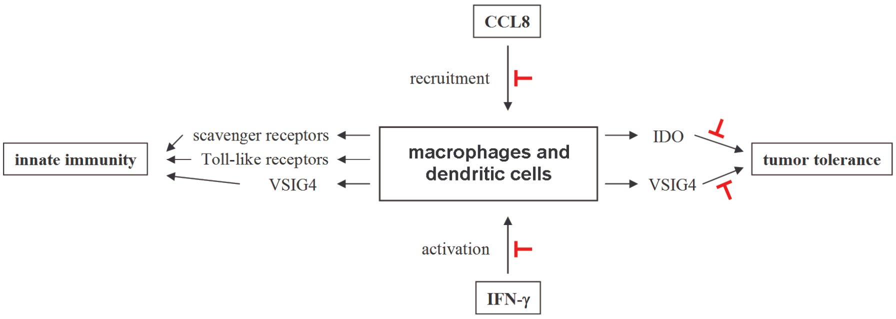


***Figure 3*. Schematical proposal on the host immune response in THRLBCL, based on our morphological and gene expression data, and on literature evidence.** By morphology, the microenvironment of THRLBCL is hallmarked by the presence of histiocytes/macrophages. Gene expression profiling data confirm the central role of macrophages and/or dendritic cells and suggest that these cells may be recruited by CCL8 (21, 22). IFN-γ activates these cells to produce IDO (23, 24). High levels of IDO, together with VSIG4, suppress the proliferation of effector T cells (such as CD8+ cytotoxic T cells), resulting in tumour tolerance (26, 28). Macrophages and dendritic cells also express receptors involved in innate immunity, including scavenger and Toll-like receptors, and VSIG4 as a complement receptor (29). Blocking the production and/or the function of CCL8, IFN-γ, and in particular IDO and VSIG4 may abrogate the induction of tumour tolerance. It is encouraging to note that inhibitors to target IDO are available (30).

The THRLBCL signature shows significant overlap with the signature Dave *et al*. (13) found to be related to an unfavourable immune response in part of the follicular lymphomas (9 of 23 genes, $p = 8.7 \times 10^{-10}$, Supplementary Table 3B). In addition, the signature Monti *et al*. (14) found to be related to host response in a subgroup of DLBCL, enriched in THRLBCL cases, was also significantly overrepresented in our THRLBCL signature (14 of 59 genes, $p = 3.8 \times 10^{-11}$, Supplementary Table 3C). In contrast, the favourable immune response of Dave *et al.* (13), as well as the oxidative phosphorylation and B cell receptor/proliferation signatures of Monti *et al*. (14) did not overlap more than randomly expected with our THRLBCL signature.

Absence of T cell genes in the NLPHL and THRLBCL gene expression signatures

Neither the NLPHL nor the THRLBCL gene expression signature contained a significant component of T cell-associated genes. As shown in Supplementary Table 4, this absence of T cell genes is not due to our strict statistical cut-off, as even with a cut-off of $p < 0.05$ after correction for multiple testing, none of the tested T cell-associated genes showed a significant difference. In addition, the ratio between $CD4^+$ and $CD8^+$ T cells, described to change in favour of the $CD8^+$ cells in THRLBCL (7, 31), was not reflected in the expression profiles either, although we did observe a (non-significant) tendency toward higher expression of $CD8\alpha$ (Supplementary Table 4). A partial explanation for this unexpected result can be sought in the presence of residual non-neoplastic T cell areas in all NLPHL cases. Furthermore, no difference in expression of CD57, described as a typical feature of the T cells surrounding the tumour cells in NLPHL, was found. The number of $CD57^+$ T cells present in NLPHL is variable and $CD57^+$ T cells are also found in THRLBCL (7, 31). Indeed, our immunohistochemical stainings confirm this profound difference in location of $CD57^+$ cells, but show that the number of these $CD57^+$ cells is approximately equal in both lymphoma entities (data not shown), in agreement with our expression profiling results.

Altogether, this absence of T cell-associated genes in the THRLBCL expression signature might be regarded as a confirmation that not the T cells, but the macrophages/histiocytes represent the functionally important component of the microenvironment in THRLBCL. Therefore, we hypothesize that the tolerogenic immune response, mainly orchestrated by macrophages/histiocytes and dendritic cells in the THRLBCL microenvironment, is the feature responsible for the adverse prognosis of this lymphoma entity, in comparison to NLPHL.

THRLBCL and NLPHL gene expression signatures in additional cases

As we observed large differences between the expression profiles of NLPHL and THRLBCL, we were wondering if a simple and intuitive classifier, based on real-time quantitative RT-PCR measurements of a very limited number of genes, would be able to discriminate both entities in additional cases. Therefore, we selected three genes from the gene expression signatures and we assayed those in 69 additional NLPHL and THRLBCL cases, both in-house and external, using quantitative RT-PCR (see Supplementary Materials and Methods). In all cases, the quantitative RT-PCR classification agreed with our morphological diagnosis (Supplementary Table 5, Supplementary Figure 3), although in a proportion of the 14 external cases (4 cases, all diagnosed with NLPHL), the difference between the NLPHL score and the THRLBCL score was less pronounced than in the in-house cases. These results underline that the NLPHL and THRLBCL cases selected for microarray expression are representative for these lymphoma entities and suggests that this classifier might have some diagnostic use, e.g. in supporting morphological findings.

Conclusion

We demonstrated that the gene expression profile of THRLBCL, in comparison with that of NLPHL, is hallmarked by an increased expression of IFN-γ, CCL8, IDO and VSIG4. Based on their function as described in the literature, the products of these genes point towards a distinct tolerogenic host immune response that may play a key role in the aggressive behaviour of this lymphoma. This leads us to put these mediators forward as potential targets for therapy.

Acknowledgments

We are indebted to Andreas Rosenwald for careful review of the manuscript. PVL is supported by a PhD research fellowship of the Research Foundation – Flanders (FWO).

Supplementary Data

The supplementary data accompanying this study can be accessed at http://homes.esat.kuleuven.be/~pvanloo/THRLBCL_sup.pdf.

References

(1) Ramsay AD, Smith WJ, Isaacson PG. T-cell-rich B-cell lymphoma. Am J Surg Pathol 1988 Jun;12(6):433-43.

(2) Chittal SM, Brousset P, Voigt JJ, Delsol G. Large B-cell lymphoma rich in T-cells and simulating Hodgkin's disease. Histopathology 1991 Sep;19(3):211-20.

(3) Delabie J, Vandenberghe E, Kennes C, Verhoef G, Foschini MP, Stul M, et al. Histiocyte-rich B-cell lymphoma. A distinct clinicopathologic entity possibly related to lymphocyte predominant Hodgkin's disease, paragranuloma subtype. Am J Surg Pathol 1992 Jan;16(1):37-48.

(4) Abramson JS. T-cell/histiocyte-rich B-cell lymphoma: biology, diagnosis, and management. Oncologist 2006 Apr;11(4):384-92.

(5) Gatter KC, Warnke R. Diffuse large B-cell lymphomas. In: Jaffe ES, Harris NL, Stein H, Vardiman J, editors. World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues.Lyon, France: IARC Press; 2001. p. 171-4.

(6) Rudiger T, Gascoyne RD, Jaffe ES, de JD, Delabie J, De Wolf-Peeters C, et al. Workshop on the relationship between nodular lymphocyte predominant Hodgkin's lymphoma and T cell/histiocyte-rich B cell lymphoma. Ann Oncol 2002;13 Suppl 1:44-51.

(7) Achten R, Verhoef G, Vanuytsel L, De Wolf-Peeters C. Histiocyte-rich, T-cell-rich B-cell lymphoma: a distinct diffuse large B-cell lymphoma subtype showing characteristic morphologic and immunophenotypic features. Histopathology 2002 Jan;40(1):31-45.

(8) Brauninger A, Kuppers R, Spieker T, Siebert R, Strickler JG, Schlegelberger B, et al. Molecular analysis of single B cells from T-cell-rich B-cell lymphoma shows the derivation of the tumor cells from mutating germinal center B cells and exemplifies means by which immunoglobulin genes are modified in germinal center B cells. Blood 1999 Apr 15;93(8):2679-87.

(9) Braeuninger A, Kuppers R, Strickler JG, Wacker HH, Rajewsky K, Hansmann ML. Hodgkin and Reed-Sternberg cells in lymphocyte predominant Hodgkin disease represent clonal populations of germinal center-derived tumor B cells. Proc Natl Acad Sci U S A 1997 Aug 19;94(17):9337-42.

(10) Franke S, Wlodarska I, Maes B, Vandenberghe P, Achten R, Hagemeijer A, et al. Comparative genomic hybridization pattern distinguishes T-cell/histiocyte-rich B-cell lymphoma from nodular lymphocyte predominance Hodgkin's lymphoma. Am J Pathol 2002 Nov;161(5):1861-7.

(11) Achten R, Verhoef G, Vanuytsel L, De Wolf-Peeters C. T-cell/histiocyte-rich large B-cell lymphoma: a distinct clinicopathologic entity. J Clin Oncol 2002 Mar 1;20(5):1269-77.

(12) Diehl V, Sextro M, Franklin J, Hansmann ML, Harris N, Jaffe E, et al. Clinical presentation, course, and prognostic factors in lymphocyte-predominant Hodgkin's disease and lymphocyte-rich classical Hodgkin's disease: report from the European Task Force on Lymphoma Project on Lymphocyte-Predominant Hodgkin's Disease. J Clin Oncol 1999 Mar;17(3):776-83.

(13) Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N Engl J Med 2004 Nov 18;351(21):2159-69.

(14) Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, Mihm M, et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood 2005 Mar 1;105(5):1851-61.

(15) Boudova L, Torlakovic E, Delabie J, Reimer P, Pfistner B, Wiedenmann S, et al. Nodular lymphocyte-predominant Hodgkin lymphoma with nodules resembling T-cell/histiocyte-rich B-cell lymphoma: differential diagnosis between nodular lymphocyte-predominant Hodgkin lymphoma and T-cell/histiocyte-rich B-cell lymphoma. Blood 2003 Nov 15;102(10):3753-8.

(16) Jaffe ES, Harris NL, Stein H, Vardiman J. World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues. In press; 2008.

(17) Franke S, Wlodarska I, Maes B, Vandenberghe P, Delabie J, Hagemeijer A, et al. Lymphocyte predominance Hodgkin disease is characterized by recurrent genomic imbalances. Blood 2001 Mar 15;97(6):1845-53.

(18) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.

(19) Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statistical Science 2003 Feb;18(1):71-103.

(20) Billiau A. Interferon-gamma: biology and role in pathogenesis. Adv Immunol 1996;62:61-130.

(21) Mantovani A, Sica A, Sozzani S, Allavena P, Vecchi A, Locati M. The chemokine system in diverse forms of macrophage activation and polarization. Trends Immunol 2004 Dec;25(12):677-86.

(22) Van Damme J, Proost P, Put W, Arens S, Lenaerts JP, Conings R, et al. Induction of monocyte chemotactic proteins MCP-1 and MCP-2 in human fibroblasts and leukocytes by cytokines and cytokine inducers. Chemical synthesis of MCP-2 and development of a specific RIA. J Immunol 1994 Jun 1;152(11):5495-502.

(23) Taylor MW, Feng GS. Relationship between interferon-gamma, indoleamine 2,3-dioxygenase, and tryptophan catabolism. FASEB J 1991 Aug;5(11):2516-22.

(24) Chon SY, Hassanain HH, Gupta SL. Cooperative role of interferon regulatory factor 1 and p91 (STAT1) response elements in interferon-gamma-inducible expression of human indoleamine 2,3-dioxygenase gene. J Biol Chem 1996 Jul 19;271(29):17247-52.

(25) Munn DH, Mellor AL. Indoleamine 2,3-dioxygenase and tumor-induced tolerance. J Clin Invest 2007 May;117(5):1147-54.

(26) Mellor AL, Munn DH. IDO expression by dendritic cells: tolerance and tryptophan catabolism. Nat Rev Immunol 2004 Oct;4(10):762-74.

(27) Curti A, Pandolfi S, Valzasina B, Aluigi M, Isidori A, Ferri E, et al. Modulation of tryptophan catabolism by human leukemic cells results in the conversion of CD25$^-$ into CD25$^+$ regulatory T cells. Blood 2007 Apr 1;109(7):2871-7.

(28) Vogt L, Schmitz N, Kurrer MO, Bauer M, Hinton HI, Behnke S, et al. VSIG4, a B7 family-related protein, is a negative regulator of T cell activation. J Clin Invest 2006 Oct;116(10):2817-26.

(29) Helmy KY, Katschke KJ, Jr., Gorgani NN, Kljavin NM, Elliott JM, Diehl L, et al. CRIg: a macrophage complement receptor required for phagocytosis of circulating pathogens. Cell 2006 Mar 10;124(5):915-27.

(30) Muller AJ, Scherle PA. Targeting the mechanisms of tumoral immune tolerance with small-molecule inhibitors. Nat Rev Cancer 2006 Aug;6(8):613-25.

(31) Fraga M, Sanchez-Verde L, Forteza J, Garcia-Rivero A, Piris MA. T-cell/histiocyte-rich large B-cell lymphoma is a disseminated aggressive neoplasm: differential diagnosis from Hodgkin's lymphoma. Histopathology 2002 Sep;41(3):216-29.

Study 4

Polysomy 17 in breast cancer: clinicopathological significance and impact on HER2 testing

Contribution of the doctorandus: statistical data analysis and interpretation of the results.
This study has been accepted for publication in the Journal of Clinical Oncology.

Polysomy 17 in breast cancer: clinicopathological significance and impact on HER2 testing

Isabelle Vanden Bempt[1,*], Peter Van Loo[2,3,4,*], Maria Drijkoningen[1,8], Patrick Neven[5,8], Ann Smeets[6,8], Marie-Rose Christiaens[6,8], Robert Paridaens[7,8] and Christiane De Wolf-Peeters[1]

[1]Department of Pathology, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Belgium; [2]Department of Molecular and Developmental Genetics, Flanders Institute for Biotechnology, Belgium; [3]Department of Human Genetics, Katholieke Universiteit Leuven, Belgium; [4]Bioinformatics group, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium; [5]Department of Obstetrics and Gynecology, [6]Department of Surgery-Senology, [7]Department of General Medical Oncology and [8]Multidisciplinary Breast Centre, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Belgium

* Both authors contributed equally to this work

Corresponding author:

Name: Isabelle Vanden Bempt

E-mail: Isabelle.vandenbempt@uz.kuleuven.ac.be

Department of Pathology, Minderbroedersstraat 12, 3000 Leuven, Belgium

Phone: +32 16 336644

Fax: +32 16 336640

Running head:

Significance of polysomy 17 in breast cancer

ABSTRACT

Purpose Polysomy 17 is frequently found in breast cancer and may complicate the interpretation of HER2 testing results. We investigated the impact of polysomy 17 on HER2 testing and studied its clinicopathological significance in relation to *HER2* gene amplification.

Patients and Methods In 226 patients with primary invasive breast carcinoma, *HER2* gene and chromosome 17 copy numbers were determined by dual-color FISH. The interpretation of FISH results was based on either absolute *HER2* gene copy number or the ratio *HER2*/Chromosome 17. Results were correlated to HER2 protein expression on IHC, *HER2* mRNA expression by RT-PCR and to various clinicopathological parameters.

Results All cases with an equivocal HER2 result by FISH, either by absolute *HER2* copy number (44/226, 19.5%) or by the ratio *HER2*/Chromosome 17 (3/226, 1.3%), displayed polysomy 17. On its own, polysomy 17 was not associated with HER2 overexpression on IHC or increased *HER2* mRNA levels by RT-PCR. Moreover and in contrast to *HER2* gene amplification, polysomy 17 was not associated with high tumor grade, hormone receptor negativity or reduced disease-free survival.

Conclusion Polysomy 17 affects HER2 testing in breast cancer and is a major cause of equivocal results by FISH. We show that tumors displaying polysomy 17 in the absence of *HER2* gene amplification resemble more HER2 negative than HER2 positive tumors. These findings urge the need for clinical trials to investigative whether or not polysomy 17 tumors benefit from HER2-targeted therapy.

INTRODUCTION

The search for prognostic markers and therapeutic targets in human breast cancer has revealed a major role for the HER2 oncoprotein. Overexpression of HER2 has been reported in 15 to 25% of invasive breast carcinomas.[1,2] In most cases, this can be attributed to amplification of the *HER2* oncogene located on the long arm of chromosome 17 (17q12).[3] Both HER2 overexpression and *HER2* gene amplification have been correlated with poor clinical outcome.[4-6] Apart from its prognostic value, the HER2 status has major therapeutic implications. Not only does HER2 overexpression predict response to certain chemotherapeutic agents such as anthracyclines or paclitaxel, it is also considered to be a strong predictive marker for clinical benefit from HER2-targeted therapy (trastuzumab) in the metastatic and more recently also in the adjuvant setting.[7-12] While tumors not expressing HER2 have virtually no chance of responding to trastuzumab, moderate or even high levels of expression are not always associated with a therapeutic success. Moreover, treating breast cancer patients with trastuzumab is expensive and not without risk since serious cardiac toxicity has been observed in approximately 1 to 4% of patients.[13] Therefore, correct identification of patients that will benefit from trastuzumab therapy is of utmost importance.

A wide variety of techniques can be applied to determine the HER2 status in breast cancer tissue. Of these, immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH) are most frequently used. Although both methods have shown high concordance in some studies, reproducibility remains poor in others.[14,15] Recently, an expert team assembled by the American Society of Clinical Oncology (ASCO) and the College of American Pathologists (CAP) has developed guidelines for HER2 testing in breast cancer.[16] Accordingly, HER2 testing results should be reported as either positive, negative or equivocal. The latter group represents a gray area of breast tumors scoring 2+ on IHC or having a modest increase in *HER2* gene copy number by FISH. Interestingly, equivocal HER2 testing results have been related to chromosome 17 polysomy.[17-19] Indeed, tumors featuring an increased chromosome 17 copy number will contain more copies of the *HER2* gene which could result in elevated HER2 expression. At present, it remains unknown to what extent polysomy 17 obscures the interpretation of HER2 testing results and whether or not polysomy 17 tumors share biological characteristics with true HER2 positive breast cancers. In the present study, we aimed to clarify this issue.

MATERIALS AND METHODS

Cases

Since 2002, routine HER2 FISH analysis (PathVysion, Vysis, Downers Grove, IL, USA) has been performed at the Pathology Department of the University Hospital Gasthuisberg on breast cancer cases showing an equivocal or positive HER2 result on IHC (score 2+/3+). From this series of 751 cases, non-invasive breast carcinomas, metastatic lesions, cases lacking clinical or pathological data and referral cases were excluded. The remaining 171 cases as well as a control series of 55 consecutive cases with a negative HER2 result on IHC were recruited into the present study. Table 1 summarizes the clinicopathological characteristics of the 226 included cases. Patients underwent mastectomy or local wide excision of their primary breast tumor with an axillary lymph node dissection at least at level I and II. Histopathological examination was performed on HE stained sections and tumors were classified and graded according to the World Health Organization Classification and the Elston and Ellis grading system respectively.[20,21] Disease-free survival was defined as the time period (in months) between the date of surgery and the date of recurrence or distant metastasis. All patients had the best standard of care for local and systemic treatment; trastuzumab was not yet standard of care in the adjuvant or neo-adjuvant setting. This study was approved by the Local Commission for Medical Ethics and Clinical Studies.

Immunohistochemistry

Immunohistochemical HER2 assessment was performed using the CB11 mouse monoclonal antibody (1/40 diluted, Novocastra Laboratories, Newcastle-upon-Tyne, UK). Staining results were scored as described previously.[22] For the present study, we validated our IHC assay against HercepTest (DakoCytomation, Glostrup, Denmark) in a subset of 50 cases (35 score 0/1+; 15 score 3+) according to the ASCO/CAP guidelines for HER2 testing[16], revealing 98% concordance between both IHC assays. For all cases, immunohistochemical data were available on the estrogen receptor (ER) and progesterone receptor (PR) status (mouse monoclonal antibodies NCL-ER-6F11, 1/30 diluted and NCL-PgR-312, 1/40 diluted) (Novocastra Laboratories).

Table 1. Clinicopathological characteristics of 226 invasive breast carcinoma cases included in our study.

Parameter	N (%)
Age	
≤ 50 years	75 (33.2)
> 50 years	151 (66.8)
Tumor grade	
I	12 (5.3)
II	75 (33.2)
III	139 (61.5)
Tumor size	
≤ 2 cm	103 (45.6)
> 2 cm	123 (54.4)
NPI	
I	44 (19.5)
II	102 (45.1)
III	80 (35.4)
LN status	
Negative	120 (53.1)
Positive	106 (46.9)
LVI	
Negative	180 (79.6)
Positive	46 (20.4)
ER status	
Negative	47 (20.8)
Positive	179 (79.2)
PR status	
Negative	90 (39.8)
Positive	136 (60.2)
HER2 status on IHC	
Negative (0/1+)	55 (24.3)
Equivocal (2+)	99 (43.8)
Positive (3+)	72 (31.8)
Follow-up	
Mean (months)	42.2
Metastasis	33 (14.6)

NPI, Nottingham Prognostic Index; LN, lymph node; LVI, lymphovascular invasion; ER, estrogen receptor; PR, progesterone receptor; IHC, Immunohistochemistry; N, number of cases; (%), percentage of cases

Fluorescence *in situ* Hybridization (FISH)

FISH analysis (PathVysion, Vysis) was performed manually, according to the manufacturer's recommendations. Using a Zeiss Axioplan 2 epifluorescence microscope (Carl Zeiss, Germany), we counted signals in at least 100 tumor nuclei in 2 or more separate regions of the tissue section. Averages of *HER2* gene and chromosome 17 copy number counts were rounded off to the nearest whole number; in case *HER2* gene amplification appeared as clusters of uncountable *HER2* signals, we estimated the average *HER2* gene copy number. FISH results were interpreted according to 2 different scoring methods: (1) based on absolute *HER2* gene copy number (*HER2* absolute) or (2) based on the ratio *HER2* gene/Chromosome 17 copy number (*HER2*/Chr17 ratio). Actually, these scoring methods were developed to be used in combination with respectively the Oncor INFORM *HER-2/neu* test kit (Ventana Medical Systems) or the PathVysion test kit. We have previously compared both FISH test kits in a series of 20 breast cancer cases and found that the count of absolute *HER2* gene copy number was nearly identical for both kits.[22] Therefore, we applied both scoring methods in combination with the PathVysion test kit. As proposed by the ASCO/CAP guidelines[16], an absolute *HER2* gene copy number lower than 4 or a *HER2*/Chr17 ratio of less than 1.8 was considered HER2 negative; an absolute *HER2* copy number between 4 and 6 or a *HER2*/Chr17 ratio between 1.8 and 2.2 was considered HER2 equivocal, and an absolute *HER2* gene copy number higher than 6 or a *HER2*/Chr17 ratio higher than 2.2 was considered HER2 positive. Polysomy 17 was defined as an average chromosome 17 copy number ≥ 3.[19,23] Lymphocytes, (myo)fibroblasts and normal epithelial cells served as internal control.

Quantitative Reverse Transcriptase Polymerase Chain Reaction

In 157 out of 226 cases, representative frozen tumor tissue was available for quantitative RT-PCR analysis. For each case, total RNA was extracted from 20µm sections using the RNeasy mini kit (Qiagen, Hilden, Germany). RNA purity and concentration were checked spectrophotometrically (Nanodrop Technologies, Wilmington DE, USA). One µg total RNA was reverse transcribed and PCR reactions on the resulting cDNA were performed in the ABI-Prism 7900 HT Sequence Detector (Applied Biosystems, Lennik, Belgium). PCR primers and probes for *HER2* and housekeeping gene *GAPDH* were obtained from Applied Biosystems (Taqman® Gene Expression Assays). Each sample was analyzed in triplicate in a MicroAmp^{TM} optical 96-well reaction plate (Applied Biosystems). A sample of normal breast tissue was used as a calibrator and the ∆∆Ct-method was applied to determine relative gene expression levels.[24]

Statistical analysis

Differences in *HER2* mRNA expression levels between different subgroups were assayed by a Wilcoxon rank sum test. Differences in clinicopathological variables between subgroups were checked using chi-square tests. The Bonferroni method was used for multiple testing correction. Survival analysis was performed using the Kaplan-Meier method. Survival differences between subgroups were assayed by log-rank tests. A p-value of < 0.05 was considered statistically significant.

RESULTS

Comparison between IHC and FISH for HER2 testing

This comparison is outlined in Table 2. An equivocal HER2 status by FISH was found in 44/226 cases (19.5%) based on absolute *HER2* gene copy number and in 3/226 (1.3%) based on the ratio *HER2*/Chr17. Note that none of these cases showed overexpression on IHC (score 3+). Remarkably, all cases with an equivocal HER2 status by FISH as well as those cases showing discordant HER2 testing results displayed polysomy 17.

Table 2. Comparison between IHC and FISH for determination of the HER2 status in breast cancer.

FISH **IHC**	**Positive**		**Equivocal**		**Negative**	
	***HER2* > 6**	**R > 2.2**	**4≤*HER2*≤ 6**	**1.8≤R≤2.2**	***HER2* < 4**	**R < 1.8**
Positive (3+)	72/72 (100)	72/72 (100)	0	0	0	0
Equivocal (2+)	27/99 (27.3)	25/99 (25.3)	37/99 (37.4)	1/99 (1.0)	35/99 (35.3)	73/99 (73.7)
Negative (0/1+)	3/55 (5.5)	0	7/55 (12.7)	2/55 (3.6)	45/55 (81.8)	53/55 (96.4)
	102/226 *(45.1)*	***97/226*** *(42.9)*	***44/226*** *(19.5)*	***3/226*** *(1.3)*	***80/226*** *(35.4)*	***126/226*** *(55.8)*

IHC, immunohistochemical determination of HER2 protein expression scored on a 0 to 3+ scale; FISH, fluorescence *in situ* hybridization to determine the HER2 status based on either absolute *HER2* gene copy number (*HER2*) or the ratio of *HER2*/Chromosome 17 copy number (R); Percentages are given between brackets

Impact of polysomy 17 on HER2 testing results by IHC and FISH

Polysomy 17 was observed in 104/226 cases (46.0%), either on its own (62/104) or in combination with *HER2* gene amplification (42/104). As shown in Table 3, polysomy 17 did not affect the interpretation of HER2 testing results by FISH when it was accompanied by *HER2* gene amplification. Furthermore, most of these cases showed overexpression on IHC (78.6% score 3+). By contrast, a score 3+ on IHC was not found in tumors displaying polysomy 17 in the absence of *HER2* gene amplification. Moreover, in cases where polysomy 17 on its own resulted in an absolute *HER2* gene copy number higher than 3, the interpretation of HER2 FISH results was obscured. Indeed, 44 cases showed a modest increase in *HER2* gene copy number due to polysomy 17 (4 to 6 copies) and were interpreted as equivocal by FISH if only *HER2* copies were counted. However, when chromosome 17 copy number was taken into account (*HER2*/Chr17 ratio), all these cases turned out to be HER2 negative. Five cases showed a relatively high increase in *HER2* gene copy number due to polysomy 17 (7 to 10 copies) and were interpreted as positive based on absolute *HER2* gene copy number. According to the *HER2*/Chr17 ratio however, 2 of these cases were interpreted as HER2 negative (ratio 7/4 and 7/5, both < 1.8) whereas 3 cases fell in the equivocal range (ratio 10/5 = 2.0). These data illustrate how polysomy 17 can be interpreted as HER2 positive or HER2 negative, depending on which scoring method is applied to interpret HER2 FISH results.

Table 3. Polysomy 17 in relation to HER2 testing results.

	PS 17 + *HER2* GA N (%)	**PS 17 - *HER2* GA N (%)**
HER2 status by IHC		
Positive (score 3+)	33 (78.6)	0
Equivocal (score 2+)	9 (21.4)	46 (74.2)
Negative (score 0/1+)	0	16 (25.8)
HER2 status by FISH (*HER2*)		
Positive (*HER2* > 6)	42 (100)	5 (8.1)
Equivocal (4 ≤ *HER2* ≤ 6)	0	44 (71.0)
Negative (*HER2* < 4)	0	13 (20.9)
HER2 status by FISH (R)		
Positive (R > 2.2)	42 (100)	0
Equivocal (1.8 ≤ R ≤ 2.2)	0	3 (4.8)
Negative (R < 1.8)	0	59 (95.2)
	N = 42	N = 62

IHC, immunohistochemical determination of HER2 protein expression scored on a 0 to 3+ scale; FISH, fluorescence *in situ* hybridization to determine the HER2 status based on either absolute *HER2* gene copy number (*HER2*) or the ratio of *HER2/C*hromosome 17 copy number (R); PS 17 + *HER2* GA, polysomy 17 accompanied by *HER2* gene amplification; PS 17 - *HER2* GA, polysomy 17 on its own, in the absence of *HER2* gene amplification; N, number of cases; (%), percentage of cases

Stratification

To investigate whether tumors displaying polysomy 17 in the absence of *HER2* gene amplification should be regarded as HER2 negative or HER2 positive, we compared

HER2 mRNA levels and clinicopathological characteristics in the following 3 groups: (1) HER2 negative tumors (normal *HER2* gene and chromosome 17 copy number, n = 67), (2) Polysomy 17 tumors (polysomy 17 in the absence of *HER2* gene amplification, n = 62) and (3) HER2 positive tumors (*HER2* gene amplification defined as a ratio *HER2*/Chr17 ≥ 2.2, n = 97).

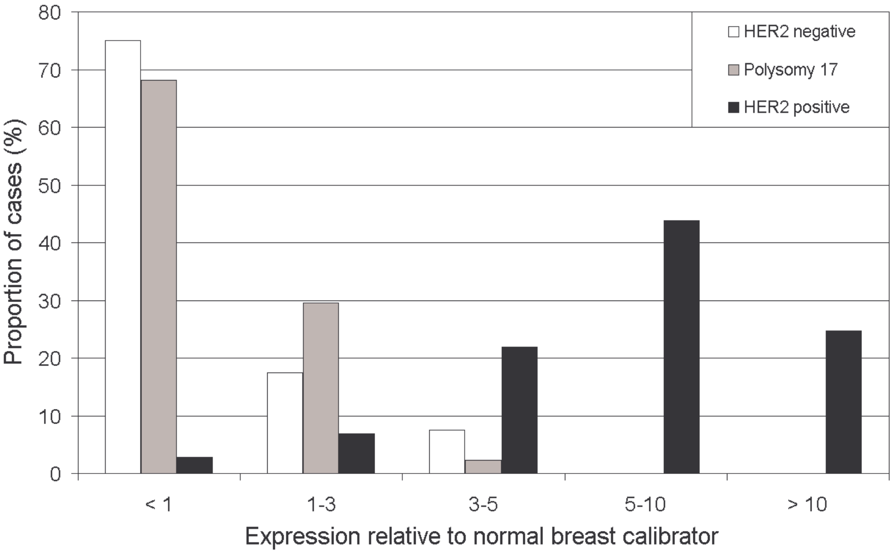


Figure 1. Distribution of HER2 mRNA expression values in polysomy 17 tumors compared to HER2 negative and HER2 positive cases. Polysomy 17 tumors (gray bars) and HER2 negative tumors (white bars) show a similar distribution pattern of low HER2 mRNA expression values. By contrast, HER2 positive cases (black bars) frequently show elevated expression values with most cases having at least a 5-fold increase in HER2 mRNA expression compared to normal breast tissue.

HER2 mRNA expression by quantitative RT-PCR

As illustrated in Figure 1, polysomy 17 tumors had low relative *HER2* mRNA expression values comparable to those found in the HER2 negative group (mean expression 0.914 versus 0.912, p = 0.1865). By contrast, HER2 positive tumors generally had increased relative expression values, with most cases showing at least a 5-fold increase in *HER2* mRNA expression compared to normal breast tissue. In HER2 positive cases, *HER2* mRNA expression levels were significantly higher than those in HER2 negative (mean 7.831 versus 0.912, $p < 10^{-15}$) and polysomy 17 tumors (mean 7.831 versus 0.914, $p < 10^{-16}$).

Table 4. Distribution of clinicopathological features in polysomy 17 tumors compared to HER2 negative and HER2 positive cases.

Parameter	HER2 negative N (%)	p-value^{1}	Polysomy 17 N (%)	p-value^{2}	HER2 positive N (%)	p-value^{3}
Age		1.0		0.18		1.0
≤ 50 years	21 (31.3)		14 (22.6)		42 (43.3)	
> 50 years	46 (68.7)		48 (77.4)		55 (56.7)	
Tumor grade		1.0		**3.1×10^{-3}**		**7.5×10^{-9}**
I	6 (8.9)		5 (8.1)		1 (1.0)	
II	38 (56.7)		23 (37.1)		14 (14.5)	
III	23 (34.4)		34 (54.8)		82 (84.5)	
Tumor size		1.0		1.0		1.0
≤ 2 cm	29 (43.3)		24 (38.7)		41 (42.3)	
> 2 cm	38 (56.7)		38 (61.3)		56 (57.7)	
NPI		1.0		0.71		**0.030**
I	20 (29.8)		15 (24.2)		9 (9.3)	
II	30 (44.8)		27 (43.5)		45 (46.4)	
III	17 (25.4)		20 (32.3)		43 (44.3)	
LN status		1.0		1.0		1.0
Negative	41 (61.2)		35 (56.5)		44 (45.4)	
Positive	26 (38.8)		27 (43.5)		53 (54.6)	
LVI		1.0		1.0		1.0
Negative	55 (82.1)		53 (85.5)		72 (74.2)	
Positive	12 (17.9)		9 (14.5)		25 (25.8)	
ER status		1.0		**1.2×10^{-4}**		**1.4×10^{-4}**
Negative	5 (7.5)		4 (6.5)		38 (39.2)	
Positive	62 (92.5)		58 (93.5)		59 (60.8)	
PR status		1.0		**0.024**		**6.2×10^{-3}**
Negative	18 (26.9)		18 (29.0)		54 (55.7)	
Positive	49 (73.1)		44 (71.0)		43 (44.3)	
Follow-up		NA		NA		NA
Mean (months)	47.7		48.6		33.3	
Metastasis	3 (4.5)		9 (14.5)		21 (21.6)	
	N = 67		N = 62		N = 97	

NPI, Nottingham Prognostic Index; LN, lymph node; LVI, lymphovascular invasion; ER, estrogen receptor; PR, progesterone receptor; N, number of cases; (%), percentage of cases; p-value^{1}, HER2 negative versus Polysomy 17; p-value^{2}: Polysomy 17 versus HER2 positive; p-value^{3}: HER2 positive versus HER2 negative; statistically significant differences are highlighted in bold; NA, not applicable; all p-values given are the result of a Chi-square test, corrected for multiple testing by the Bonferroni method.

Clinicopathological characteristics of polysomy 17 tumors

Table 4 shows the distribution of clinicopathological parameters in HER2 negative, polysomy 17 and HER2 positive tumors. Compared to HER2 negative cases, HER2 positive tumors showed higher tumor grade ($p < 10^{-8}$), higher NPI risk group (p =

0.030) and were more frequently ER negative (p < 0.001) and PR negative (p = 0.0062). Polysomy 17 tumors were more similar to HER2 negative than HER2 positive cases. While tumor grade (p = 0.0031), ER status (p < 0.001) and PR status (p = 0.024) differed significantly between polysomy 17 tumors and HER2 positive cases, no differences were found between polysomy 17 tumors and HER2 negative cases in any of the clinicopathological parameters investigated. Kaplan-Meier survival curves (Figure 2) illustrate shorter disease-free survival in HER2 positive tumors compared to HER2 negative cases (p < 0.001). Survival in polysomy 17 tumors was intermediate between HER2 negative (not significant, p = 0.056) and HER2 positive cases (p = 0.031).

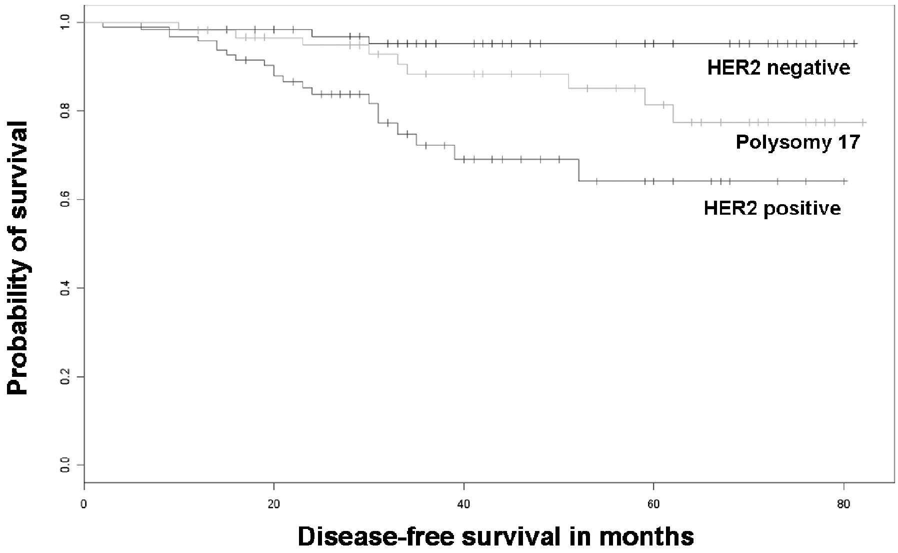


Figure 2. Kaplan-Meier plot of polysomy 17 tumors compared to HER2 negative and HER2 positive cases.

DISCUSSION

Polysomy 17 is common in breast cancer, with reported frequencies ranging from 13 to 46% depending on the study population and the definition of polysomy 17.[17-19,23,25] In our series, including 171 cases with an IHC score 2+/3+ and 55 with a score 0/1+, polysomy 17 was found in 46.0%. Since polysomy 17 implies extra copies of the *HER2* gene, is conceivable that polysomy 17 might lead to increased HER2 expression levels. However, it remains unclear whether or not polysomy 17 results in HER2 overexpression in a way similar to *HER2* gene amplification and whether or not polysomy 17 tumors should be regarded as HER2 positive.[17,18,25-29]

In the present study, we provide evidence that polysomy 17 and *HER2* gene amplification are two distinct genetic aberrations with a different clinicopathological significance in breast cancer. First, we show that polysomy 17 on its own does not result in HER2 overexpression, neither at the protein nor at the mRNA level. Indeed, we did not encounter any breast tumor showing HER2 overexpression on IHC (score 3+) and polysomy 17 in the absence of *HER2* gene amplification. Moreover and in line with Dal Lago *et al*[26], we did not find increased *HER2* mRNA expression by quantitative RT-PCR in polysomy 17 cases, not even in those tumors showing up to 10 *HER2* gene copies. Second, we could demonstrate that *HER2* gene amplification and polysomy 17 have a different clinicopathological impact in breast cancer. While *HER2* gene amplification was clearly associated with high tumor grade, hormone receptor negativity and reduced disease-free survival, polysomy 17 did not show any significant association with adverse clinicopathological parameters. Nevertheless, a trend toward shorter disease-free survival was observed in polysomy 17 tumors. Since polysomy 17 may reflect aneuploidy and increased chromosomal instability, it can be expected that tumors harboring this anomaly will behave more aggressively than those without it.[30-32] Still, our data suggest that the clinicopathological impact of polysomy 17 is not as strong as that of *HER2* gene amplification in breast cancer and that tumors displaying polysomy 17 in the absence of *HER2* gene amplification behave more similar to HER2 negative than to HER2 positive tumors.

Our current findings could have important clinical implications. Since polysomy 17 on its own is not associated with HER2 overexpression and since it does not have the same clinicopathological significance as true *HER2* gene amplification, one may wonder whether polysomy 17 tumors benefit from HER2-targeted therapy such as trastuzumab, which targets the HER2 protein at the tumor cell membrane. Indeed, the best therapeutic response rates have been observed in breast cancers showing HER2 overexpression by IHC.[28,33] Recently, trastuzumab response has been reported in two cases showing polysomy 17 in the absence of *HER2* gene amplification and in one case showing neither polysomy 17 nor *HER2* gene amplification. Of interest, all 3 cases showed HER2 overexpression on IHC.[28] We speculate that in such rare cases HER2 overexpression might result from deregulated gene transcription. Further phase III trials are needed to elucidate whether or not polysomy 17 tumors benefit from HER2-targeted therapy.

It is important to realize that polysomy 17 has a substantial impact on the interpretation of HER2 testing results, especially in those cases with an equivocal HER2 status on IHC (score 2+). Indeed, in those cases where polysomy 17 results in a moderate increase in *HER2* gene copy number (4 to 6), HER2 FISH results could be interpreted as equivocal if only absolute *HER2* copies are counted. As such, we found that 37/99 (37.4%) cases with an IHC score 2+ were still considered “equivocal” after FISH analysis based on absolute *HER2* copy number whereas only

1 case (1.0%) remained equivocal based on the *HER2*/chromosome 17 ratio. Based on these data and given that about 15% of newly diagnosed breast cancers show an IHC score 2+[16], we estimate that 5.6% and 0.15% of breast carcinomas remain equivocal after FISH testing, depending on whether or not a control probe for chromosome 17 is used. Remarkably, the one case (0.15%) showing an equivocal HER2 status on both IHC and FISH, even after correction for chromosome 17 copy number, also displayed polysomy 17 with a mean ratio of 10/5 or 2.0. In this particular case, quantitative RT-PCR indicated no increase in *HER2* mRNA expression, suggesting a negative HER2 status after all. In the end, one may wonder whether quantitative RT-PCR could be a valuable alternative for HER2 testing in routine clinical practice. Still, the need for representative fresh or frozen breast cancer tissue for optimal RT-PCR testing results, as well as inevitable dilution of invasive tumor cells with normal and stromal cell populations or non-invasive breast lesions limits the use of RT-PCR for routine HER2 testing.

In conclusion, polysomy 17 is a major cause of equivocal HER2 testing results by FISH. We provide evidence that polysomy 17 and *HER2* gene amplification have a distinct impact on the clinicopathological parameters in breast cancer and that polysomy 17 tumors should be regarded HER2 negative. Indeed, *HER2* gene amplification usually results in excessive HER2 expression levels and defines a distinct clinicopathological breast cancer entity characterized by high tumor grade, reduced hormone receptor expression and a poor prognosis. By contrast, polysomy 17 is not related to HER2 overexpression or adverse clinicopathological features but may rather reflect increased chromosomal instability in breast cancer. These findings underscore the importance of using dual-color systems for HER2 (F)ISH testing and urge the need for clinical trials to investigate whether or not polysomy 17 tumors benefit from HER2-targeted therapy.

Acknowledgements

We are grateful to Drs. Peter Vandenberghe and Iwona Wlodarska from the Centre of Human Genetics for helpful collaboration. We thank Miet Vanherck, Johan Van Even, Helga Van Den Bosch and Lieve Ophalvens for excellent technical assistance.

REFERENCES

(1) Revillion F, Bonneterre J, Peyrat JP: ERBB2 oncogene in human breast cancer and its clinical significance. Eur J Cancer 34:791-808, 1998

(2) van de Vijver MJ, Mooi WJ, Peterse JL, et al: Amplification and over-expression of the neu oncogene in human breast carcinomas. Eur J Surg Oncol 14:111-114, 1988

(3) Pauletti G, Godolphin W, Press MF, et al: Detection and quantitation of HER-2/neu gene amplification in human breast cancer archival material using fluorescence in situ hybridization. Oncogene 13:63-72, 1996

(4) Kallioniemi OP, Holli K, Visakorpi T, et al: Association of c-erbB-2 protein over-expression with high rate of cell proliferation, increased risk of visceral metastasis and poor long-term survival in breast cancer. Int J Cancer 49:650-655, 1991

(5) Press MF, Bernstein L, Thomas PA, et al: HER-2/neu gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas. J Clin Oncol 15:2894-2904, 1997

(6) Slamon DJ, Clark GM, Wong SG, et al: Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235:177-182, 1987

(7) Joensuu H, Kellokumpu-Lehtinen PL, Bono P, et al: Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. N Engl J Med 354:809-820, 2006

(8) Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al: Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 353:1659-1672, 2005

(9) Romond EH, Perez EA, Bryant J, et al: Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. N Engl J Med 353:1673-1684, 2005

(10) Slamon DJ, Leyland-Jones B, Shak S, et al: Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 344:783-792, 2001

(11) Smith I, Procter M, Gelber RD, et al: 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. Lancet 369:29-36, 2007

(12) Vogel CL, Cobleigh MA, Tripathy D, et al: Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 20:719-726, 2002

(13) Hayes DF, Picard MH: Heart of darkness: the downside of trastuzumab. J Clin Oncol 24:4056-4058, 2006

(14) Paik S, Bryant J, Tan-Chiu E, et al: Real-world performance of HER2 testing--National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst 94:852-854, 2002

(15) Roche PC, Suman VJ, Jenkins RB, et al: Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. J Natl Cancer Inst 94:855-857, 2002

(16) Wolff AC, Hammond ME, Schwartz JN, et al: American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 25:118-145, 2007

(17) Hyun CL, Lee HE, Kim KS, et al: The effect of chromosome 17 polysomy on HER2-/neu status in breast cancer. J Clin Pathol [Epub ahead of print]

(18) Merola R, Mottolese M, Orlandi G, et al: Analysis of aneusomy level and HER-2 gene copy number and their effect on amplification rate in breast cancer specimens read as 2+ in immunohistochemical analysis. Eur J Cancer 42:1501-1506, 2006

(19) Salido M, Tusquets I, Corominas JM, et al: Polysomy of chromosome 17 in breast cancer tumors showing an overexpression of ERBB2: a study of 175 cases using fluorescence in situ hybridization and immunohistochemistry. Breast Cancer Res 7: R267-R273, 2005

(20) Elston EW, Ellis IO: Method for grading breast cancer. J Clin Pathol 46:189-190, 1993

(21) Tavassoli FA, Devilee P: World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Breast and Female Genital Organs. Lyon, France, IARC Press, 2003

(22) Vanden Bempt I, Vanhentenrijk V, Drijkoningen M, et al: Real-time reverse transcription-PCR and fluorescence in-situ hybridization are complementary to understand the mechanisms involved in HER-2/neu overexpression in human breast carcinomas. Histopathology 46:431-441, 2005

(23) Torrisi R, Rotmensz N, Bagnardi V, et al: HER2 status in early breast cancer: Relevance of cell staining patterns, gene amplification and polysomy 17. Eur J Cancer 43:2339-2344, 2007

(24) Winer J, Jung CK, Shackel I, et al: Development and validation of real-time quantitative reverse transcriptase-polymerase chain reaction for monitoring gene expression in cardiac myocytes in vitro. Anal Biochem 270:41-49, 1999

(25) Ma Y, Lespagnard L, Durbecq V, et al: Polysomy 17 in HER-2/neu status elaboration in breast cancer: effect on daily practice. Clin Cancer Res 11:4393-4399, 2005

(26) Dal Lago L., Durbecq V, Desmedt C, et al: Correction for chromosome-17 is critical for the determination of true Her-2/neu gene amplification status in breast cancer. Mol Cancer Ther 5:2572-2579, 2006

(27) Downs-Kelly E, Yoder BJ, Stoler M, et al: The influence of polysomy 17 on HER2 gene and protein expression in adenocarcinoma of the breast: a fluorescent in situ hybridization, immunohistochemical, and isotopic mRNA in situ hybridization study. Am J Surg Pathol 29:1221-1227, 2005

(28) Hofmann M, Stoss O, Gaiser T, et al: Central HER2 IHC and FISH analysis in a trastuzumab (Herceptin(R)) Phase II monotherapy study: assessment of test sensitivity and impact of chromosome 17 polysomy. J Clin Pathol 61:89-94, 2008

(29) Lal P, Salazar PA, Ladanyi M, et al: Impact of polysomy 17 on HER-2/neu immunohistochemistry in breast carcinomas without HER-2/neu gene amplification. J Mol Diagn 5:155-159, 2003

(30) Carlson JA, Healy K, Tran TA, et al: Chromosome 17 aneusomy detected by fluorescence in situ hybridization in vulvar squamous cell carcinomas and synchronous vulvar skin. Am J Pathol 157:973-983, 2000

(31) Ichikawa D, Hashimoto N, Hoshima M, et al: Analysis of numerical aberrations of specific chromosomes by fluorescent in situ hybridization as a diagnostic tool in breast cancer. Cancer 77:2064-2069, 1996

(32) Watters AD, Going JJ, Cooke TG, et al: Chromosome 17 aneusomy is associated with poor prognostic factors in invasive breast carcinoma. Breast Cancer Res Treat 77:109-114, 2003

(33) Seidman AD, Fornier MN, Esteva FJ, et al: Weekly trastuzumab and paclitaxel therapy for metastatic breast cancer with analysis of efficacy by HER2 immunophenotype and gene amplification. J Clin Oncol 19:2587-2595, 2001

General discussion and perspectives

1 Computational methods for gene prioritization and *cis*-regulatory module detection

In this work, we developed two novel computational methods: Endeavour and ModuleMiner. ModuleMiner detects similar *cis*-regulatory modules in a set of co-regulated or co-expressed genes, while the purpose of Endeavour is candidate gene prioritization. We validated both methods extensively *in silico* and we performed an integrated case-study combining *cis*-regulatory module detection and gene prioritization.

1.1 ModuleMiner: *cis*-regulatory module detection

In the general introduction to this work, we divided the available *cis*-regulatory module detection algorithms into three classes: (i) methods that detect CRMs as clusters of binding sites for a user-defined set of PWMs (Type I CRM detection algorithms), (ii) methods that detect similar CRMs in co-regulated genes (Type II CRM detection methods) and (iii) methods that detect CRMs genome-wide using no prior combination of PWMs. Type I algorithms are the best performing of the three, but the paucity of available data limits their practical applicability. Type II algorithms show an intermediate performance, and are more generally applicable, as sets of co-regulated or co-expressed genes can easily be obtained, e.g. from microarray experiments. Type III algorithms tackle a considerably more difficult problem than Type I or Type II algorithms, resulting in lower performance. However, they can easily be applied in a high-throughput manner to find CRMs in the complete genome.

We reasoned Type II CRM detection algorithms show a favorable balance between general applicability and performance limitations. Furthermore, only a limited number of "first-generation" methods pre-existed for this subtype, allowing opportunities for improvement. For these reasons, we aimed to develop a novel Type II CRM algorithm, ModuleMiner. Contrary to the existing approaches, ModuleMiner specifically looks for the combination of PWMs that shows maximum specificity for the given set of co-regulated genes, compared to all other genes in the genome. This whole genome optimization procedure allows the elimination of a number of critical parameters of the existing algorithms. Indeed, most of the existing algorithms require the definition of the length of the CRMs and the number of PWMs involved (tables 3 and 5 in the General introduction), while ModuleMiner's whole genome optimization procedure allows an optimization over these parameters as well.

ModuleMiner can be considered a wrapper algorithm around ModuleScanner, a Type I CRM detection algorithm (tables 1 and 2 in the General introduction). Given a Transcriptional Regulatory Model (TRM, a combination of PWMs, supplemented by a number of parameters), ModuleScanner scans the genome for CRMs that fit this TRM. ModuleMiner uses the best CRM prediction near each gene to rank all genes in the genome. By combining the ranks of the given co-regulated genes (i.e. the input to the algorithm), ModuleMiner can assign a "fitness score" to that TRM. By varying these TRMs and repeating this ModuleScanner-centered procedure a large number of times (in a genetic algorithm-based optimization strategy), ModuleMiner obtains the optimal TRM for that given set of co-regulated genes. The choice of ModuleScanner as the Type I CRM detection algorithm was determined mainly by reasons of computational complexity: as for each TRM the complete genome needs to be scanned, and typically about 50000 TRMs are evaluated in the optimization process, the speed (and as a consequence also the limited complexity) of the underlying Type I algorithm is of paramount importance. It would be interesting to evaluate to what extent ModuleMiner's performance can be increased by replacing ModuleScanner by more advanced Type I algorithms, most particularly methods based on hidden Markov models (e.g. Ahab (Rajewsky *et al.*, 2002), Stubb (Sinha *et al.*, 2003, 2004), table 1 in the introductory chapter) and the recent Enhancer Element Locator algorithm which aligns sequences in the motif domain (Hallikas *et al.*, 2006). However, in practice, this will likely only be possible in a few years time, when more computational power becomes available.

We validated ModuleMiner by direct comparison with other state-of-the-art Type II CRM detection algorithms on benchmark data, and we observed a consistent higher performance of our novel algorithm. We also evaluated the sensitivity of ModuleMiner to noise in its input genes, leading to the conclusion that provided a "critical mass" of co-regulated genes is available, ModuleMiner is highly insensitive to additional non-co-regulated genes.

We applied ModuleMiner on a larger scale, to (i) sets of genes involved in specific embryonic development processes and (ii) microarray clusters containing genes co-expressed in specific adult tissues. In most of these gene sets, ModuleMiner was able to identify similar CRMs, as confirmed by five-fold cross-validation and/or leave-one-out cross-validation. In total, 209 CRMs were identified in 9 (of 10) presented microarray clusters, and 48 CRMs were identified in 5 (of 5) custom-build embryonic development gene sets.

When we regarded the positions of both sets of CRMs, we noticed a pronounced difference. CRMs predicted to direct expression in terminally differentiated tissues are highly enriched close to the transcription start site, while CRMs predicted to direct expression during embryonic development are more evenly distributed and may even be depleted in the proximal promotor region. Indeed, 63 % of the "adult tissue" CRMs, and only 8 % of the "embryonic development" CRMs are within 200 base pairs of the transcription start site. This led us to hypothesize that transcription regulation in adult tissue expression is mostly exerted by proximal promotors, while transcription regulation during embryonic development is mostly controlled by more distal enhancers.

The recently developed Type I (and Type III) CRM detection algorithm Enhancer Element Locator (Hallikas *et al.*, 2006) is based on the novel principle of aligning predicted transcription factor binding sites (not sequences), and shows good performance. It would be interesting to apply this same principle in a Type II CRM detection algorithm.

One long term goal of these CRM detection methods is a complete annotation of all transcriptional regulatory elements in the human genome. Because of the reasons mentioned above, we believe Type II CRM detection algorithms are the most likely to deliver major contributions to this in the near and mid-term future. We expect that the protein binding microarray technique (Mukherjee *et al.*, 2004) will make PWMs available for a large part of the human transcription factor repertoire, and that this will have a strong positive effect on the performance of Type II(a) CRM detection methods.

1.2 Endeavour: gene prioritization by genomic data fusion

In the field of computational gene prioritization, three novel well performing systems biology methods have recently been developed: Prioritizer (Franke *et al.*, 2006), the method by Lage *et al.* (2007), and our own method Endeavour (this work; Aerts *et al.* (2006)). While the former two are network-based methods integrating multiple data sources into one network, Endeavour is not: it uses each data source separately to prioritize the candidate genes, and then fuses the obtained individual prioritizations into one global prioritization. An advantage of both network-based prioritization methods is the biologically attractive representation of the underlying data. However, the Endeavour framework is more modular and able to handle more heterogeneous data sources. Indeed, data sources that cannot easily be converted into networks (e.g. literature data, microarray data and sequence similarity) represent challenges for these network-based methods. Although often these challenges can be overcome, converting these data into networks will likely result in a loss of information and a drop in performance.

Endeavour now integrates data from more than 10 different data sources: (i) text mining, (ii) Gene Ontology annotations, (iii) protein domains (InterPro), (iv) pathway information (the Kyoto Encyclopedia of Genes and Genomes, KEGG), (v) anatomical EST expression data, (vi) microarray gene expression data, (vii) *cis*-regulatory motif data, (viii) *cis*-regulatory modules (ModuleSearcher, Aerts *et al.* (2004)), (ix) sequence similarity (BLAST), (x) protein-protein interactions (the Biomolecular Interaction Network Database, BIND) and (xi) general disease probabilities (Lopez-Bigas and Ouzounis, 2004; Adie *et al.*, 2005). Each of these data sources is used separately to prioritize the candidate genes, based on similarities with a given set of training genes. Finally, order statistics are used to integrate these prioritizations into one overall prioritization.

We validated Endeavour by a large-scale leave-one-out cross-validation on 627 disease genes from the Online Mendelian Inheritance in Man (OMIM) database and 76 pathway genes from Gene Ontology, also including sets of random genes as a control. Our results showed that each data source separately performs better than random (both for diseases and for pathways), and hence information useful for candidate gene prioritization can be extracted from each data source. In addition, the data fusion of all data sources resulted in high performance disease and pathway gene prioritization, showing AUC values of 87 % and 90 % respectively. In addition, case-studies using recently identified monogenic and polygenic disease genes confirmed a high performance for prioritizing monogenic disease genes (average rank was 11 out of 200 candidate genes) and a better then random performance for prioritizing polygenic disease genes (average rank 40 of 200 candidate genes). In addition, these analyses showed that our text mining data source can also extract information not explicitly present in the literature.

Our collaborators applied Endeavour to the search for a gene involved in a recurrent chromosomal deletion containing 58 candidate genes in DiGeorge syndrome. Endeavour put forward Ypel1 as a candidate gene, which was subsequently validated by morpholino knockdown studies in zebrafish (Aerts *et al.*, 2006). Apart from this in-house application of our gene prioritization tool, external researcher have applied Endeavour to search for disease genes as well. These applications include obesity and Type 2 diabetes (Elbers *et al.*, 2007), pulmonary fibrosis (Tzouvelekis *et al.*, 2007) and cleft lip and cleft palate (Osoegawa *et al.*, 2008). In addition, Adachi *et al.* (2007) used Endeavour to study adipocyte biology and Windelinckx *et al.* (2007) applied Endeavour to the search for genes involved in muscle strength.

The strengths of Endeavour and the two network-based gene prioritization methods (Franke *et al.*, 2006; Lage *et al.*, 2007) are at least partly complementary. While the latter methods can obtain the best performance from data that can easily be represented as a network, Endeavour obtains a higher performance from data that is not that suitable for representation as a network (e.g. text mining, microarray data) and is able to handle more and more heterogeneous data sources. However, it might be that precisely in some of these network-based data sources (e.g. protein-protein interactions), more information that is not explicitly part of the scientific knowledge is present, in contrast to e.g. text-mining the literature, where mostly (although not exclusively) known data can be extracted. Therefore, it would be interesting to combine both approaches to gain increased strength. The modular structure of Endeavour suggests one natural way to achieve this end.

Endeavour, as well as the other gene prioritization methods, are largely limited to protein-coding genes. This effectively ignores a growing number of non-protein-coding genes, such as microRNAs (Bartel, 2004) and other non-coding RNAs. It is becoming increasingly clear that these genes can also play an important role in disease (He *et al.*, 2005; Lu *et al.*, 2005; Calin *et al.*, 2007). Options for the prioritization of these non-coding RNA genes are currently limited because of the limited amount of data available regarding these genes, although specifically for miRNAs this is becoming less of a limitation. Once more data will become available, it would be instructive to also include non-coding RNA genes in our gene prioritization framework.

Finally, Endeavour was initially developed as a prioritization method for human candidate genes and as such all implemented data sources contain information about human genes. The system would benefit from increased versatility and most likely increased performance as well when information can also be integrated cross-species.

1.3 Integrated case-study: macrophage differentiation

As an additional validation of Endeavour, we applied the gene prioritization method in a case-study looking for targets of myeloid differentiation. In this study, we combined a CRM detection method with Endeavour, and our results can be used to assess the performance and to identify the limits of both approaches.

This case-study can be considered in the following broader context: given the sequence of the potential regulatory regions of a gene, can we determine how this gene is expressed? This question has been tackled in a landmark study by Beer and Tavazoie (2004). In this study, the authors confirmed that based on upstream sequences, the expression of yeast genes can be estimated with at least partial confidence. However, in the considerably more complex human genome, this question remains largely unanswered. In this case-study, we tackle the above question in a simplified form: given only sequence information, can we predict genes differentially regulated in a given process?

The process we consider is hematopoietic differentiation, and more precisely the final stages of differentiation of myeloblasts (hematopoietic progenitor cells, from which macrophages and neutrophils can develop) to macrophages. We model this process by incubating the leukemic cell line HL-60 (which is arrested at approximately the myeloblast stage) with phorbol 12-myristate 13-acetate (PMA, also called TPA), which will induce differentiation towards the macrophage lineage.

As shown in figure 1, we first used the Type II CRM detection method ModuleSearcher (Aerts *et al.*, 2003b, 2004) to construct a transcriptional regulatory model of 18 genes known to be up-regulated during HL-60 differentiation (Tamayo *et al.*, 1999). Subsequently, we applied the Type I CRM detection method ModuleScanner (Aerts *et al.*, 2003b) to scan the human genome for target genes of the TRM. As candidate targets of the TRM, we selected the 100 best scoring putative *cis*-regulatory modules. Of these 100, nine were located near genes used for building the TRM, while 91 were located near 90 distinct new target genes (near one new target genes, two high scoring CRMs were predicted).

We evaluated the predictive accuracy of this approach by measuring the

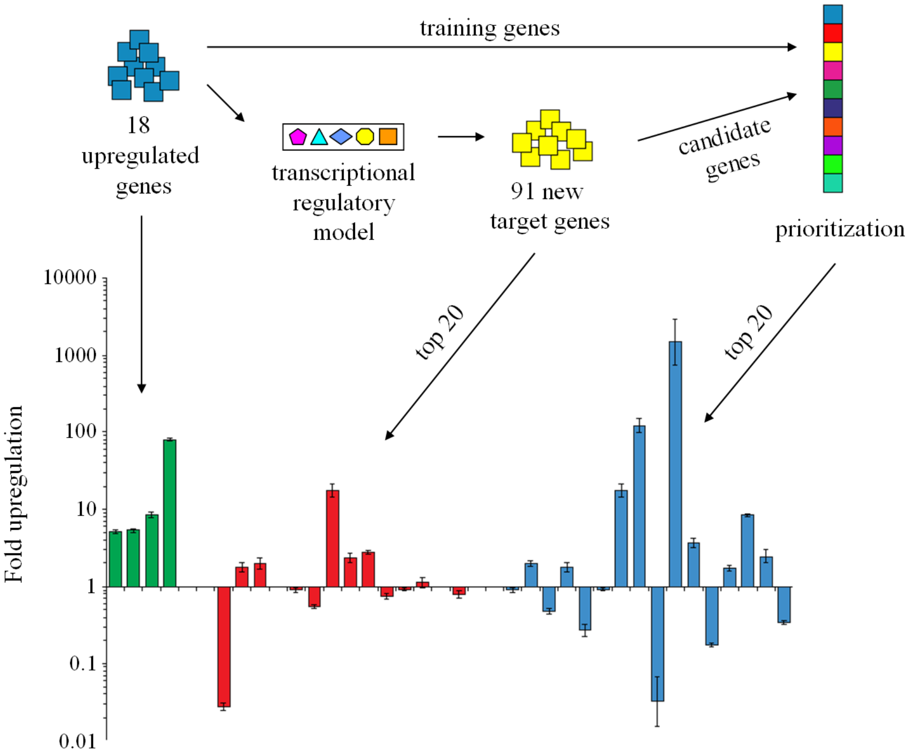


Figure 1: Overview of the procedure integrating computational *cis*-regulatory module detection and Endeavour, to predict genes differentially regulated during HL-60 differentiation.

expression difference in differentiated versus undifferentiated HL-60 cells, for the 20 best scoring new target genes by real-time quantitative RT-PCR. The expression of Eya1, the gene showing two high scoring CRMs, was assayed as well.

By non-quantitative PCR, we could confirm expression of 16 of the 20 genes in differentiated HL-60 cells (data not shown). Although reliable data about the number of genes expressed in a specific cell type is difficult to obtain, available data suggests that less than 50 % of the genes in the genome is expressed in macrophages or HL-60 cells (Velculescu *et al.*, 1999; Tamayo *et al.*, 1999; Stegmaier *et al.*, 2004). As the TRM was trained using genes expressed in differentiated HL-60 cells, this enrichment of expressed genes among the predicted new target genes suggests that our procedure was at least partly successful in predicting additional genes expressed in differentiated HL-60 cells.

The results of our quantitative measurement of the relative expression levels in differentiated versus undifferentiated HL-60 cells (figure 1) indicate that 7 of the 14 genes that could be measured quantitatively were found to be differentially expressed (up- or down-regulated). Together with the data from Tamayo *et al.* (1999), which indicates that about 15 % of the genes on their microarray chip are differentially expressed in this system, this shows that our predicted new target genes are indeed enriched in differentially expressed genes. This is further strengthened by the observation that Eya1, the one gene for which two CRMs were predicted, was found to be up-regulated 29.9 fold. Although these results indicate that the transcriptional regulation of the predicted new target genes show similarities to that of the genes used for training the model, there are still clear differences. Most notably, while all of the training genes were measured by real-time quantitative PCR were up-regulated at least 3-fold, only 2 of the measured new target genes showed a higher than 3 fold difference (figure 1).

Aiming to increase the performance of new target gene selection, we used our Endeavour gene prioritization method to prioritize the predicted new target genes (figure 1). We then measured the expression difference of the 20 genes at the top of the prioritized list, by real-time quantitative RT-PCR. Our results showed that of the 16 genes that could be measured quantitatively, 8 were found to be differentially expressed by more than 3 fold (figure 1). This suggests that among the predicted new target genes, a number of genes with an expression pattern similar to the genes used for constructing the TRM can be found, and that Endeavour is able to select these correct targets.

2 Towards a better stratification and understanding of cancer

Systems biology methods, among which prominently microarray expression profiling and associated analysis methods, have contributed significantly to the stratification and outcome prediction of cancer, and to the understanding of the mechanisms underlying cancer (Perou *et al.*, 2000; Yeoh *et al.*, 2002; Rosenwald *et al.*, 2002; Lamb *et al.*, 2003; Dave *et al.*, 2004; Carrasco *et al.*, 2006; Pujana *et al.*, 2007). In this work, we used microarray expression profiling and associated systems biology analysis methods to gain insight into the role of the microenvironment in two specific subtypes of lymphoma. In another study regarding breast cancer, we investigated the effect of polysomy 17 on methods for testing HER2 amplification, and we correlated this with clinicopathological parameters. Although the methods we applied here were primarily statistical (instead of systems biology methods), the goal is similar: to gain a better understanding of cancer.

2.1 Microarray expression profiling to gain insight into the lymphoma microenvironment

We performed microarray expression profiling, comparing two lymphoma entities with a prominent microenvironment and a markedly different prognosis: the indolent nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL) and the aggressive T cell/histiocyte rich B cell lymphoma (THRLBCL). Our results confirmed clearly different expression profiles of both lymphomas. The

NLPHL expression profile contained mainly B cell genes, consistent with the B cell rich composition of the lymph node derived stroma in this lymphoma. In the THRLBCL expression profile, genes related to macrophages/histiocytes were prominently present. Genes related to T cells were nearly completely absent in both expression profiles, most likely because our samples of both lymphomas contained an approximately equal T cell component.

Comparing the gene expression signatures to those obtained in other studies, we found that the THRLBCL signature shows statistically significant overlap with the signature Monti *et al.* (2005) found to be related to the host response in a subgroup of diffuse large B cell lymphoma, enriched in THRLBCL cases. In addition, this THRLBCL signature showed similarities to the signature of an unfavorable immune response Dave *et al.* (2004) observed in a subset of follicular lymphomas.

Detailed analysis of the THRLBCL signature revealed up-regulation of IFN-γ and genes up-regulated by IFN-γ in macrophages, in particular CCL8 and IDO, as well as VSIG4 and multiple toll-like receptors and scavenger receptors. Correlating this THRLBCL signature with literature, we hypothesize that the activation IFN-γ leads to the recruitment of histiocytes/macrophages in THRLBCL, which are subsequently activated by CCL8. These macrophages (along with possibly additional immune cells) produce (i) IDO and VSIG4 which drastically increases tumour tolerance and (ii) toll-like receptors, scavenger receptors and VSIG4, indicative of innate immunity. Altogether, the THRLBCL gene expression signature reflects the recruitment and activation of histiocytes/macrophages, innate immune responses and a tumour tolerogenic microenvironment. These particular characteristics may explain the bad prognosis of these lymphoma patients.

Our expression profiling experiment provided mechanistic hypotheses regarding the microenvironment in the aggressive THRLBCL. Similar approaches may be useful as well in elucidating the pathogenetic mechanisms of other cancer entities. However, these analyses are often complicated by heterogeneity, even in very specific cancers entities. Indeed, we believe that precisely the choice of two very specific (but also rare), carefully selected, relatively homogeneous lymphoma entities has potentiated these mechanistic insights. We therefore hypothesize that after detailed stratification of more heterogeneous cancer entities, and careful selection of specific subentities, similar systems biology analyses may lead to additional mechanistic insights.

2.2 Stratification of breast cancer by HER2 amplification status and presence of polysomy 17

We studied the effect of polysomy 17 on the interpretation of different HER2 testing methods in breast cancer, and we investigated whether polysomy 17 breast cancers share biological characteristics with HER2 amplified breast cancers. We found that all cases with an equivocal HER2 result by FISH (absolute HER2 copy number of the HER2/chromosome 17 ratio) are polysomic for chromosome 17. Polysomy 17 can occur on its own or in combination with HER2 gene amplification. Polysomy 17 without HER2 amplification was not associated with HER2 overexpression, neither on the mRNA level nor on the protein level. In addition, in contrast to HER2 amplification, polysomy 17 was not associated with high tumour grade, hormone receptor negativity or reduced disease-free survival. Based on these results, we state that tumours showing polysomy 17 without HER2 amplification are more similar to HER2 negative than HER2 positive tumours.

2.3 Clinicopathological implications

The findings of our breast cancer study could have important clinical implications. As polysomy 17 is clinicopathologically distinct from true HER2 gene amplification, it may not be likely that specific HER2-targeted therapies such as trastuzumab will be effective in these tumours. Indeed, polysomy 17 tumours without HER2 amplification do not show HER2 overexpression by immunohistochemistry, while it is in these tumours with HER2 overexpression that the best therapeutic response rates have been obtained (Seidman *et al.*, 2001; Hofmann *et al.*, 2008). Whether or not polysomy 17 tumours benefit from HER2-targeted therapy should be further investigated by phase III trials.

The tumour tolerogenic immune response we observed in the expression signature of THRLBCL may offer potential targets for therapy. Specifically, blocking the production and/or the function of CCL8, IFN-γ, and in particular IDO and VSIG4 may abrogate the induction of tumour tolerance. It is encouraging that inhibitors to target IDO and IFN-γ are available (Muller and Scherle, 2006; Sigidin *et al.*, 2001).

By selecting the three most significantly differentially expressed genes from our THRLBCL and NLPHL expression signatures, we constructed a real-time quantitative RT-PCR classifier that can discriminate between both lymphomas in a set of additional cases. In a small pilot study, we applied this classifier to diffuse large B cell lymphoma (DLBCL) cases (data not shown). Our results indicate that most DLBCL cases could be discriminated from THRLBCL, but a subgroup of DLBCL showed a similar (three-gene) profile to THRLBCL. Upon morphological revision, these DLBCL cases also appeared to show a T cell and histiocyte rich microenvironment (although contrary to real THRLBCL cases, the clonal tumor cells still represent the majority of the tumor cell mass in these cases). It would therefore be interesting to further explore if this subset of DLBCL also shows a similar tolerogenic immune response, and hence might also be targeted by IDO or IFN-γ inhibitors.

3 Long term perspectives

3.1 Systems biology methods for the identification of regulatory regions

The large-scale detection of *cis*-regulatory modules in the human genome is a complex and largely unsolved question. Although we expect a high boost in performance once protein binding microarrays (Mukherjee *et al.*, 2004) identify position weight matrices for most human transcription factors, we still question if this will be sufficient to decisively unravel the gene regulatory code. We believe the large-scale identification of transcriptional regulatory sequences will remain a major challenge for at least the coming decade. It is unclear to us whether the solution to this problem will eventually come from the computational or the experimental field, although we would hazard to predict that both field will contribute synergistically.

3.2 Systems biology methods for disease gene identification

Gene prioritization methods such as Endeavour and both network-based methods (Franke *et al.*, 2006; Lage *et al.*, 2007) are well performing systems biology methods that make effective use of the plethora of data in the public domain. As such, we believe these data-fusion methods represent a breakthrough compared to earlier gene prioritization methods based on only one or two data sources. As discussed above, increased data integration, as well as cross-species data integration will likely deliver another boost in performance. In addition, the quantity as well as the diversity of available data will continue to rise exponentially, again increasing the power of gene prioritization methods.

Finally, we believe the combination of the CGH-array technique to identify candidate genes, and gene prioritization methods to prioritize those candidate genes, will prove to be a powerful method to identify novel disease genes, both for inheritable disorders and for cancer.

3.3 Systems biology methods for disease mechanism elucidation

For the identification of disease causing mechanisms, multiple systems biology methods exist. However, there is little consensus on how to tackle this problem. Microarray gene expression profiling has received a lot of attention in this regards, yet after the initial successes, many scientist have become more sceptical, given the difficulties associated with extracting the real information from the overflow of data generated by this technique. However, we believe that microarrays still have a great potential for the understanding of disease mechanisms, when experiments are carefully designed. Firstly, focussing to very specific disease subtypes or carefully stratified cancer entities allows asking very specific questions and may hence deliver useful answers. Secondly, we believe the combination of microarrays with other (computational or wet lab) systems biology approaches will prove very powerful. In this regard it is instructive to mention the Connectivity Map (Lamb *et al.*, 2006), as well as gene prioritization methods. Intelligent integration with other data, at e.g. the DNA level (array-CGH, SNP microarrays) and with expert knowledge may be the key to success here as well.

3.4 Data-driven identification of disease genes and mechanisms

In the previous sections as well as throughout this work, we discussed the identification of disease genes and disease mechanisms as a question-drive process. Here, we used both generated and public domain data to answer specific questions. The exponential rise in high throughput data may also make it possible to address these issues in general through a data-driven process. E.g. because of advances in sequencing technology, it is not unthinkable that the complete genome sequence of millions of individuals will become available within the next few decades. Datamining this information is expected to be very instructive in linking phenotypes to genotypes. We believe that (in analogy with microarrays) the efficient extraction of information may prove to be the key challenge.

Bibliography

Adachi, J., Kumar, C., Zhang, Y. and Mann, M. (2007). In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol. Cell Proteomics*, 6:1257–1273.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. *et al.* (2000). The genome sequence of Drosophila melanogaster. *Science*, 287:2185–2195.

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. and Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55.

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y. and De Moor, B. (2003a). Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31:1753–1764.

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

Aerts, S., Van Loo, P., Moreau, Y. and De Moor, B. (2004). A genetic algorithm for the detection of new *cis*-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20:1974–1976.

Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y. and De Moor, B. (2005). TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res*, 33:393–396.

Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. and De Moor, B. (2003b). Computational detection of *cis*-regulatory modules. *Bioinformatics*, 19 Suppl 2: II5–II14.

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. and Staudt, L. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

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

Bader, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T. and Hogue, C. (2001). BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 29:242–245.

Bailey, T. L. and Noble, W. S. (2003). Searching for statistically significant regulatory modules. *Bioinformatics*, 19 Suppl 2:II16–II25.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116:281–297.

Beer, M. A. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, 117:185–198.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci U S A*, 99:757–762.

Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B. and Celniker, S. E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol*, 5:R61.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T. *et al.* (2004). An overview of Ensembl. *Genome Res*, 14:925–928.

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., Coulombe, B. and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, 16:656–668.

Bluthgen, N., Kielbasa, S. M. and Herzel, H. (2005). Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res*, 33:272–279. Evaluation Studies.

Bullinger, L., Dohner, K., Bair, E., Frohling, S., Schlenk, R. F., Tibshirani, R., Dohner, H. and Pollack, J. R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*, 350:1605–1616.

Bussemaker, H., Li, H. and Siggia, E. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.*, 27:167–171.

Calin, G., Liu, C., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E., Wojcik, S., Shimizu, M., Tili, E., Rossi, S., Taccioli, C., Pichiorri, F., Liu, X., Zupo, S., Herlea, V., Gramantieri, L., Lanza, G., Alder, H., Rassenti, L., Volinia, S., Schmittgen, T., Kipps, T., Negrini, M. and Croce, C. (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, 12:215–229.

Carrasco, D., Tonon, G., Huang, Y., Zhang, Y., Sinha, R., Feng, B., Stewart, J., Zhan, F., Khatry, D., Protopopova, M., Protopopov, A., Sukhdeo, K., Hanamura, I., Stephens, O., Barlogie, B., Anderson, K., Chin, L., Shaughnessy, J., Brennan, C. and Depinho, R. (2006). High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell*, 9:313–325.

C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282:2012–2018.

Das, D., Banerjee, N. and Zhang, M. Q. (2004). Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A*, 101:16234–16239.

Dave, S. S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R. D., Chan, W. C., Fisher, R. I., Braziel, R. M., Rimsza, L. M., Grogan, T. M., Miller, T. P., LeBlanc, M., Greiner, T. C., Weisenburger, D. D., Lynch, J. C., Vose, J., Armitage, J. O., Smeland, E. B., Kvaloy, S., Holte, H., Delabie, J., Connors, J. M., Lansdorp, P. M., Ouyang, Q., Lister, T. A., Davies, A. J., Norton, A. J., Muller-Hermelink, H. K., Ott, G., Campo, E., Montserrat, E., Wilson, W. H., Jaffe, E. S., Simon, R., Yang, L., Powell, J., Zhao, H., Goldschmidt, N., Chiorazzi, M. and Staudt, L. M. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N Engl J Med*, 351:2159–2169.

Elbers, C., Onland-Moret, N., Franke, L., Niehoff, A., van der Schouw, Y. and Wijmenga, C. (2007). A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol. Metab.*, 18:19–26.

Elnitski, L., Hardison, R., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M., Schwartz, S., Miller, W. and Chiaromonte, F. (2003). Distinguishing regulatory DNA from neutral sites. *Genome Res.*, 13:64–72.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B. and Merrick, J. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:496–512.

Fortunel, N., Otu, H., Ng, H., Chen, J., Mu, X., Chevassut, T., Li, X., Joseph, M., Bailey, C., Hatzfeld, J., Hatzfeld, A., Usta, F., Vega, V., Long, P., Libermann, T. and Lim, B. (2003). Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science*, 302:393.

Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M. and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78:1011–1025.

Freudenberg, J. and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18 Suppl 2:110–115. Evaluation Studies.

Frith, M. C., Li, M. C. and Weng, Z. (2003). Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*, 31:3666–3668.

Frith, M. C., Spouge, J. L., Hansen, U. and Weng, Z. (2002). Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*, 30:3214–3224.

Frith, M., Hansen, U. and Weng, Z. (2001). Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17:878–889.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E. *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428:493–521.

Grad, Y. H., Roth, F. P., Halfon, M. S. and Church, G. M. (2004). Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D.pseudoobscura*. *Bioinformatics*, 20:2738–2750.

GuhaThakurta, D. and Stormo, G. D. (2001). Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608–621.

Gupta, M. and Liu, J. S. (2005). De novo *cis*-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci U S A*, 102:7079–7084.

Halfon, M. S., Grad, Y., Church, G. M. and Michelson, A. M. (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res*, 12:1019–1028.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E. and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124:47–59.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32: 258–261.

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. and Hammond, S. M. (2005). A microRNA polycistron as a potential human oncogene. *Nature*, 435:828–833.

Hofmann, M., Stoss, O., Gaiser, T., Kneitz, H., Heinmller, P., Gutjahr, T., Kaufmann, M., Henkel, T. and Rschoff, J. (2008). Central HER2 IHC and FISH analysis in a trastuzumab (Herceptin) phase II monotherapy study: assessment of test sensitivity and impact of chromosome 17 polysomy. *J. Clin. Pathol.*, 61:89–94.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945.

Jegga, A. G., Sherwood, S. P., Carman, J. W., Pinski, A. T., Phillips, J. L., Pestian, J. P. and Aronow, B. J. (2002). Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res*, 12:1408–1417.

Johansson, O., Alkema, W., Wasserman, W. W. and Lagergren, J. (2003). Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19 Suppl 1:169–176.

Kent, W. J., Hsu, F., Karolchik, D., Kuhn, R. M., Clawson, H., Trumbower, H. and Haussler, D. (2005). Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res*, 15:737–741.

King, D. C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W. and Hardison, R. C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res*, 15:1051–1060.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R. and Chiaromonte, F. (2004). Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res*, 14:700–707.

Kreiman, G. (2004). Identification of sparsely distributed clusters of *cis*-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res*, 32: 2889–2900. Evaluation Studies.

Krivan, W. and Wasserman, W. W. (2001). A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*, 11:1559–1566.

Lage, K., Karlberg, E., Størling, Z., Olason, P., Pedersen, A., Rigina, O., Hinsby, A., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y. and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, 25:309–316.

Lamb, J., Crawford, E., Peck, D., Modell, J., Blat, I., Wrobel, M., Lerner, J., Brunet, J., Subramanian, A., Ross, K., Reich, M., Hieronymus, H., Wei, G., Armstrong, S., Haggarty, S., Clemons, P., Wei, R., Carr, S., Lander, E. and Golub, T. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313:1929–1935.

Lamb, J., Ramaswamy, S., Ford, H. L., Contreras, B., Martinez, R. V., Kittrell, F. S., Zahnow, C. A., Patterson, N., Golub, T. R. and Ewen, M. E. (2003). A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*, 114:323–334.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Lee, J., Heo, J., Libbrecht, L., Chu, I., Kaposi-Novak, P., Calvisi, D., Mikaelyan, A., Roberts, L., Demetris, A., Sun, Z., Nevens, F., Roskams, T. and Thorgeirsson, S. (2006). A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nat. Med.*, 12: 410–416.

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

Lopez-Bigas, N. and Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32: 3108–3114.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R. and Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838.

Monti, S., Savage, K. J., Kutok, J. L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R. C. T., Dal Cin, P., Ladd, C., Pinkus, G. S., Salles, G., Harris, N. L., Dalla-Favera, R., Habermann, T. M., Aster, J. C., Golub, T. R. and Shipp, M. A. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105:1851–1861.

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A. and Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36: 1331–1339.

Muller, A. and Scherle, P. (2006). Targeting the mechanisms of tumoral immune tolerance with small-molecule inhibitors. *Nat. Rev. Cancer*, 6:613–625.

Osoegawa, K., Vessere, G., Utami, K., Mansilla, M., Johnson, M., Riley, B., L'Heureux, J., Pfundt, R., Staaf, J., van der Vliet, W., Lidral, A., Schoenmakers, E., Borg, A., Schutte, B., Lammer, E., Murray, J. and de Jong, P. (2008). Identification of novel candidate genes associated with cleft lip and palate using array comparative genomic hybridisation. *J. Med. Genet.*, 45: 81–86.

Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31:316–319.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O. and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406:747–752.

Philippakis, A. A., Busser, B. W., Gisselbrecht, S. S., He, F. S., Estrada, B., Michelson, A. M. and Bulyk, M. L. (2006). Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. *PLoS Comput Biol*, 2:e53.

Philippakis, A. A., He, F. S. and Bulyk, M. L. (2005). Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac Symp Biocomput*, p. 519–530.

Pujana, M., Han, J., Starita, L., Stevens, K., Tewari, M., Ahn, J., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W., Rual, J., Levine, D., Rozek, L., Gelman, R., Gunsalus, K., Greenberg, R., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Solé, X., Hernández, P., Lázaro, C., Nathanson, K., Weber, B., Cusick, M., Hill, D., Offit, K., Livingston, D., Gruber, S., Parvin, J. and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.*, 39:1338–1349.

Rajewsky, N., Vergassola, M., Gaul, U. and Siggia, E. D. (2002). Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, 3:30.

Ramaswamy, S., Ross, K., Lander, E. and Golub, T. (2003). A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, 33:49–54.

Rebeiz, M., Reeves, N. L. and Posakony, J. W. (2002). SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc Natl Acad Sci U S A*, 99:9888–9893.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6:1–6.

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., Gascoyne, R., Muller-Hermelink, H., Smeland, E., Giltnane, J., Hurt, E., Zhao, H., Averett, L., Yang, L., Wilson, W., Jaffe, E., Simon, R., Klausner, R., Powell, J., Duffey, P., Longo, D., Greiner, T., Weisenburger, D., Sanger, W., Dave, B., Lynch, J., Vose, J., Armitage, J., Montserrat, E., López-Guillermo, A., Grogan, T., Miller, T., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, 346:1937–1947.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32 Database issue:91–94.

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

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D. and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*, 2:E271.

Segal, E. and Sharan, R. (2005). A discriminative model for identifying spatial cis-regulatory modules. *J Comput Biol*, 12:822–834.

Seidman, A., Fornier, M., Esteva, F., Tan, L., Kaptain, S., Bach, A., Panageas, K., Arroyo, C., Valero, V., Currie, V., Gilewski, T., Theodoulou, M., Moynahan, M., Moasser, M., Sklarin, N., Dickler, M., D'Andrea, G., Cristofanilli, M., Rivera, E., Hortobagyi, G., Norton, L. and Hudis, C. (2001). Weekly trastuzumab and paclitaxel therapy for metastatic breast cancer with analysis of efficacy by HER2 immunophenotype and gene amplification. *J. Clin. Oncol.*, 19:2587–2595.

Sharan, R., Ben-Hur, A., Loots, G. G. and Ovcharenko, I. (2004). CREME: *Cis*-Regulatory Module Explorer for the human genome. *Nucleic Acids Res*, 32:253–256.

Sharan, R., Ovcharenko, I., Ben-Hur, A. and Karp, R. M. (2003). CREME: a framework for identifying *cis*-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19 Suppl 1:283–291.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T., Mesirov, J., Neuberg, D., Lander, E., Aster, J. and Golub, T.

(2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, 8:68–74.

Sigidin, Y., Loukina, G., Skurkovich, B. and Skurkovich, S. (2001). Randomized, double-blind trial of anti-interferon-gamma antibodies in rheumatoid arthritis. *Scand. J. Rheumatol.*, 30:203–207.

Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U. and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, 5:129.

Sinha, S., van Nimwegen, E. and Siggia, E. D. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1:292–301. Evaluation Studies.

Slamon, D., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J. and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.*, 344:783–792.

Smith, A. D., Sumazin, P., Xuan, Z. and Zhang, M. Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103:6275–6280.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

Son, C. G., Bilke, S., Davis, S., Greer, B. T., Wei, J. S., Whiteford, C. C., Chen, Q.-R., Cenacchi, N. and Khan, J. (2005). Database of mRNA gene expression profiles of multiple human organs. *Genome Res*, 15:443–450.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E. and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98: 10869–10874.

Stegmaier, K., Ross, K. N., Colavito, S. A., O'Malley, S., Stockwell, B. R. and Golub, T. R. (2004). Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat Genet*, 36:257–263.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R. and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101:6062–6067.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102:15545–15550.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96:2907–2912.

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

Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S. and Lawrence, C. E. (2004). Decoding Human Regulatory Circuits. *Genome Res*, 14:1967–1974.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B. and Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*, 33:1544–1552. Evaluation Studies.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23:137–144.

Turner, F. S., Clutterbuck, D. R. and Semple, C. A. M. (2003). POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4:R75.

Tzouvelekis, A., Harokopos, V., Paparountas, T., Oikonomou, N., Chatziioannou, A., Vilaras, G., Tsiambas, E., Karameris, A., Bouros, D. and Aidinis, V. (2007). Comparative expression profiling in pulmonary fibrosis suggests a role of hypoxia-inducible factor-1alpha in disease pathogenesis. *Am. J. Respir. Crit. Care Med.*, 176:1108–1119.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.

Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., Cook, B. P., Dufault, M. R., Ferguson, A. T., Gao, Y., He, T. C., Hermeking, H., Hiraldo, S. K., Hwang, P. M., Lopez, M. A., Luderer, H. F., Mathews, B., Petroziello, J. M., Polyak, K., Zawel, L. and Kinzler, K. W. (1999). Analysis of human transcriptomes. *Nat Genet*, 23:387–388. Letter.

Vogel, C., Cobleigh, M., Tripathy, D., Gutheil, J., Harris, L., Fehrenbacher, L., Slamon, D., Murphy, M., Novotny, W., Burchmore, M., Shak, S., Stewart, S. and Press, M. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.*, 20:719–726.

von Mering, C., Jensen, L., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B. and Bork, P. (2007). STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, 35:D358–362.

Wasserman, W. W. and Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278:167–181.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562.

Windelinckx, A., Vlietinck, R., Aerssens, J., Beunen, G. and Thomis, M. (2007). Selection of genes and single nucleotide polymorphisms for fine mapping starting from a broad linkage region. *Twin Res Hum Genet*, 10:871–885.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W. E., Naeve, C., Wong, L. and Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143.

Yıldırım, M., Goh, K., Cusick, M., Barabási, A. and Vidal, M. (2007). Drug-target network. *Nat. Biotechnol.*, 25:1119–1126.

Zhou, Q. and Wong, W. H. (2004). CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101:12114–12119.

Curriculum Vitae

Peter Van Loo was born in Herentals on March 3, 1980. He started studying Engineering at the Katholieke Universiteit Leuven in 1998 and combined this with Bio-Engineering studies from 1999 onwards. He obtained a Master of Science in Electrotechnical Engineering in 2003 and a Master of Science in Engineering in Cell and Gene Technology in 2004. In October 2004, he started his PhD at the Katholieke Universiteit Leuven under the supervision of prof. Dr. Peter Marynen, prof. Dr. Bart De Moor and prof. Dr. Chris De Wolf-Peeters.

List of publications

Published or in press

1. Aerts, S., **Van Loo, P.**, Thijs, G., Moreau, Y. and De Moor, B. (2003). Computational detection of *cis*-regulatory modules. *Bioinformatics*, 19 Suppl 2:II5-II14.

2. Aerts, S., **Van Loo, P.**, Moreau Y. and De Moor, B. (2004). A genetic algorithm for the detection of new *cis*-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20:1974-1976.

3. Aerts, S., **Van Loo, P.**, Thijs, G., Mayer, H., de Martin, R., Moreau, Y. and De Moor, B. (2005). Toucan: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res*, 33:W393-396.

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

5. Pospisilova, H., Baens, M., Michaux, L., Stul, M., Van Hummelen, P., **Van Loo, P.**, Vermeesch, J., Jarosova, M., Zemanova, Z., Michalova, K., Van den Berghe, I., Alexander, H.D., Hagemeijer, A., Vandenberghe, P., Cools, J., De Wolf-Peeters, C., Marynen, P. and Wlodarska, I. (2007). Interstitial del(14)(q) involving IGH: a novel recurrent aberration in B-NHL. *Leukemia*, 21:2079-2083.

6. Griffith, O.L.$^{\#}$, Montgomery, S.B.$^{\#}$, Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M.C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S.M., Giardine, B., Hooghe, B., **Van Loo, P.**, Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I.J., Robertson, G., Wadelius, C., De Bleser, P., Vlieghe, D., Halfon, M.S., Wasserman, W., Hardison, R., Bergman, C.M. and Jones, S.J.M.; The Open Regulatory Annotation Consortium (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, 36:D107-D113.

7. **Van Loo, P.**, Aerts, S., Thienpont, B., De Moor, B., Moreau, Y. and Marynen, P. (2008). ModuleMiner: improved computational detection of *cis*-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues? *Genome Biology*, 9:R66.

8. Vanden Bempt, I.$^{\#}$, **Van Loo, P.**$^{\#}$, Drijkoningen, M., Neven, P., Smeets, A., Christiaens, M.-R., Paridaens, R. and De Wolf-Peeters, C. (2008). Polysomy 17 in breast cancer: clinicopathological significance and impact on HER2 testing. *Journal of Clinical Oncology*, in press. (#: equal contribution)

9. Tranchevent, L.-C.$^{\#}$, Barriot, R.$^{\#}$, Yu, S., Van Vooren, S., **Van Loo, P.**, Coessens, B., Aerts, S., Hassan, B., De Moor, B. and Moreau, Y. (2008). Endeavour: a web server for gene prioritization in multiple species based on a wide range of genomic data. *Nucleic Acids Research*, in press.

In revision

10. **Van Loo, P.**, Vanhentenrijk, V., Dierickx, D., Vanden Bempt, I., Verhoef, G., Marynen, P., Matthys, P.$^{\#}$ and De Wolf-Peeters, C.$^{\#}$ (2008). T cell/histiocyte rich large B cell lymphoma shows transcriptional features suggestive of a tolerogenic host immune response. *In revision.*