Meetings Reports

MAY 2005

Cell adhesion and signal transduction in cancer

EMBO reports

EMBO reports

meeting report

Cell adhesion and signal transduction in cancer

Conference on Cadherins, Catenins and Cancer

Walter Birchmeier
Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

The Centro Nacional de Investigaciones Oncológicas (CNIO) Cancer Conference on "Cadherins, Catenins and Cancer" was held in Madrid, Spain, between 29 November and 1 December 2004. The conference was organized by A. Cano, H. Clevers, J. Palacios and F. van Roy

Keywords: cancer; development; epithelial-mesenchymal transitions; tumours

EMBO reports (2005) 6, 413–417. doi:10.1038/sj.embor.7400408

Introduction

Cadherin cell-adhesion molecules and their intracellular binding partners, catenins, were discovered in the 1980s (for a review, see Takeichi, 1995). Interest in these molecules was sparked by the discovery that cadherins and, subsequently, catenins are important in the formation and metastasis of carcinomas (Behrens et al, 1989; Berx et al, 1998; Perl et al, 1998; Polakis, 2000). Twenty years later, interest in cadherins and catenins remains high, as seen at the Centro Nacional de Investigaciones Oncológicas (CNIO) Cancer Conference on "Cadherins, Catenins and Cancer" in Madrid. The meeting covered research on all fronts, from the molecular aspects of cadherin and catenin regulation and signalling, to their role in human carcinogenesis and subsequent clinical applications.

E-cadherin in cancer progression

In 1998, germline mutations in the E-cadherin gene were identified in New Zealand Maori families with diffuse-type gastric cancer, which firmly established the importance of E-cadherin in the progression of carcinomas (Guilford et al, 1998). At the Madrid meeting, F. Carneiro (Porto, Portugal) presented new data, which showed that hereditary diffuse gastric cancer is a dominantly inherited familial cancer syndrome that contributes up to 3% of all gastric cancers. Furthermore, E-cadherin mutations are the causal genetic defect for up to 40% of these cases. About 80% of the mutations that have been identified in the E-cadherin gene result in a truncated protein, whereas the remaining 20% are missense mutations. Germline mutations of E-cadherin are distributed throughout the whole gene, in contrast to somatic mutations in sporadic diffuse gastric cancer, which cluster in exons 7-9 (Berx et al, 1998). Inactivation of the remaining normal allele is required for carcinogenesis, and gene silencing through promoter methylation is found in 50% of cases. These studies prompted the International Gastric Cancer Linkage Consortium to develop clinical criteria and guidelines for the management of carriers of E-cadherin mutations, which include the use of intensive screening and prophylactic gastrectomies (Caldas et al, 1999).

Despite the identification of somatic mutations in the E-cadherin gene in human cancers, causal evidence for the involvement of E-cadherin in tumour formation and progression in testable animal models was previously lacking. This has now been elegantly shown by J. Jonkers (Amsterdam, the Netherlands), who introduced a conditional loss-of-function mutation in the E-cadherin gene into mice that carry p53 mutations. Although tissue-specific inactivation of E-cadherin alone did not result in tumour formation, the combined inactivation of E-cadherin and p53 led to the accelerated development of mammary gland and skin tumours. Moreover, loss of E-cadherin induced a phenotypic change from non-invasive to highly invasive mammary gland tumours and a conversion from ductal to lobular carcinomas. The latter in mice resemble their human equivalent in that they show strong stromal involvement, invasion into the surrounding tissue and distant metastases to several organs.

By contrast, the group of G. Christofori (Basel, Switzerland) has used the Rip1-Tag2 transgenic mouse model of pancreatic β-cell carcinogenesis to show that the loss of E-cadherin is causally involved in the transition from benign adenoma to malignant, invasive carcinoma (Perl et al, 1998). The group also reported that the loss of another neural cell-adhesion molecule (NCAM) during Rip1-Tag2 tumour progression results in the formation of lymph-node metastasis. They showed that the loss of NCAM results in the loss of β1 integrin function, which contributes to loss of tissue integrity,

Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany
Tel: +49 30 9406 3800; Fax: +49 30 9406 2656;
E-mail: wbirch@mdc-berlin.de

Submitted 04 March 2005; accepted 25 March 2005
published online 22 April 2005

© 2005 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

upregulated expression of lymphangiogenic vascular endothelial growth factor-C (VEGFC) and VEGFD, and subsequently increases lymphangiogenesis and the lymphogenic dissemination of tumour cells (Crnic et al, 2004). Hence, a new model was proposed, in which there are two distinct pathways to tumour metastasis: the classical pathway of active tumour cell migration and invasion, which is exemplified by the loss of E-cadherin; and a passive pathway in which the increased presence of lymphatic vessels washes tumour cell clusters out into the lymphatics, where they become trapped in the regional lymph nodes.

The regulation of E-cadherin expression

The invasion of carcinoma cells is associated with the loss-of-expression of epithelial genes, such as E-cadherin, and the gain-of-expression of mesenchymal genes (Behrens et al, 1989; Batlle et al, 2000; Cano et al, 2000). This process, which is known as epithelial-mesenchymal transition (EMT), also occurs under strict spatiotemporal control during normal embryonic development (Fig 1; Thiery, 2002). J. Palacios (Madrid, Spain) and G. Berx (Ghent, Belgium) reported that further target genes that are downregulated during EMT include keratins, desmosomal cadherins and components of the basal lamina. By contrast, the upregulated mesenchymal genes include vimentin, secreted protein acidic and rich in cysteine (SPARC), collagens, integrins, protease inhibitors and cytoskeletal components. Altered expression of a selection of these target genes could also be validated in uterine carcinosarcomas and breast cancers.

EMT is a crucial event during tumour metastasis and normal embryonic development, as indicated by two further reports at the Madrid meeting. T. Brabletz (Erlangen, Germany) showed that, in colon carcinomas, tumour cells in the central tumour mass and at the invasive fronts differ from one another: the latter undergo EMT, as characterized by a loss of E-cadherin at the membrane and translocation of β-catenin to the nucleus (Brabletz et al, 2001). This is a transient event, given that metastases in the lymph nodes and liver have adherent tumour cells in tubular structures. Genes that become activated at the tumour invasive front include those encoding the transcription factor Slug, laminin-γ2 and, as reported by A. Ben-ZeÕev (Rehovot, Israel), the cell–cell-adhesion molecule L1 and the disintegrin and metalloprotease ADAM10 (Gavert et al, 2005). L1 expression conferred increased cell motility, cell transformation and tumorigenesis in mice. By contrast, E. Dejana (Milan, Italy) showed that EMT is required for normal heart development during embryogenesis. Endocardial cells in the atrioventricular region undergo transforming growth factor-β (TGF-β)-dependent EMT and invade the underlying cardiac jelly. This process is accompanied by the activation of β-catenin/T-cell factor (TCF)/lymphocyte-enhancer factor (Lef) transcriptional activity, and is blocked by the absence of β-catenin, which implies an interaction between TGF-β and Wingless (Wnt)-signalling pathways in the induction of endothelial-mesenchymal transformation (Liebner et al, 2004).

One of the hallmarks of EMT is the loss of E-cadherin expression. Substantial progress has been made in identifying the mechanisms by which E-cadherin expression is normally regulated. It was reported that different transcription factorsÑ Snail, Slug, δ-crystallin E2-box factor-1/zinc-finger E-box-binding transcriptional repressor (δEF1/ZEB1), Smad-interacting protein-1 (SIP1)/ZEB2 and E12/E47Ñ were responsible for downregulating E-cadherin expression by binding directly to E-boxes in the E-cadherin promoter (Fig 1; Batlle et al, 2000; Cano et al, 2000). The groups of A. Cano (Madrid, Spain) and A. Garc a de Herreros (Barcelona, Spain) reported new findings on the action of the transcription factors Snail and Slug during EMT. They showed that Snail mediates E-cadherin repression by recruitmentÑ through the amino-terminal Snail/growth-factor independent-1 (SNAG) domainÑ of a Sin3A corepressor complex that also contains histone deacetylase-1/2 (HDAC1/2), which leads to chromatin inactivation. Activity of the Snail promoter and levels of Snail mRNA are dependent on the extracellular signal-regulated kinase (ERK), mitogen-activated protein (MAP) kinase and phosphatidylinositol 3-kinase (PI3K)/Akt signalling pathways, as well as Snail itself, which therefore functions in an autoregulatory loop (Bachelder et al, 2005).

β-catenin in cancer and embryogenesis

It is not only cadherins that have been implicated in tumorigenesis and metastasis; their cytoplasmic binding partners, catenins, are also important in cancer formation and progression. β-catenin has been found to be mutationally activated in several tumours (for example, in 90% of hepatoblastomas, 75% of pilomatricomas and 10% of colon cancers; for reviews, see Bienz & Clevers, 2000; Polakis, 2000). Moreover, β-catenin interacts with the tumour suppressor gene adenomatous polyposis coli (APC), which is mutated in familial and sporadic colon carcinomas.

A second armadillo family protein, p120 catenin, which is known to stabilize E-cadherin and thereby prevent its degradation, has also been implicated in cancer progression (for a review, see Reynolds & Roczniak-Ferguson, 2004). A. Reynolds (Nashville, TN, USA) reported the conditional ablation of the p120 catenin gene in mice, using MMTV-Cre or villin-Cre lines. Loss of p120 catenin caused marked E-cadherin deficiency in target tissues, which resulted in notable defects in cell adhesion, polarity and the morphology of epithelial cells. Mammary glands were almost entirely absent, and the animals developed psoriasis-like skin defects that involved massive epidermal proliferation and inflammation. Salivary glands showed hyperplastic growth that was reminiscent of the early stages of tumour progression. These data are consistent with a role of p120 catenin as a tumour modifier and raise the possibility that E-cadherin dysfunction in a subset of human tumours might be associated with p120 downregulation. p120 catenin also binds to the transcription factor Kaiso, which is a BTB/POZ (BR-C, TTK and BAB/Pox virus and zinc finger) family member, and might regulate transcription, similar to β-catenin. P. McCrea (Houston, TX, USA) showed that Kaiso is, indeed, required for gastrulation movements in *Xenopus* embryos, which are under the control of non-canonical Wnt signalling. Interestingly, p120 catenin relieves Kaiso-mediated repression of the non-canonical Wnt11 (Kim et al, 2004). F. van Roy (Ghent, Belgium) reported on a remarkable nucleocytoplasmic shuttling of Kaiso under the influence of the microenvironment of both normal and tumoral tissues.

The identification of the signalling roles of β-catenin in the Wnt signalling pathway provided a breakthrough in the understanding of the roles of cadherin and catenin molecules (Behrens et al, 1996; Bienz & Clevers, 2000; De Robertis et al, 2000; Polakis, 2000). β-catenin was found to have a dual role as a cytoplasmic-interaction partner of cadherins, which is essential for cell adhesions and as a nuclear partner of the TCF/Lef family of

©2005 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

EMBO reports

INTERNATIONAL COVERAGE AND RECOGNITION
cnio

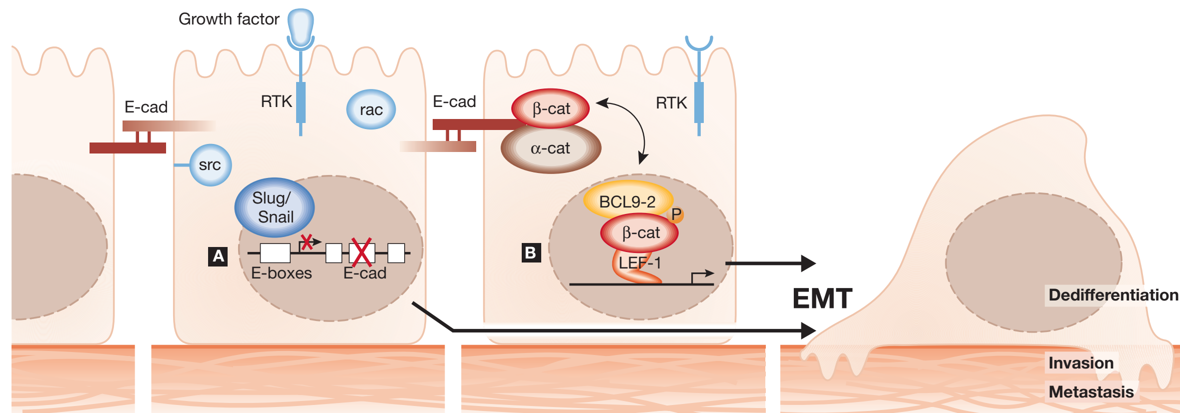


Fig 1 | Epithelial–mesenchymal transition (EMT) occurs late in tumour progression, but also in normal embryonic development (for example, during gastrulation). Epithelial cells lose the expression of epithelial-specific genes, such as E-cadherin (E-cad), that encode cell–cell adhesion molecules, and acquire the expression of mesenchymal genes (see text for examples). Carcinomas represent ~85% of human tumours—particularly those of the colon, breast and lung cancers—and all originate from epithelial cells. EMT causes cells to lose apical-basal polarity (shown on the left) and gain fibroblast-like morphology, high motility and invasive properties (shown on the right). When cells are cultured on gels of extracellular matrix proteins, such as collagen (cross-hatched), the fibroblastoid cells can invade the matrix; this process resembles the invasion and metastasis of carcinoma cells. In recent years, transcription factors (such as Snail and Slug) have been identified that control the expression of E-cadherin by binding directly to E-boxes in the gene promoter (A). Other factors, such as growth factors and their receptors (for example, hepatocyte growth factor (HGF), receptor tyrosine kinases (RTK) and transforming growth factor-β (TGF-β)), the tyrosine kinase src and cytoplasmic G-proteins (such as rac) can also promote EMT indirectly (for a general review on EMT, see Thiery, 2002). (B) β-catenin was found to exert a dual role as an essential cytoplasmic-interaction partner of cadherins, which is essential for cell–cell adhesion, and as a nuclear partner of the T-cell factor (TCF)/lymphocyte-enhancer factor (Lef) family of transcription factors that regulate genes of the canonical Wnt signalling pathway. The switch of β-catenin from its action in cell adhesion to transcriptional control in the nucleus is controlled by binding to BCL9-2, which is the homologue of a human B-cell oncogene product, and is promoted by tyrosine phosphorylation of β-catenin (Brembeck et al, 2004).

transcription factors that directly regulate gene expression in the canonical Wnt signalling pathway.

However, it was not clear how the switch between the roles of β-catenin in cell adhesion and transcription was regulated. B. Gumbiner (Charlottesville, VA, USA) reported that the participation of β-catenin in adhesion and Wnt signalling is dictated by the presence of distinct molecular forms of β-catenin that have different binding properties. A closed form of β-catenin with a folded-back carboxy terminus binds TCFs alone, while an open conformation binds both cadherins and TCFs (Gottardi & Gumbiner, 2004). W. Birchmeier (Berlin, Germany) reported that this switch can be regulated by the binding of β-catenin to BCL9-2 (the homologue of a human B-cell oncogene product, BCL9) or the segment-polarity gene product, legless, in *Drosophila*. β-Catenin/BCL9-2 binding and transcriptional activation is promoted by tyrosine phosphorylation of β-catenin, which competes with α-catenin binding and cell adhesion (Fig 1; Brembeck et al, 2004).

Research has focused on the roles of catenin molecules not only during tumorigenesis, but also during normal embryogenesis. In vertebrate embryogenesis, Wnt/β-catenin signalling participates in dorso-anterior-axis formation, primitive-streak formation and dorsoventral patterning of the mesoderm. During the later stages of development, the Wnt/β-catenin pathway regulates the patterning of the neural tube and brain, limbs, heart, hair, intestine and other organs (for reviews, see De Robertis et al, 2000; Huelsken & Birchmeier, 2001; Moon et al, 2002). R. Kemler (Freiburg, Germany) showed that β-catenin functions in the early mammalian telencephalon predominantly as a mediator of cell adhesion. β-catenin, together with N-cadherin, is localized to adhesion junctions at the apical lining of the neuroepithelium. Ablation of β-catenin in the forebrain using the FoxG1-Cre line leads to a disruption of apical adherens junctions and a breakdown of neuroepithelial structures that result in apoptosis. Consequently, β-catenin-mutant embryos have no forebrain or anterior facial structures. The lack of nuclear β-catenin and the absence of TCF/Lef–β-catenin-dependent transcriptional activity in the forebrain of wild-type mice indicates that the canonical Wnt signalling pathway is not crucial in the development of the early telencephalon.

During development, the Wnt/β-catenin pathway cooperates with several other signalling pathways, such as the TGFβ/bone morphogenetic protein (BMP), Sonic Hedgehog, Notch and Ras/MAP kinase pathways (Beddington & Robertson, 1999; Jessell, 2000; Capdevila & Izpisua-Belmonte, 2001). E. Fuchs (New York, NY, USA) reported that Wnt/β-catenin and BMP/TGFβ signalling cooperate in the generation of hair follicles. TGFβ2 is necessary to induce the transcription factor Snail, and, by doing so, transiently downregulates E-cadherin expression and activates Ras/MAP kinase signalling (Jamora et al, 2005). H. Clevers (Utrecht, the Netherlands) reported that both the Wnt/β-catenin and Notch pathways are crucial in stem-cell specification of the gut. When the Notch pathway was blocked by an intestine-specific deletion of the transcription factor CSL, all proliferative crypt-precursor cells differentiated towards the goblet-cell lineage.

108

©2005 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

Clinical implications

Efforts are underway in several laboratories to apply the knowledge that has been gained on cadherin deficiency and activation of Wnt/β-catenin signalling to the diagnosis and treatment of human conditions. At the Madrid meeting, K.-F. Becker (Munich, Germany) showed that it is possible to raise antibodies against the mutant E-cadherin protein that is present in sporadic diffuse-type gastric cancers (Gamboa-Dominguez et al, 2005). These could be linked to toxins, thereby providing a way in which to selectively kill cells that bear these mutations in culture. Conversely, in a mouse model, radiolabelled antibodies selectively targeted tumour cells that expressed mutant E-cadherin and prolonged the survival of the treated animals. Such personalized therapy could be applied to the 250,000 new gastric cancers that occur annually worldwide.

The group of P. Polakis (San Francisco, CA, USA) has adopted an alterative approach. J. Gonzalez-Sancho (Madrid, Spain) showed that the Wnt pathway inhibitor Dickkopf-1 (DKK1) is a direct β-catenin target and is downregulated in human colon cancer (Niida et al, 2004; Gonzalez-Sancho et al, 2005). Polakis produced a DKK1–Fc–immunoglobulin-G (IgG) fusion protein that is being tested for activity in cancer models that are under the control of Wnts. The purified DKK1 works well in vitro and initial attempts in vivo have focused on the MMTV-Wnt murine mammary gland tumour model. Tumours were transplanted into the cleared fat pads of naive animals and DKK1-Fc was administered daily. A 50% reduction in tumour growth was observed over a 2–3-week period.

Summary and future directions

The Madrid meeting reported clear progress in this field. The merging of cell adhesion with concomitant signal transduction has now become accepted knowledge, and the importance of cadherins and catenins in the various stages of tumour formation and progression has also been firmly established. Breakthroughs in the field have often come unexpectedly, such as the discovery of cadherins by mouse embryologists, the identification of β-catenin/armadillo by Drosophila developmental biologists and the uncovering of the β-catenin/APC connection by cancer biologists. Moreover, scientists in this field have worked with organisms ranging from Hydra, which harbours an impressive collection of Wnt-pathway genes (Kusserow et al, 2005), to humans in which several thousands of tumours have been analyzed for β-catenin mutations (Bienz & Clevers, 2000; Polakis, 2000). The Madrid meeting clearly indicated that the results of this basic science are ready to be applied to human cancer patients and several different approaches are now being tested. Certainly, there has never been a feeling of loneliness in the field during the past 20 years, but rather a sense of healthy competition.

In the future, we expect the generation and testing of various known, and as yet unknown, molecules that interfere with Wnt signalling and cancer progression (for example, see Dihlmann & von Knebel-Doeberitz, 2005). E-cadherin gene repressors might also provide new targets for anti-invasive therapy. Moreover, target genes of EMT in human cancers might be suitable as new anti-cancer targets. Finally, the discovery of crosstalk between the cadherin/catenin system and other signalling pathways and adhesion systems might yield new targets for validation in cancer diagnosis and therapy.

ACKNOWLEDGEMENTS

I thank the participants of the Madrid meeting for their important contributions to this report, and M. Ros rio (Berlin) for helpful discussions and improving the text. Space limitations prevented me from referencing many worthy publications.

REFERENCES

Bachelder RE, Yoon SO, Franci C, de Herreros AG, Mercurio AM (2005) Glycogen synthase kinase-3 is an endogenous inhibitor of Snail transcription: implications for the epithelial-mesenchymal transition. J Cell Biol 168: 29–33

Batlle E, Sancho E, Franci C, Dominguez D, Monfar M, Baulida J, Garcia De Herreros A (2000) The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. Nat Cell Biol 2: 84–89

Beddington RS, Robertson EJ (1999) Axis development and early asymmetry in mammals. Cell 96: 195–209

Behrens J, Mareel MM, van Roy FM, Birchmeier W (1989) Dissecting tumor cell invasion: epithelial cells acquire invasive properties after the loss of uvomorulin-mediated cell–cell adhesion. J Cell Biol 108: 2435–2447

Behrens J, von Kries JP, Kuhl M, Bruhn L, Wedlich D, Grosschedl R, Birchmeier W (1996) Functional interaction of β-catenin with the transcription factor LEF-1. Nature 382: 638–642

Berx G, Becker KF, Hofler H, van Roy F (1998) Mutations of the human E-cadherin (CDH1) gene. Hum Mutat 12: 226–237

Bienz M, Clevers H (2000) Linking colorectal cancer to Wnt signaling. Cell 103: 311–320

Brabletz T, Jung A, Reu S, Porzner M, Hlubek F, Kunz-Schughart LA, Knuechel R, Kirchner T (2001) Variable β-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. Proc Natl Acad Sci USA 98: 10356–10361

Brembeck FH, Schwarz-Romond T, Bakkers J, Wilhelm S, Hammerschmidt M, Birchmeier W (2004) Essential role of BCL9-2 in the switch between β-catenin's adhesive and transcriptional functions. Genes Dev 18: 2225–2230

Caldas C et al (1999) Familial gastric cancer: overview and guidelines for management. J Med Genet 36: 873–880

Cano A, Perez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, del Barrio MG, Portillo F, Nieto MA (2000) The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nat Cell Biol 2: 76–83

Capdevila J, Izpisua-Belmonte JC (2001) Patterning mechanisms controlling vertebrate limb development. Annu Rev Cell Dev Biol 17: 87–132

Crnic I, Strittmatter K, Cavallaro U, Kopfstein L, Jussila L, Alitalo K, Christofori G (2004) Loss of neural cell adhesion molecule induces tumor metastasis by up-regulating lymphangiogenesis. Cancer Res 64: 8630–8638

De Robertis EM, Larrain J, Oelgeschlager M, Wessely O (2000) The establishment of Spemann's organizer and patterning of the vertebrate embryo. Nat Rev Genet 1: 171–181

Dihlmann S, von Knebel-Doeberitz M (2005) Wnt/β-catenin-pathway as a molecular target for future anti-cancer therapeutics. Int J Cancer 113: 515–524

Gamboa-Dominguez A et al (2005) E-cadherin expression in sporadic gastric cancer from Mexico: exon 8 and 9 deletions are infrequent events associated with poor survival. Hum Pathol 36: 29–35

Gavert N, Conacci-Sorrell M, Gast D, Schneider A, Altevogt P, Brabletz T, Ben-Ze'ev A (2005) L1, a novel target of β-catenin signaling, transforms cells and is expressed at the invasive front of colon cancers. J Cell Biol 168: 633–642

Gonzalez-Sancho JM, Aguilera O, Garcia JM, Pendas-Franco N, Pena C, Cal S, Garcia de Herreros A, Bonilla F, Munoz A (2005) The Wnt antagonist Dickkopf-1 gene is a downstream target of β-catenin/TCF and is downregulated in human colon cancer. Oncogene 24: 1098–1103

Gottardi CJ, Gumbiner BM (2004) Distinct molecular forms of β-catenin are targeted to adhesive or transcriptional complexes. J Cell Biol 167: 339–349

Guilford P, Hopkins J, Harraway J, McLeod M, McLeod N, Harawira P, Taite H, Scoular R, Miller A, Reeve AE (1998) E-cadherin germline mutations in familial gastric cancer. Nature 392: 402–405

Huelsken J, Birchmeier W (2001) New aspects of Wnt signaling pathways in higher vertebrates. Curr Opin Genet Dev 11: 547–553

©2005 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

Jamora C, Lee P, Kocieniewski P, Azhar M, Hosokawa R, Chai Y, Fuchs E (2005) A signaling pathway involving TGFβ2 and snail in hair follicle morphogenesis. PLoS Biol 3: e11
Jessell TM (2000) Neuronal specification in the spinal cord: inductive signals and transcriptional codes. Nat Rev Genet 1: 20–29
Kim SW, Park JI, Spring CM, Sater AK, Ji H, Otchere AA, Daniel JM, McCrea PD (2004) Non-canonical Wnt signals are modulated by the Kaiso transcriptional repressor and p120-catenin. Nat Cell Biol 6: 1212–1220
Kusserow A et al (2005) Unexpected complexity of the Wnt gene family in a sea anemone. Nature 433: 156–160
Liebner S, Cattelino A, Gallini R, Rudini N, Iurlaro M, Piccolo S, Dejana E (2004) β-catenin is required for endothelial-mesenchymal transformation during heart cushion development in the mouse. J Cell Biol 166: 359–367
Moon RT, Bowerman B, Boutros M, Perrimon N (2002) The promise and perils of Wnt signaling through β-catenin. Science 296: 1644–1646
Niida A, Hiroko T, Kasai M, Furukawa Y, Nakamura Y, Suzuki Y, Sugano S, Akiyama T (2004) DKK1, a negative regulator of Wnt signaling, is a target of the β-catenin/TCF pathway. Oncogene 23: 8520–8526
Perl AK, Wilgenbus P, Dahl U, Semb H, Christofori G (1998) A causal role for E-cadherin in the transition from adenoma to carcinoma. Nature 392: 190–193
Polakis P (2000) Wnt signaling and cancer. Genes Dev 14: 1837–1851
Reynolds AB, Roczniak-Ferguson A (2004) Emerging roles for p120-catenin in cell adhesion and cancer. Oncogene 23: 7947–7956
Takeichi M (1995) Morphogenetic roles of classic cadherins. Curr Opin Cell Biol 7: 619–627
Thiery JP (2002) Epithelial-mesenchymal transitions in tumour progression. Nat Rev Cancer 2: 442–454

Walter Birchmeier

©2005 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

MAY 2004

CNIO Meeting report "From microarray data to results"

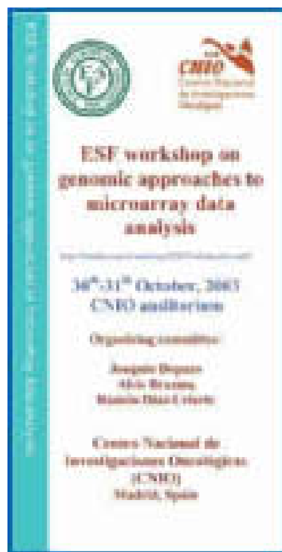
Workshop on Genomic Approaches to Microarray Data Analysis

From microarray data to results

Workshop on Genomic Approaches to Microarray Data Analysis

Thomas Schlitt[1,2+] & Patrick Kemmeren[1,3]

[1]EMBL-EBI, Cambridge, UK, [2]British Antarctic Survey, Cambridge, UK, and [3]Division of Biomedical Genetics, UMC Utrecht, Utrecht, The Netherlands



This European Science Foundation (ESF) exploratory workshop was held at the Centro Nacional de Investigaciones Oncologicas (CNIO) in Madrid on 30 and 31 October 2003, and was organized by J. Dopazo, R. Díaz-Uriarte and A. Brazma.

Keywords: data analysis; genomics; microarray; statistics

EMBO reports (2004) 5, 459–463. doi:10.1038/sj.embor.7400156

Introduction

The main focus of this workshop was to discuss the challenges involved in the collection, storage and analysis of the large amounts of biological data that are being produced by microarray technology.

[1]European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2]British Antarctic Survey, Natural Environment Research Council, High Cross, Madingley Road, Cambridge CB3 0ET, UK
[3]Genomics Laboratory, Division of Biomedical Genetics, UMC Utrecht, PO Box 85060, 3508 AB Utrecht, The Netherlands
+Corresponding author. Tel: +44 1223 221656; Fax: +44 1223 221259;
E-mail: schlitt@ebi.ac.uk

Submitted 16 February 2004; accepted 26 March 2004; published online 23 April 2004

Microarray experiments can be conceptually subdivided into material- and data-processing steps. During material processing, important information needs to be recorded, such as array design, experimental conditions and sample treatment, to enable meaningful data analysis and biological interpretation. This workshop concentrated on the subsequent microarray data processing, which can be further divided into data preprocessing, such as normalization and filtering, data-analysis steps and the biological interpretation of the results.

The first steps in microarray data preprocessing involve image scanning, and include spot finding and the selection of good quality spots. Next, data-normalization steps are necessary to correct unavoidable experimental variations, such as differences in sample preparation, dye incorporation and hybridization efficiencies. These variations are not owing to differences in gene expression in the original samples and, therefore, need to be corrected before data analysis can be carried out. Such analysis might include various methods to identify genes that are differentially expressed or conditions (cell-culture treatments, diseases and so on) that result in similar changes in gene expression. Some normalization or data-analysis methods require special arrangements, such as a particular array design. Therefore, both material- and data-processing steps need to be considered at the early stages of a microarray experiment. The biological interpretation of the data is facilitated by various tools, which place the analysis results into context with existing biological knowledge, such as the scientific literature or sequence data. Efforts to unify and standardize the way in which information is recorded are making the interpretation of large-scale experiments easier. Finally, the integration of biological information from various sources, such as large-scale data sets produced by various experimental techniques, provides a valuable platform for the exploration of regulatory networks. All of these topics were discussed during the workshop and a summary of the research that was presented is given here.

Microarray data sharing

It is important that all of the information about a microarray experiment is recorded systematically, so that meaningful data sets can be generated. A. Brazma (Cambridge, UK) showed that microarray

©2004 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

data sets can be complex, so it is of particular importance to establish standards to enable microarray data to be shared efficiently. Such a standard has been defined by the Microarray Gene Expression Data Society (MGED; http://www.mged.org) and is now widely accepted by biological journals (Brazma et al, 2001). The European Bioinformatics Institute (EBI) offers a public data repository known as ArrayExpress that conforms to the MGED requirements and stores information about microarray experiments, including material-processing aspects such as experimental design, sample treatment and array designs. Furthermore, the EBI provides two web-based tools that allow scientists to analyse microarray data (Expression Profiler) and to submit microarray data to ArrayExpress (MiameExpress). Brazma also discussed the specia arrangements that have been made for the submission of extremely large microarray data sets.

Experimental design and normalization

Normalization is a particular type of preprocessing that is applied to correct systematic variations both in and between data sets, such as differences in labelling efficiencies. Choosing the appropriate experimental and array design facilitates data normalization and further downstream analysis. P. Kemmeren (Utrecht, The Netherlands) presented a normalization method to accurately determine differential gene-expression levels using external controls. Most normalization methods assume that the messenger RNA (mRNA) expression levels of only a few genes change in each condition, or that changes in mRNA content are balanced (that is, a similar number of genes are upregulated and downregulated in each particular condition). Changes are calculated relative to the majority of transcripts but if global shifts in mRNA occur, such as in the yeast stationary phase, these methods can be misleading (Fig 1). Kemmeren showed that global changes in mRNA levels can be monitored more accurately with the use of external RNA controls, such as *Bacillus subtilis* mRNA, which are added to the samples in known concentrations (van de Peppel et al, 2003).

High background noise in the measurements can cause further problems. For low-intensity signals, background noise can be close to the signal intensity itself, leading to increased variance that confounds the detection of gene-expression changes for weakly expressed genes. A. von Heydebreck (Berlin, Germany) presented a computational method called ‘vsn’ to stabilize the variance across the intensity range. This variance-stabilization method uses the dependence between the variance and mean intensities to derive a transformation such that the variance becomes approximately independent of the mean intensities (Huber et al, 2002; an implementation of vsn is available online as an R package at http://www.dkfz.de/abt0840/whuber). The transformed ratio provides a more reliable measure for differential gene expression that can be used in downstream analyses regardless of the intensity range.

Data analysis, clustering and gene selection

Many different data-analysis methods can be applied to a microarray data set after the normalization step, depending on the particular questions being studied. The Gene Expression Pattern Analysis Suite (GEPAS), which was developed by J. Herrero (Madrid, Spain) and co-workers (Herrero et al, 2003), contains many tools to identify functionally related genes. It allows preprocessing of the data, execution of pairwise comparisons, gene

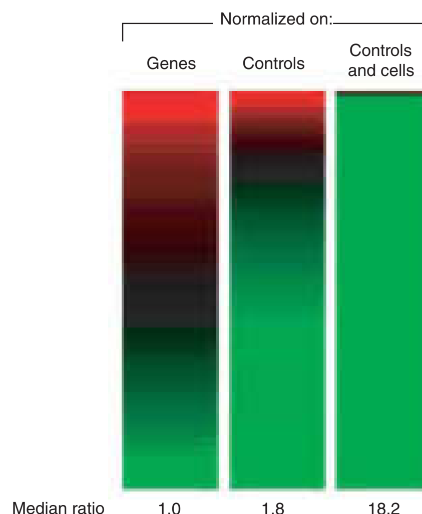


Fig 1 | The perception of changes in gene expression can depend on the normalization method used. The same data set and normalization method was used based on endogenous genes (left column), external controls (middle column), and external controls and cell count (right column). Upregulated genes are shown in red and downregulated genes in green. Numbers below the bars indicate the median change of all genes after normalization (as a ratio).

selection, unsupervised clustering and molecular classification, and is linked to the web-based tool FatiGO (http://fatigo.bioinfo.cnio.es; see later) for gene-annotation retrieval.

An important task in data analysis is the detection of differentially expressed genes. However, multiple tests for differentially expressed genes raise the probability of making false discoveries, as is the case for multiple tests in general, and this problem increases with the number of comparisons made. For example, an error rate of 5% might be acceptable in an individual test, but if 100 such tests are performed, there will probably be five false-positive results and, furthermore, their identity will be unknown. A solution could be to choose smaller significance levels for the individual tests to reduce the total number of false positives, but this is not feasible for microarray experiments given the large number of genes and limited precision of the technology. Different approaches can be taken to decide which tests provide acceptable error rates. S. Dudoit (Berkeley, USA) proposed general multiple-testing procedures for controlling false-positive rates, such as the generalized family-wise error rate (FWER). Using this procedure, the probability of finding at least one false positive is minimized. This new framework covers a broad range of testing problems that cannot be handled by traditional procedures, such as tests concerning parameters in survival models, pairwise correlations and Gene Ontology (GO) annotation. In a different approach, Y. Benjamini and A. Reiner (Tel Aviv, Israel) proposed the use of false-discovery rate (FDR) to control the number of false positives (Reiner et al, 2003). The FDR is defined as the expected proportion of false positives among the rejected hypotheses; that is, the proportion of falsely discovered differentially expressed genes. The expected proportion of false positives is estimated from running the same analysis on randomized data. The FDR is less conservative

©2004 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

than the FWER. The FWER is more appropriate for analyses in which a single false positive is unacceptable, such as comparing various drug treatments with a control. By contrast, the FDR is more applicable in screens for candidate genes, in which a certain proportion of false positives among the discovered genes is acceptable.

M. van der Laan (Berkeley, CA, USA) proposed a new method to find the optimal predictor for multivariate regression analysis, which involves predicting the outcome of a certain experiment on the basis of a number of variables, such as gene-expression levels. The deletion/substitution/addition algorithm (DSA) minimizes the residual sum of squared errors over a subset of basis functions. After training, the algorithm can be used as a black-box algorithm for multivariate regression; for example, to detect transcription-factor binding sites using yeast gene-expression data. G. Valentini (Milan, Italy) proposed a method for cancer classification using support vector machines (SVMs) as a classifier (Valentini, 2003). As not all genes are relevant for distinguishing normal from malignant tissues, feature-selection methods are used to select only those genes that are necessary for correct classification, therefore reducing the effect of background noise on the outcome of the prediction. Preliminary results indicate that the low-bias bagging (Lobag) approach used by Valentini and colleagues in association with feature selection outperforms other SVMs for the detection of normal and malignant tissues.

The work of R. Díaz-Uriarte (Madrid, Spain) addresses the identification of molecular signatures from biological data. Gene-expression signatures are sets of genes that are co-ordinately expressed and are related to the phenotypic condition. Most existing models for the identification of molecular signatures fail to address both of these requirements. An alternative approach is to identify a seed gene with good predictive abilities and then to iteratively look for groups of genes that are highly correlated both with the seed gene and among themselves, which also have good predictive abilities. Genes are eliminated from the group if they show only a small correlation with the seed gene or do not improve the prediction accuracy. According to tests with simulated and real data sets, Díaz-Uriarte showed that the performance of this algorithm is comparable with other state-of-the-art methods. The features learned (for example, the identification of a group of predictive genes) are interpretable, and the algorithm can be easily applied to other classifiers and other types of dependent variable (for example, survival analysis).

Information mining and automatic annotation methods

It is important for successful data mining to keep the information about genes that are represented on a microarray up to date, but this can be difficult for commercially produced arrays. J. De Las Rivas (Salamanca, Spain) introduced the tool Dynamic Annotation of GeneChip probe sets from Affymetrix (DAGA) for the identification and annotation of genes that are included on the oligonucleotide arrays from Affymetrix. The program generates a consensus sequence on the basis of the combination of the probes that form each set, and uses this consensus to search with BLAST for homologous sequences in mouse or human genome databases. The tool allows the validation of each probe set and their reassignment to genes, therefore keeping the annotations up to date with the information in the sequence databases.

Another problem faced by many scientists after performing cluster analysis of microarray data is how to identify biologically meaningful clusters and put them into context with the published literature to develop new hypotheses. This task is confounded by problems at several levels. Information retrieval—for example, locating all of the articles about one specific gene—can be difficult because many genes have several names. Unfortunately, in some cases, the same gene name refers to several different genes and, similarly, some acronyms are used to abbreviate several different terms. J. Tamames (Madrid, Spain) discussed the available text-mining tools, including TextDetective, which supports the retrieval of articles on particular genes, proteins, drugs or diseases, and TextMiner, which identifies the most relevant information from a set of articles. These tools are based on statistical methods, such as comparing word frequencies between articles of interest and all other articles. A. Valencia (Madrid, Spain) proposed integrating literature information into clustering analysis of microarray data to identify biologically meaningful clusters (Blaschke et al, 2001). He described one such approach, which integrates the clustering of expression profiles with the text-mining system Gene Expression Information System for Human Analysis (GEISHA). Valencia also pointed out that text-mining competitions are held to compare the performance of automated literature-mining systems (for further details, see http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04).

J. Dopazo (Madrid, Spain) described the web-based tool FatiGO, which supports the biological interpretation of clusters on the basis of the incorporation of biological knowledge derived from GO (Ashburner et al, 2000). GO is a hierarchical system of controlled vocabularies that is used by many biological databases to annotate proteins in a standardized hierarchical fashion. FatiGO finds GO terms that describe a group of genes with respect to a reference set, such as the remainder of the genome, and estimates the significance of the results (Al-Shahrour et al, 2004). J. Komorowski (Uppsala, Sweden) presented a method to use microarray data to infer the participation of genes in biological processes (Lagreid et al, 2003). Templates, such as constant expression, increasing expression or decreasing expression, are used to describe the expression patterns. Time-series expression profiles are divided into possible subintervals, each of which is assigned to an expression template using Boolean reasoning. On the basis of the expression templates, genes are assigned to a biological process using a 'guilt-by-association' approach. For a given profile, all applied rules are examined (over all possible subintervals) and a functional annotation is assigned on the basis of a majority vote.

In addition to looking for literature information on clusters, shared regulatory mechanisms can be explained by looking for *cis*-regulatory sequence motifs. This is performed under the assumption that coexpression at the transcriptional level is associated with transcriptional coregulation, which, in turn, is reflected at the sequence level by the presence of transcription-factor binding sites. Y. Moreau (Leuven, Belgium) presented a method for discovering *cis*-regulatory motifs using microarray data to identify known motifs using position-weight matrices and unknown motifs using a Gibbs motif sampling method, which searches for the statistically most probable motifs in a set of nucleotide sequences, and can find the optimal width and number of these motifs in each sequence (Lawrence et al, 1993). However, individual motifs often cannot explain the patterns in gene-expression data, and combinations of the motifs might be more appropriate; for example, in cases of synergistic effects between transcription factors. A genetic algorithm can therefore be used to search efficiently through all possible combinations of motifs. Although this approach works well for yeast, there are

©2004 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

problems with large non-compact genomes in which too many potential motifs are being found. The results for these genomes can be improved by restricting the analysis to evolutionarily conserved regions. All of these methods are part of the Java application TOUCAN, which is being developed by Y. Moreau and co-workers (Aerts *et al*, 2003).

Gene networks

The availability of high-throughput technologies makes it possible to explore large-scale regulatory networks, but also highlights the limitations of the existing modelling techniques and data sets. New approaches are necessary to analyse gene networks on a genome-wide scale. One particular problem that was addressed by several speakers is the integration of high-throughput data from heterogeneous sources, such as data on gene expression, mutant phenotypes and protein complexes. These integrated data sets form the basis for the subsequent analysis of gene networks. R. Shamir's group (Tel Aviv, Israel) is focusing on the modelling and analysis of networks that involve transcription regulation and metabolism. General steps in network inference include the definition of a class of possible networks and a scoring function, which scores how well a solution fits the data. Ideally, the search algorithm should find all possible solutions to the problem, but frequently the search space is too big. Shamir therefore proposed that a partially known network should be taken and the model should be refined by adding further levels of detail. He presented a model for lysine biosynthesis in the yeast *Saccharomyces cerevisiae* on the basis of data from the literature. Hypotheses about the regulation of lysine biosynthesis derived from this model were then compared with biological data to validate the model. He also presented a computational model in which changes in mRNA concentration are computed on the basis of transcription-factor concentrations, transcription-factor–DNA affinity and DNA signals in the promoter (Tanay & Shamir, 2003). Y. Barash (Jerusalem, Israel) described the use of probabilistic models for the identification of regulatory networks and improved modelling of DNA-binding sites within proteins. He presented several applications in which graphical models have been helpful to integrate expression data with other genomic data to extract meaningful biological hypotheses and associate statistical confidence with them. These models have been used to identify new binding sites for transcription factors, interactions among transcription factors, co-regulated gene modules and to predict expression profiles for genes under various conditions. An example is the identification of a respiration module in *S. cerevisiae*, in which transcription factor PHD1 activates its target genes, including the gene for the transcription factor HAP4, and this factor then activates secondary target genes such as *COX4*, *COX6* and *ATP17* (Segal *et al*, 2003). T. Schlitt (Cambridge, UK) compared experimental data sets for *S. cerevisiae* using a

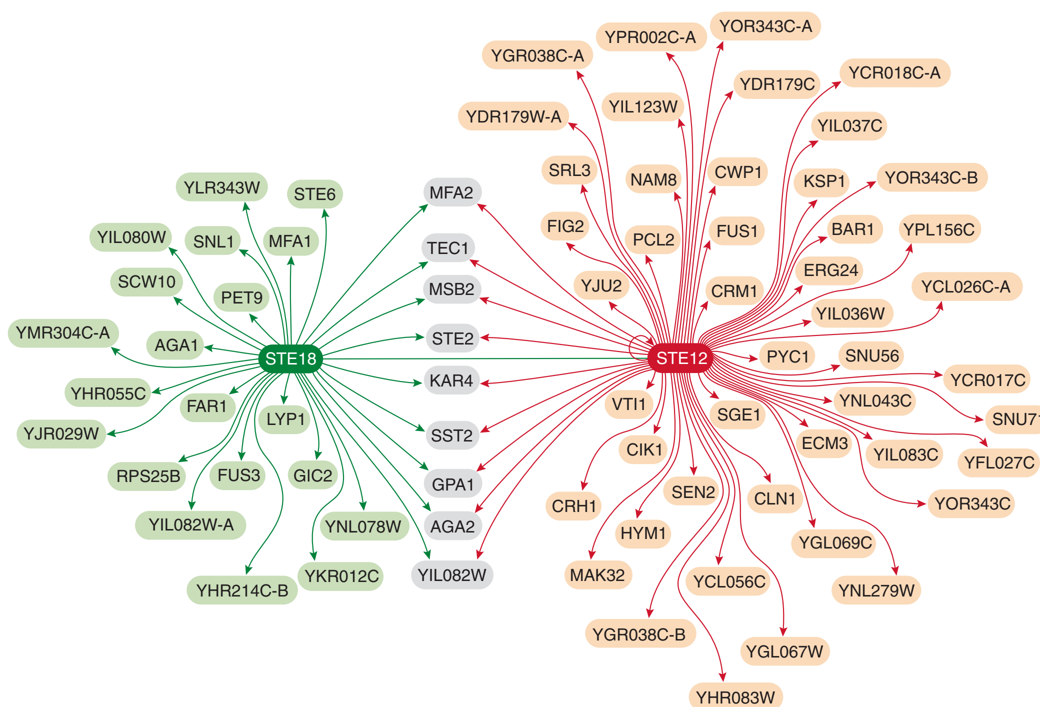


Fig 2 | Functional relationships between genes can be identified using a graph-based approach. A functional relationship between the genes *STE12* and *STE18* is indicated by the small but significant overlap (grey nodes) between the yeast genes that have a binding site for the factor STE12 in their promoter, and expression changes observed in the *ste18* deletion mutant. Here, genes are represented as nodes and are connected to *STE18* (dark green node) if their expression differs between the *ste18* mutant and the wild-type strain (green connections), or to *STE12* (red node) if STE12 binds to their putative promoters (red connections). Both *STE12* and *STE18* are involved in the pheromone response in yeast.

©2004 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

graph-based approach, in which arcs (A→B) are used to represent information such as "transcription factor A regulates gene B" (Fig 2). The comparison includes data sets from chromatin-immunoprecipitation experiments, computational analyses of transcription-factor binding sites and microarray experiments on single-gene-deletion mutants (Schlitt et al, 2003). It is possible to predict functional relationships between genes by comparing gene neighbourhoods in these graphs (Fig 2). The relationships that were identified correspond to known protein–protein interactions and/or their co-occurrence in abstracts of scientific articles. F. Falciani (Birmingham, UK) used relevance networks to model the interaction between tumour cells and normal cells in prostate cancer. Crosstalk is thought to influence many important aspects of tumour biology. Falciani has developed a strategy to identify new genes that are involved in this crosstalk on the basis of gene-expression profiling and statistical modelling. The analysis revealed several genes, such as the repulsive factor Slit-2, that have a potential role in tumour growth and metastasis formation, and these have been verified experimentally.

Conclusion

During this workshop, numerous topics were discussed, ranging from data preprocessing to machine-learning approaches. Many challenges are still being faced with regard to the proper annotation of both the material-processing steps and the genes themselves. New insights are also being gained with respect to data normalization and preprocessing, in which methods to deal with changes in global mRNA expression and ratio statistics allow gene-expression changes to be measured more accurately. The talk dealing with *cis*-regulatory module detection showed that more sophisticated methods are available for this task, which are able to address large non-compact genomes. During the network session, it became clear that many challenges still exist in this area. Important aspects of how to deal with large data sets, different data qualities and different types of data are actively being explored. Progress has already been made in many of these areas and should continue in the future, with new developments and technologies becoming available. From this workshop, it was clear that the interplay between different areas of expertise will have a crucial role in advancing our understanding of biological processes.

ACKNOWLEDGEMENTS

The authors thank A. Brazma, M. Clark, F.C.P. Holstege and the speakers for their comments on the manuscript, as well as the organizers of the workshop. P.K. is supported by a grant from the Netherlands Organisation for Scientific Research (NWO; grant no. 05050205).

REFERENCES

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B (2003) Computational detection of *cis*-regulatory modules. *Bioinformatics* 19 (Suppl. II): 5–14

Al-Shahrour F, Díaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20: 578–580

Ashburner M et al (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29

Blaschke C, Oliveros JC, Valencia A (2001) Mining functional information associated with expression arrays. *Funct Integr Genomics* 1: 256–268

Brazma A et al (2001) Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nat Genet* 29: 365–371

Herrero J, Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res* 31: 3461–3467

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

Lagreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK (2003) Predicting gene ontology biological process from temporal gene expression patterns. *Genome Res* 13: 965–979

Lawrence CE, Altschul SF, Bogouski MS, Liu JS, Neuwald AF, Wooten JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262: 208–214

Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19: 368–375

Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A (2003) From gene networks to gene function. *Genome Res* 13: 2568–2576

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34: 166–176

Tanay A, Shamir R (2003) in *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 2003)* (eds Vingron M, Istrail S, Pevzner P, Waterman M), 301–310. ACM, New York, USA. doi:10.1145/640075.640115

Valentini G (2003) An application of low bias bagged SVMs to the classification of heterogeneous malignant tissues. *Lect Notes Comput Sci* 2859: 316–321

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

Thomas Schlitt

Patrick Kemmeren

©2004 EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

Bio-IT World
TECHNOLOGY FOR THE LIFE SCIENCES
An International Data Group Publication

MARCH 2004

CNIO Symposium report, "Profiles in Carcinogenesis"

Microarrays, proteomics, and targeted cancer therapeutics reign in Spain at molecular taxonomy in Madrid

News&Analysis

Profiles in Carcinogenesis

MICROARRAYS, PROTEOMICS, AND TARGETED CANCER THERAPEUTICS REIGN IN SPAIN AT MOLECULAR TAXONOMY IN MADRID

By Kevin Davies

"Consider, by the time you reach 40 ... your 30 trillion or so cells have each divided themselves a few thousand times. How could it possibly not be that a few of those cells would not have banded together in that state of cytological anarchy which leads to cancer and death?" — JOHN DIAMOND

MADRID — "Cancer Jab Soon" was a British newspaper headline that greeted the discovery of the first human oncogene. Media hype has changed little in 25 years, nor has the annual number of casualties of cancer. But speakers at the Spanish National Cancer Centre (CNIO) 2004 Symposium* presented copious signs of a detailed new taxonomic understanding of tumor cell biology, thanks largely to advances in DNA profiling of gene expression, producing new clues about the hallmarks of cancer and new strategies in treating the disease.

Census of Cancer

Genome census has been conducted by Michael Stratton (Wellcome Trust Sanger Institute) and colleagues of all cancer-causing mutations in the two decades since the first such mutation was described in the *RAS* oncogene by Mariano Barbacid (CNIO). Mutations have been logged in 291 genes, or 1 percent of the genome, although most are restricted to sporadic, rather than inherited, cancers. Twenty-seven of these are genes for protein kinases (enzymes that attach phosphate groups to proteins), whereas only six would be expected by chance, which helps explain why the kinase family is such a popular therapeutic target. Stratton's Cancer Genome Project ultimately aims to screen the entire genome for cancer mutations.

Cancer is often associated with changes in gene expression as well as sequence, spawning interest in the field of epigenetics. Manel Esteller (CNIO) spoke of "the five bases of DNA," referring to 5-methylcytosine, which makes up 4 percent to 8 percent of the genome. The use of demethylating agents to re-activate tumor suppressor genes reduces lymphoma in mice, and has "big potential for cancer treatment," Esteller said.

*The CNIO 2004 Symposium: The Molecular Taxonomy of Cancer; Madrid, Spain: March 3-6, 2004. Co-organized by Kevin Davies, Todd Golub, Miguel-Angel Piris, and Amanda Wren.

Despite the intense focus on genetic factors, environmental triggers remain shrouded in mystery. Mary-Claire King (University of Washington), who pinpointed the breast and ovarian cancer gene *BRCA1* in 1990, reported that the risk of breast cancer in relatives of *BRCA1/2* carriers is significantly higher in women born after 1940, than in those born before 1940. The key environmental determinant turns out to be teenage physical activity. Girls who were active tend to have a later age-of-onset of cancer, presumably because of the effect of physical activity on regulating estrogen levels (a known cancer risk factor). King also described a new technology called FLASH (fosmid library allele separation and haplotyping), which she is using to identify new mutations and susceptibility genes.

Express Yourself

A cursory analysis of Medline citation trends by Olli Kallioniemi (VTT Technical Research Centre of Finland) reveals that interest in microarrays has outstripped all other postgenomic buzzwords (bioinformatics, systems biology, genomics, and so on). Kallioniemi's team is studying numerous types of biochips, including cell carpet assays — overlaying cells on slides printed with biomolecules such as Qiagen's "druggable genome" siRNA library. Tissue microarrays, with 1,000 tissue specimens per slide, and lysate microarrays, are also paving the move "from pathology to pathomics."

However, most presenters focused on the diagnostic and therapeutic applications of DNA chips. In 2002, Laura van't Veer (Netherlands Cancer Institute) described a breast cancer gene profile that correlated with poor prognosis. New profiling studies of primary breast tumors reveal close similarities to distant metastases, suggesting that metastatic potential is an inherent property of breast tumors (Weigelt, B. et al. *PNAS* 100, 15901-5; 2003).

Louis Staudt (National Cancer Institute [NCI]) has extended profiling studies in lymphoma to follicular lymphoma, seeking the "supervised discovery" of key genes linked to clinical outcome and survival. Staudt found that the synergism of three groups of gene signatures — two in immune response, one in B-cell differentiation — was the best predictor of survival outcome. Miguel-Angel Piris (CNIO) also described novel gene signatures in non-Hodgkin's lymphoma. And Chris Boshoff (University College London) presented striking findings

(CONTINUED ON PAGE 20)

Briefs

SHARPER IMAGING
The National Institute of Biomedical Imaging and Bioengineering (at the NIH) and the Center for Devices and Radiological Health (at the FDA) have formed the Laboratory for the Assessment of Medical Imaging Systems, which will assess and optimize high-resolution medical imaging systems.

WHAT'S THE FREQUENCY?
A new industry effort will focus on using radio frequency ID (RFID) technology to improve drug manufacturing and distribution operations. Participating companies include Accenture, Abbott Laboratories, Barr Pharmaceuticals, Johnson & Johnson, McKesson, Pfizer, and Procter & Gamble. This group will examine the application of RFID technology to enhance the safety and security of supply chains.

TRANSPLANT GENETICS
A new public database on blood and marrow stem cell transplants gives researchers a tool to make more informed decisions about transplants. The NIH stem cell transplant database contains patient age, gender, and ethnicity data from more than 1,300 transplant donors and recipients.

IT'S A DOG'S HEART
Gene Network Sciences (GNS), Cornell University, and the University of California at San Diego have been awarded a $2-million bioengineering grant by the National Heart, Lung, and Blood Institute to develop a computer model of the canine ventricle. GNS will use physiological data from Cornell to develop computer models of various heart cells, leading to a 3-D computer model of a canine ventricle.

News&Analysis

Profiles in Carcinogenesis

(CONTINUED FROM PAGE 18)
on the gene-expression changes associated with HIV-induced Kaposi's sarcoma.

Todd Golub (Dana-Farber Cancer Institute/Broad Institute) expressed surprise that his group's molecular taxonomy studies should yield a therapeutic target — FLT3 kinase in leukemia — and highlight the broad role of kinases in cancer. Lately, Golub's team has shown that some primary lung tumors exhibit metastatic gene signatures, using either Affymetrix or Rosetta inkjet arrays. But Golub cautioned that the current studies "are not robust enough to hallucinate about clinical significance."

THREE OF A KIND: Iya Khalil (Gene Network Sciences), Emmanuel Petricoin (FDA), and Mary-Claire King (University of Washington) show that there's more to molecular taxonomy than microarrays.

In another vignette, Golub asked: "Can gene-expression signatures be used directly as a starting point for a small-molecule screen?" The answer is yes. Taking a small-molecule library of 2,000 compounds produced by Brent Stockwell, Golub has revealed 10 compounds that reproducibly induced a neutrophil signature in leukemia cells. Further studies have used a technique called gene set enrichment analysis, which Golub's colleague David Altshuler has used successfully to highlight gene signatures in diabetes, to classify gene signatures in lung adenocarcinoma.

Not every speaker enthused about DNA microarrays, however. "DNA is an information archive," said the FDA's Emmanuel Petricoin, "but proteins do all the work!"

Petricoin stridently advocated serum proteomics for the early detection of ovarian cancer, but rather than focus on a single biomarker (such as the current CA-125 test), Petricoin and Lance Liotta (NCI) have developed a diagnostic test called OvaChek — a serum proteomic pattern revealed by mass spectrometry (MS) in conjunction with Correlogic Systems — without identifying the protein fragment peaks. Early diagnosis of ovarian cancer would save lives and prevent unnecessary prophylactic oophorectomies. The first bioinformatic analysis relied on MS patterns produced by the Ciphergen SELDI system (Petricoin, E.F. et al. *Lancet* 359, 572; 2002), but now relies on the new high-resolution Applied Biosystems QSTAR time-of-flight system. "These are real features that are being predicted," Petricoin said, noting the excellent profile reproduction after 90 days, and hopes the test could be in the clinic within five years.

Mathematical Mojo

John Quackenbush (The Institute for Genomic Research) is "building resources to do science, to ask fundamental biological questions." Some of his open-source bioinformatics tools (see www.tigr.org/software) include MADAM (microarray data manager), MIDAS (Java data analysis), and MeV (data mining). The former physicist stressed the need to link genes to proteins because of the generally poor correlation between RNA and protein levels. "We're at a point where we can develop a theoretical biology, moving from biology being deterministic to being stochastic," he said.

Supporting that assertion, Iya Khalil, co-founder of Gene Network Sciences, delivered a spellbinding overview of the company's colon cancer cell model using its diagrammatic cell language (DCL). The model incorporates more than 500 genes to model processes including apoptosis, cell cycle, and proliferation in five cellular compartments. For one unspecified pharma customer, GNS could predict the top 100 and lower 100 targets from a group of 775 targets. Khalil hopes to publish a portion of the *in silico* model later this year.

Cracking the 'Kinome'

The success of Gleevec has focused attention on the kinase family as enticing drug targets. Some 50 kinase inhibitors are currently in clinical trials, although only two have been approved so far. But building on that success will not come easy — or cheaply.

"Drug discovery takes about five years," said GlaxoSmithKline's Peter Goodfellow. One-and-a-half years are required for high-throughput screening, conducted in state-of-the-art facilities such as Tres Cantos, just outside Madrid; the remainder is devoted to "chemical tinkering to produce the optimal structure." Clinical testing could take seven years, but the prevalence of related kinases means many drugs could exhibit unexpected toxicity. Nevertheless, with the full kinase family now known, GSK can test drug candidates against the "kinome."

One drug in the clinic, GW572016, targets both EGFR and ErbB2 (the targets of Herceptin and Iressa/Tarceva). Sixty percent of tumors overexpress either of these proteins, raising the intriguing idea of targeting both simultaneously. Positive responses have been recorded in more than half the patients.

Ultimately, success will hinge on the cost-benefit ratio — if the molecule is safe, it could be taken by a large number of people. "Aspirin is the world's most dangerous drug," said Goodfellow, yet still very safe, considering the number of people who take it.

Carlos Garcia-Echeverria (Novartis Institutes for Biomedical Research) also focused on kinase drugs, while reiterating the numerous "druggability hurdles" facing his medicinal chemistry team, including potency, cell permeability, toxicity, stability, and ease of manufacture. Novartis' pipeline includes PKC412, a drug that targets FLT3 in acute myelogenous leukemia, having yielded results in a mouse leukemia model comparable to Gleevec. Another co-targeting drug, AEE-788, looks promising as an inhibitor of angiogenesis (new blood vessel growth) in solid tumors.

Nick Dracopoli (Bristol-Myers Squibb) lobbied for pharmacogenomics in designing clinical trials, particularly in population screening, to combat the low efficacy of most oncogenic drugs. Finding good responders may not reduce the time of enrollment in clinical trials, but the success of "Herceptin shows that pharmacogenomics is economically practical in the industry," he said.

Sir John Maddox closed the meeting with the observation that the "public needs a sense of cultivated stoicism ... the bad luck of cancer is something we have to live with. Eventually it will be tractable, treatable, manageable ... But not yet."

Curtailing the cytological anarchy that claims one life in the United States every 60 seconds will take decades more work. But deep within the new molecular taxonomy of cancer almost certainly lies the answer. ●