UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE CIÊNCIAS BÁSICAS DA SAÚDE
DEPARTAMENTO DE BIOQUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CB: BIOQUÍMICA

ESTUDO DA DIVERSIDADE TUMORAL

E DESENVOLVIMENTO DE FERRAMENTAS DE BIOINFORMÁTICA PARA ANÁLISE CITOGENÉTICA E MOLECULAR DE NEOPLASIAS SÓLIDAS

Mauro Antônio Alves Castro

Porto Alegre
2009

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE CIÊNCIAS BÁSICAS DA SAÚDE
DEPARTAMENTO DE BIOQUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CB: BIOQUÍMICA

ESTUDO DA DIVERSIDADE TUMORAL E DESENVOLVIMENTO DE FERRAMENTAS DE BIOINFORMÁTICA PARA ANÁLISE CITOGENÉTICA E MOLECULAR DE NEOPLASIAS SÓLIDAS

Doutorando: Mauro Antônio Alves Castro

Orientadores: Dr. José Cláudio Fonseca Moreira (orientador)
Dra. Rita Maria Cunha de Almeida (co-orientadora)

Tese submetida ao Programa de Pós-graduação em Ciências Biológicas: Bioquímica, como requisito para obtenção do grau de Doutor em Bioquímica.

Porto Alegre, 2009

DEDICO

Aos meus pais
À minha esposa

"No great discovery was ever made without a bold guess"

(Isaac Newton)

AGRADECIMENTOS

Ao professor **José Cláudio F. Moreira**, pela orientação, apoio incansável, incentivo, confiança e amizade.

À Professora **Rita M. C. de Almeida**, pela co-orientação, apoio, generosa dedicação, valiosa contribuição intelectual e amizade.

Ao Professor **José C. M. Mombach**, pelo incentivo, auxílio constante, apoio na condução deste trabalho e amizade.

Ao Professor **Fábio Klamt**, pela colaboração na condução dos estudos experimentais, discussão dos mais diversos assuntos e amizade.

Ao Professor **Tor G. H. Onsten**, pela colaboração na condução dos estudos em citogenética.

Aos colegas **Rodrigo J. S. Dalmolin**, **José L. Rybarczyk Filho** e **Marialva Sinigaglia**, pelo companheirismo, aberto envolvimento e colaboração na condução dos estudos de bioinformática.

Aos colaboradores que participaram na caracterização dos biomarcadores tumorais.

A todos os amigos e colegas do CEEO, pela amizade e momentos de descontração.

A todos os colegas do Instituto de Física, pela oportunidade de cooperação interdisciplinar.

Ao PPG Bioquímica (UFRGS), professores e funcionários, e às agências de fomento CNPq e CAPES.

A todos que de alguma forma contribuíram para o desenvolvimento deste trabalho.

Muito obrigado!

RESUMO

A progressão das neoplasias sólidas é um processo complexo, não segue uma seqüência universal e é caracterizada pela marcada heterogeneidade tumoral na ocasião do diagnóstico. Anormalidades cromossômicas, mutações somáticas e alterações epigenéticas estão entre as principais alterações celulares decorrentes da instabilidade do genoma destas neoplasias. Essa instabilidade tem importantes implicações clínicas, pois impõe dificuldades ao desenvolvimento de novos biomarcadores tumorais, tais como a baixa recorrência de mutações somáticas causalmente relacionadas ao desenvolvimento do câncer e a grande diversidade amostral. Neste trabalho estudamos alterações citogenéticos e moleculares com objetivo de caracterizar padrões de diversidade tumoral. Os resultados demonstraram que a diversidade cariotípica é específica para cada tipo de tumor e que os tumores de menor diversidade apresentam estatísticas populacionais com melhor sobrevida (79 tipos de tumores; n=12787). Além disso, desenvolvemos novos métodos computacionais baseados na teoria da informação com objetivo de mapear a diversidade de expressão gênica dos mecanismos de estabilização do genoma. Os resultados obtidos sugerem que redes de genes de apoptose e do sistema de reparo por excisão de nucleotídeos estão funcionalmente alteradas em neoplasias sólidas, com aumento da diversidade de expressão gênica e diminuição da abundância de transcritos (14 tipos de tumores; n=492). A implicação deste achado é que ele fornece evidências em favor da instabilidade genômica ao nível dos nucleotídeos, um tipo de disfunção que causa aumento da taxa de mutação somática e que está caracterizada apenas em modelos teóricos e experimentais ou em raros casos de desordens hereditárias. Por fim, a padronização dos métodos computacionais desenvolvidos resultou em duas patentes de invenção (aplicadas ao diagnóstico/prognóstico de neoplasias sólidas) e em dois produtos tecnológicos na forma de programas de computador distribuídos com licença de código aberto (aplicados ao estudo da diversidade citogenética e molecular).

Endereço de acesso: http://lief.if.ufrgs.br/pub/biosoftwares/

ABSTRACT

The progression of solid tumors does not follow a universal sequence and it is characterized by marked tumor heterogeneity. Chromosomal abnormalities, somatic mutations and epigenetic alterations are among the major cellular changes associated to genome stability impairments, which contribute substantially to this heterogeneity. The cancer genome instability has important clinical implications, since it imposes technical problems for tumor biomarker development, such as the low recurrence of somatic mutations causally related to cancer and the great sample diversity. Here, we considered different types of cytogenetic and molecular changes in order to characterize different patterns of tumor diversity. The results showed that the karyotypic diversity is specific to each tumor type and that tumors with lower diversity present better population statistics – five-year survival rates (79 types of tumors, n = 12787). Furthermore, we developed new computational methods based on information theory concepts in order to map the diversity of gene expression networks of genome maintenance mechanisms. The results suggest that two gene expression networks - apoptosis and nucleotide excision repair (NER) - are functionally altered in solid tumors, with increased gene expression diversity and decreased transcript abundance (14 types of tumors; n=492). The implication of this finding is that it provides evidence in favor of genomic instability at nucleotide level, a type of dysfunction that can increase the somatic mutation rate, which is soundly characterized only in theoretical and experimental models, or in rare inherited disorders. Finally, the standardization of the computational methods developed here gave rise to two patents (applied to the diagnosis / prognosis of solid tumors) and two bioinformatics applications released under open source license (designed for cytogenetic and molecular diversity studies).

Availability: http://lief.if.ufrgs.br/pub/biosoftwares/

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AFP	alfa-feto-proteína
APC	gene da polipose cólica adenomatosa
BER	*base-excision repair* (reparo por excisão de bases)
BRAF	gene v-raf
CFL1	gene cofilina-1, isoforma não-muscular
DHL	desidrogenase láctica
ADN	ácido desoxirribonucléico
FDR	*false discovery rate* (taxa de falsa descoberta)
H'	diversidade ou entropia de Shannon
HCG	gonadotrofina coriônica humana
INCA	Instituto Nacional do Câncer

KRAS	gene v-Ki-ras2
MMR	*mismatch repair* (reparo de erros de pareamento do ADN)
NER	*nucleotide-excision repair* (reparo por excisão de nucleotídeos)
NSCLC	*non-small cell lung cancer* (câncer de pulmão de células não pequenas)
OMS	Organização Mundial de Saúde
RER	reparo por recombinação
RNT	região não traduzida
SUS	Sistema Único de Saúde
TNM	T: tamanho do tumor; N: presença de metástase em linfonodos regionais; M: presença metástases à distância
TNM+S	TNM + marcadores tumorais séricos
TP53	gene supressor tumoral p53
XP	xeroderma pigmentosum
CpG	citosina-guanina dinucleotídeo
5-aza	5-aza-2'-desoxicitidina

Obs.: siglas em inglês foram mantidas quando de uso corrente ou quando associadas a nomenclaturas normatizadas (*e.g.* símbolo de genes).

SUMÁRIO

APRESENTAÇÃO

Esta tese está estruturada em três partes. A ***Parte I*** apresenta uma introdução geral com ênfase nas múltiplas fases da progressão tumoral e na caracterização dos diversos tipos de alterações citogenéticas e moleculares responsáveis pela grande heterogeneidade das neoplasias sólidas, seguido dos objetivos gerais e específicos.

A ***Parte II*** apresenta oito capítulos na forma de artigos científicos, patentes de invenção e programas de computador. O primeiro capítulo apresenta uma nova abordagem para a análise da instabilidade genômica de neoplasias sólidas com base no estudo da diversidade citogenética. O segundo capítulo propõe generalizações desta nova abordagem, fornece evidências em favor de algumas teorias que concorrem para explicar a progressão tumoral e descreve formalmente um novo método para interpretação de cariótipos complexos. O terceiro capítulo amplia o estudo da heterogeneidade das neoplasias sólidas com base na caracterização de redes de expressão gênica. O quarto capítulo trata algumas proposições levantadas a partir dos trabalhos anteriores. O quinto capítulo sintetiza os métodos previamente desenvolvidos na forma de uma nova ferramenta de bioinformática capaz de analisar a diversidade de expressão gênica em escala genômica. Aqui exemplificamos o seu uso e discutimos as possíveis aplicações. O sexto capítulo propõe um novo biomarcador prognóstico para o câncer de pulmão. O capítulo sete e os anexos apresentam os produtos tecnológicos resultantes desta tese.

Por fim, a ***Parte III*** apresenta a discussão geral sobre os resultados de cada artigo científico, seguido de conclusões e perspectivas.

PARTE I

INTRODUÇÃO

Epidemiologia e caracterização de neoplasias sólidas

A progressão do câncer (ou neoplasia maligna) é um processo complexo envolvendo varias escalas temporais e espaciais, um grande número de rotas metabólicas e diferentes tipos celulares (Anderson & Quaranta, 2008). Em função disso, o termo "câncer" designa um conjunto altamente heterogêneo de doenças que podem afetar qualquer parte do corpo e que têm em comum o crescimento celular desordenado, a invasão de tecidos adjacentes e a formação de metástases (Kufe et al., 2005). Entretanto, poucos tipos de neoplasias respondem pela maior parte dos casos de cânceres fatais, sendo a grande maioria formadora de tumores sólidos. No Brasil cerca de 90,0% do total de mortes por câncer em ambos os sexos está associada ao desenvolvimento de neoplasias sólidas (**Figuras 1-2**), enquanto que neoplasias de origem hematológica (*e.g.* leucemias e linfomas) respondem por menos de 7,0% dos casos.

As neoplasias de origem hematológica têm sido historicamente tratadas e classificadas separadamente das demais neoplasias, as quais são coletivamente referidas como neoplasias sólidas (ou tumores sólidos) (Bennett et al., 1976; Harris et al., 1994; Harris et al., 1999; ICD-O-1, 1976; Rappaport et al., 1971). Diferentemente das neoplasias hematológicas, as neoplasias sólidas formam massas tumorais que em geral ficam restritas às barreiras teciduais por um longo período durante os estágios iniciais da progressão tumoral até adquirirem características invasivas quando então conseguem irromper o compartimento de origem e colonizar novos sítios (Jones et al., 2008). Além disso, no decorrer da progressão tumoral, as neoplasias sólidas adquirem um número muito maior de alterações citogenéticas e moleculares (Hoglund et al., 2002b). Como resultado, nos estágios mais avançados do desenvolvimento elas exibem acentuada heterogeneidade intratumoral – com formação de múltiplos clones aberrantes (Albertson et al., 2003; Backvall et al., 2005; Glockner et al.,

2002; Harada et al., 1998; Heng et al., 2006; Losi et al., 2005; Mitelman, 2000; Mitelman et al., 1997; Orndal et al., 1994).

Ao longo desta introdução será apresentada uma linha de argüição a fim de caracterizar a heterogeneidade das neoplasias sólidas. Além disso, serão apresentadas algumas implicações clínicas desta heterogeneidade em relação ao prognóstico dos tumores sólidos, bem como padrões citogenéticos e moleculares descritos pelas principais teorias que concorrem para explicar a progressão tumoral. Estes conceitos serão posteriormente utilizados para identificar os pontos das teorias de progressão tumoral que serão objetivos de estudo desta tese.

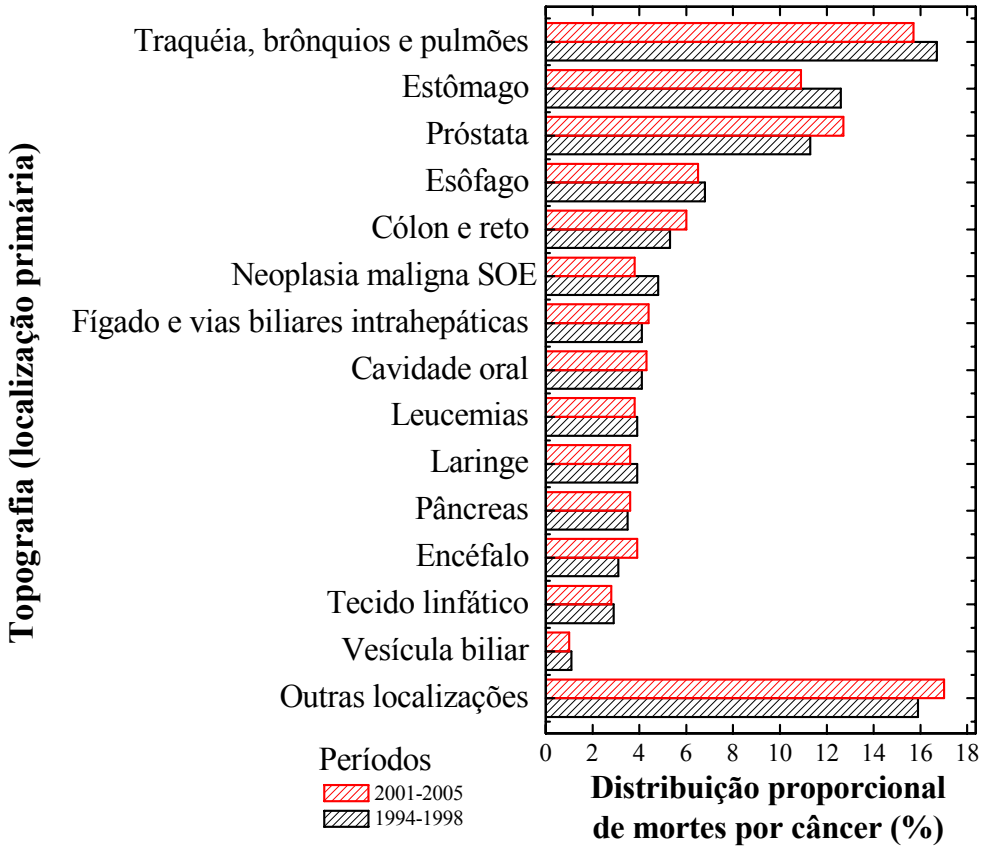


Figura 1. Distribuição proporcional do total de mortes por câncer no Brasil, segundo a localização primária do câncer, em Homens, para os períodos entre 1994-1998 e 2001-2005. Fonte: INCA/BR (Atlas de Mortalidade por Câncer, 2009). SOE: Sem Outras Especificações.

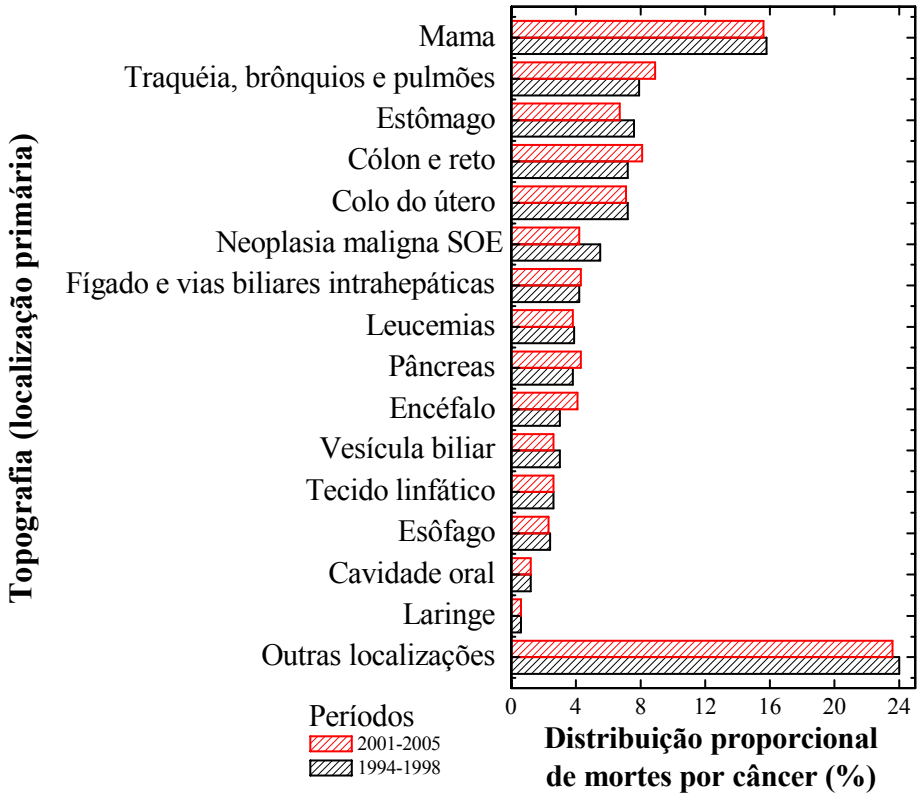


Figura 2. Distribuição proporcional do total de mortes por câncer no Brasil, segundo a localização primária do câncer, em Mulheres, para os períodos entre 1994-1998 e 2001-2005. Fonte: INCA (Atlas de Mortalidade por Câncer, 2009).

Um aspecto que pode ilustrar o grau de heterogeneidade das neoplasias sólidas é a evolução dos sistemas de normatização empregados para a classificação e o registro epidemiológico. Segundo a Organização Mundial de Saúde (OMS), as primeiras tentativas de classificar as neoplasias sólidas datam de 1950 a 1960 e foram baseadas em dois aspectos fundamentais: no sítio topográfico (local de incidência) e no comportamento celular (se a neoplasia é maligna, benigna, ou não especificada) (ICD-6, 1948; ICD-7, 1955). Contudo, estas primeiras tentativas não consideravam as características histológicas dos tumores, uma deficiência grave uma vez que as taxas de incidência e sobrevida variam de acordo com o tipo de tecido. Tal deficiência foi corrigida nas edições subseqüentes do sistema de Classificação Internacional das Doenças.

Segundo o atual sistema de classificação editado pela OMS (ICD-O-3, 2000) um espectro de mais de 600 tipos histológicos é utilizado para classificar as neoplasias sólidas. Isso reflete o alto grau de heterogeneidade intratumoral do ponto de vista citogenético (Duesberg et al., 2005; Shiomi et al., 2003), metabólico (Fidler, 1978), de receptores hormonais (Leers & Nap, 2001; Roudier et al., 2003; Sklarew et al., 1991), de susceptibilidade a radio e quimioterapia (Britten et al., 1996; Yung et al., 1982), de propriedades angiogênicas (Ahlgren et al., 2002), e de potencial invasivo e metastático (Glockner et al., 2002; Harada et al., 1998; Losi et al., 2005; Roudier et al., 2003; Spremulli & Dexter, 1983).

Apesar desse alto grau de heterogeneidade, o surgimento das neoplasias sólidas segue um padrão bastante simples do ponto de vista epidemiológico: a maior parte incide em tecidos epiteliais tardiamente na vida. O padrão de incidência do câncer tem sido descrito com base na teoria dos múltiplos estágios postulada há mais de 50 anos por Armitage & Doll (Armitage & Doll, 1954) segundo a qual a taxa de incidência Q em populações humanas aumenta na potência k, tal que $Q_t = c \times t^k$, onde t representa a faixa etária e c uma constante que especifica o tipo de câncer. A interpretação é que k +1 estágios (ou mutações) são necessários para a carcinogênese, sendo que a natureza dos estágios não afeta o modelo de incidência, o qual assume que (i) a natureza das mudanças é discreta e randômica, (ii) cada estágio é estável e (iii) a doença progride em uma única direção. Assim, os indivíduos avançariam estocasticamente através dos estágios à medida que envelhecem, gerando uma distribuição de probabilidades para o surgimento do câncer (**Figura 3**).

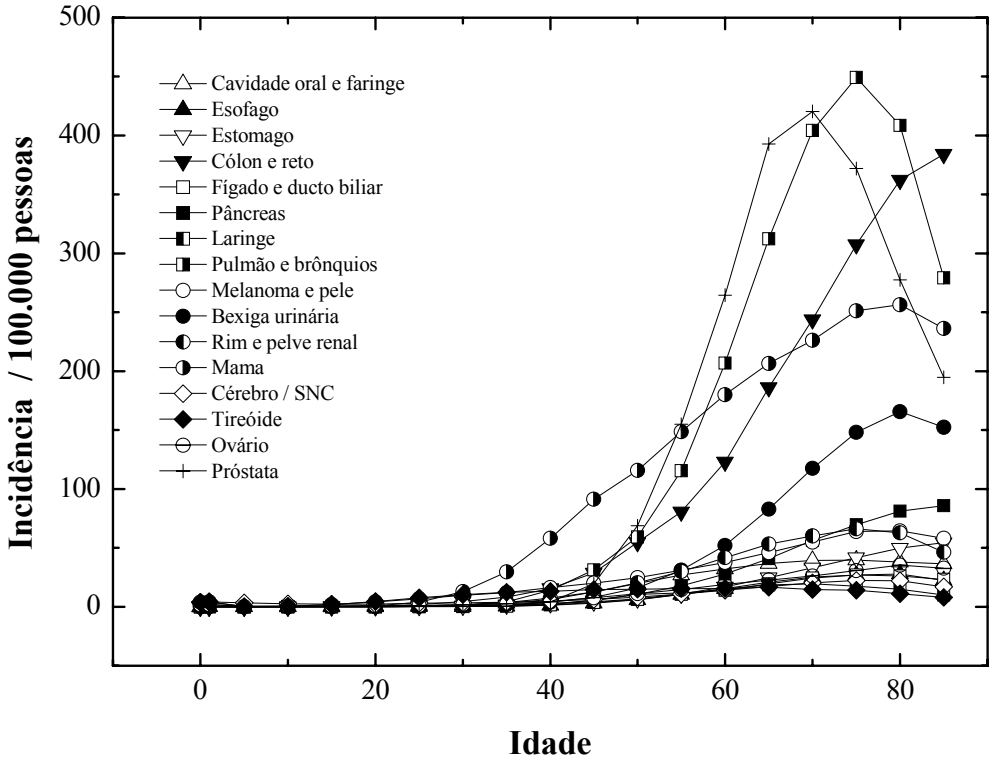


Figura 3. Curvas de incidência de câncer invasivo nos Estados Unidos, ajustada por idade e por 100 mil pessoas, entre 2002 e 2004. Contagem cobre aproximadamente 93% da população norte-americana. Fonte: NCI/SEER/EUA (Ries et al., 2009).

Embora o modelo de Armitage & Doll assuma um caráter randômico, emerge ao nível populacional um padrão previsível, o qual tem sido amplamente utilizado para tentar validar uma possível seqüência de eventos que explique a progressão tumoral. Para muitos cânceres o número de estágios foi inicialmente estimado entre quatro e sete. Posteriormente o modelo foi modificado com incorporação de diversos parâmetros, tais como expansão clonal (Luebeck & Moolgavkar, 2002), cinética de crescimento celular (Gatenby & Vincent, 2003), alterações citogenéticas (Hoglund et al., 2001), mutações em oncogenes e genes supressores tumorais (Halazonetis et al., 2008; Nowak et al., 2004), instabilidade cromossômica (Wang et al., 2004) e modificações no microambiente celular (Breivik, 2001; Breivik & Gaudernack, 2004).

Contudo, a progressão do câncer não segue uma seqüência universal e a distinção entre os diferentes estágios é um assunto controverso (Quirke et al., 2007). Por exemplo, mesmo sendo a progressão do câncer colorretal uma das mais bem descritas na literatura – e sob vários aspectos (*e.g.*, moleculares, epigenéticos e citogenéticos) (Lengauer et al., 1997; Rajagopalan et al., 2003; Vogelstein & Kinzler, 2004; Wang et al., 2004), na prática clínica a descrição do avanço desta neoplasia ainda está baseada unicamente na extensão anatômica da doença, sem considerar possíveis marcadores moleculares para complementar o estadiamento tumoral (Greene et al., 2002).

Ainda assim, do ponto de vista clínico, a busca por um modelo de progressão tumoral se justifica da seguinte forma: uma vez conhecida a seqüência de eventos que leva à transformação maligna, então indivíduos poderiam ser estratificados com base nos estágios da progressão do câncer e medidas poderiam ser tomadas para evitar o avanço da doença. Esse raciocínio está por trás dos sistemas de estadiamento clínico-patológico dos tumores sólidos, largamente utilizados para fins prognósticos e para auxílio na decisão terapêutica.

Modelos de progressão e estadiamento tumoral

Como a carcinogênese está mais bem descrita para o câncer de origem colorretal, ilustraremos nossos objetivos seguindo os autores que estudam este modelo de progressão tumoral, mas sem nos restringirmos a eles, visto que a história natural das neoplasias sólidas segue alguns padrões recorrentes (Hahn & Weinberg, 2002; Kufe et al., 2005).

A evolução molecular do câncer colorretal está fundamentada no modelo adenoma-carcinoma, conforme postulado por Fearon & Vogelstein (Fearon & Vogelstein, 1990). O modelo adenoma-carcinoma descreve a progressão molecular do câncer com base numa seqüência provável de eventos em que os autores mapearam (para cada estágio) um conjunto de genes com maior probabilidade de mutação. Este mapeamento foi feito considerando a

morfologia do tecido normal – representado pelas criptas colônicas (Cotran et al., 2000) – até a formação no câncer. A morfologia normal de uma cripta colônica está apresentada na **Figura 4**. Em condições normais, as células epiteliais da base migram para a superfície da cripta mantendo um fluxo contínuo de divisão celular, a partir da base, e substituição das células do ápice, as quais vão sendo esfoliadas pela a dinâmica de funcionamento do intestino (Humphries & Wright, 2008). Este processo de renovação leva de três a seis dias e quando as taxas de mitose superam as taxas de perda celular então ocorre o desenvolvimento de uma massa tumoral (Fodde et al., 2001).

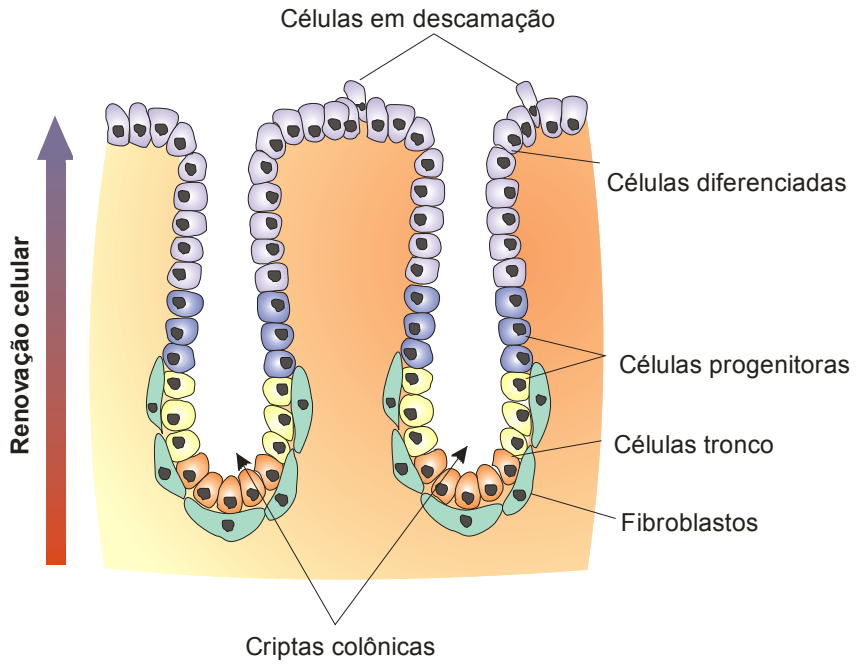


Figura 4. Organização das criptas colônicas. Adaptado de Humphries *et al.* (Humphries et al., 2008).

Em relação ao aspecto histológico, o primeiro estágio da progressão tumoral é caracterizado pelo acúmulo de células na superfície da cripta (hiperplasia), seguido por modificações na organização tecidual (displasia). Mantido esse processo, focos de tecido aberrante projetam-se para fora da superfície da mucosa normal, em direção à luz do tubo do tubo digestivo, originando pólipos. Pólipos displásicos já são considerados lesões pré-malignas (adenomas) e podem evoluir para neoplasias malignas (Cotran et al., 2000). Tal

arranjo topográfico da proliferação celular da mucosa colônica facilita a obtenção de amostras em diferentes estágios, o que explica em parte o fato do câncer colorretal ter se tornado um modelo de estudo de progressão tumoral.

Em relação ao aspecto molecular, a mutação do gene *APC* parece ser um dos primeiros eventos da progressão tumoral, o qual está alterado em cerca de 40% dos adenomas e carcinomas de cólon e reto (**Figura 5**).

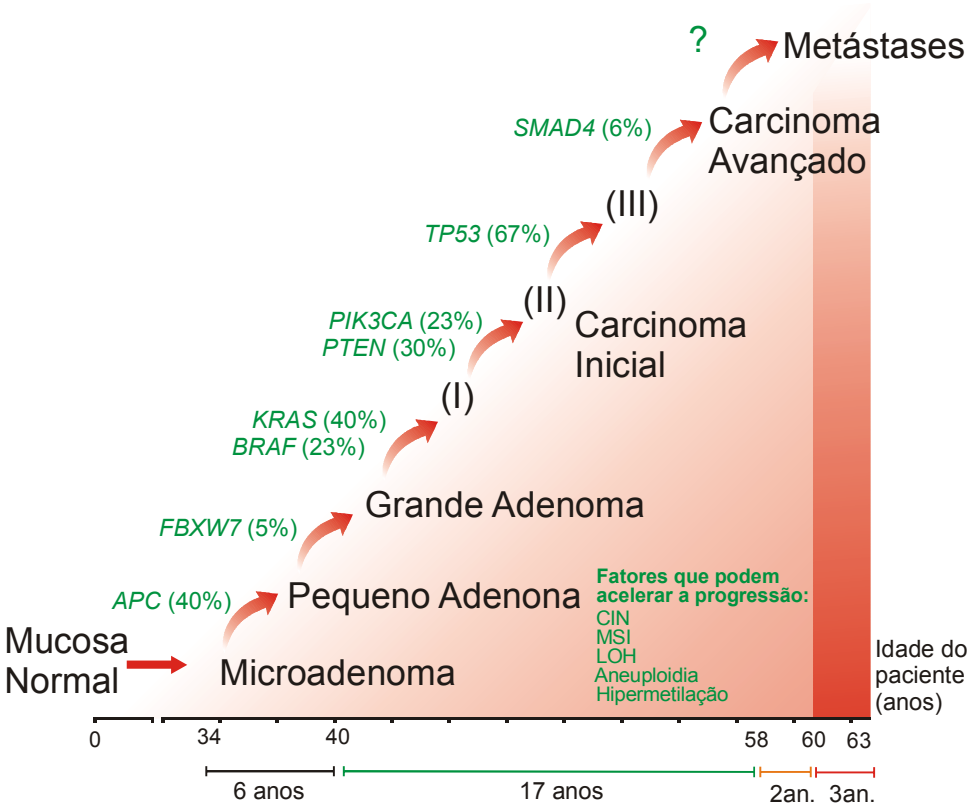


Figura 5. Estágios da progressão do câncer colorretal. Adaptado de Jones *et al.*, (Jones et al., 2008). CIN: instabilidade ao nível dos cromossomos; MSI: instabilidade microssatélite; LOH: perda de heterozigosidade. Percentuais indicam fração de tumores com genes mutados ao final da progressão, segundo o projeto COSMIC (Forbes et al., 2008). Obs.: no estágio de adenoma o gene *APC* está mutado em 42% dos 117 tumores avaliados naquele estudo.

Nesse modelo de progressão, o gene *APC* atua como inibidor da atividade da proteína β-catenina, a qual é capaz de induzir a expressão de genes envolvidos em processos de adesão e divisão celular, entre outros. A inativação da via APC/β-catenina estaria associada à progressão tumoral até o estágio de microadenoma (Ceol et al., 2007; Fodde et al., 2001). Mutações na família RAS de oncogenes (*i.e.* genes *KRAS*, *HRAS* e *NRAS*) e ativação da via

KRAS/BRAF seriam os próximos eventos genéticos (Kinzler & Vogelstein, 1996), sendo que aquisição de um fenótipo instável aceleraria a progressão (*e.g.* instabilidade ao nível dos cromossomos) (Jefford & Irminger-Finger, 2006). Os passos seguintes envolveriam mutações em genes que controlam a expressão do gene *TP53*, entre outros, bem como alterações em vias responsáveis pela transição de tumores benignos (adenomas) para tumores malignos (carcinomas) (Jones et al., 2008).

Para acompanhar a progressão da neoplasia sólida do ponto de vista clínico é necessário avaliar o estágio de evolução da doença – referido como estadiamento clínico-patológico (Greene et al., 2002). Para isso três aspectos fundamentais são levados em consideração. O primeiro deles diz respeito ao diagnóstico do tumor, conforme discutido anteriormente, envolvendo a identificação e classificação da origem topográfica e histológica do tumor primário (ICD-O-3, 2000), e se a neoplasia é benigna ou maligna.

O segundo aspecto se refere à graduação histológica do tumor (ou grau de diferenciação), podem ser classificado em grau I (bem diferenciado), grau II (moderadamente diferenciado), grau III (pouco diferenciado), ou grau IV (indiferenciado ou anaplásico) (ICD-O-3, 2000). Essa classificação fornece uma estimativa de agressividade. Por exemplo, neoplasias pouco diferenciadas são usualmente mais agressivas e tendem a ter um crescimento mais rápido que tumores bem diferenciados (de prognóstico mais favorável).

O terceiro aspecto (e seguramente o mais relevante do ponto de vista clínico) se refere à extensão anatômica da doença descrita pelo sistema TNM (T: tamanho do tumor; N: presença de metástase em linfonodos regionais; M: presença metástases à distância) (Greene et al., 2002). A **Tabela 1** apresenta um resumo do sistema TNM para o câncer colorretal. Além disso, outros parâmetros podem ser adicionados para melhorar a precisão do estadiamento, tais como marcadores histológicos e marcadores moleculares (*i.e.* marcar o estágio de evolução da doença).

Todas essas categorias são formalmente condensadas em um número conveniente de grupos para gerar o estadiamento da neoplasia (**Tabela 2**). No estágio inicial (*Estádio I*) a doença se encontra localizada, geralmente restrita ao órgão de origem, sem metástases. No estágio regional (*Estádio II*) o câncer estende-se para fora do órgão de origem, mas sem comprometer a cadeia de linfonodos regionais. No estágio regional extenso (*Estádio III*) o câncer estende-se para fora do órgão de origem através de metástase em linfonodos regionais e no estágio avançado (*Estádio IV*) o câncer está disseminado pelo corpo através de metástases à distância. Esta opção de agrupamento, tanto quanto possível, foi proposta para assegurar que cada grupo seja mais ou menos homogêneo em termos de sobrevida, o que permite estratificar os pacientes pela gravidade do caso, predizer sobrevida, escolher o tratamento inicial, uniformizar o registro do desfecho das doenças e evitar ambigüidades na troca de informações entre os sistemas de saúde (Greene et al., 2002).

Embora o modelo de progressão do câncer colorretal seja um dos mais bem descritos na literatura, marcadores de progressão molecular não são utilizados para o estadiamento deste tipo de neoplasia (Greene et al., 2002; Ludwig & Weinstein, 2005). Segundo o atual sistema de estadiamento para neoplasias sólidas (Greene et al., 2002), apenas tumores de testículo são estadiados com auxílio de marcadores moleculares (*i.e.* sistema TNM+S) (**Tabela 3**). Nas demais neoplasias sólidas o estadiamento é feito exclusivamente por critérios anatômicos, apesar de todos os avanços na caracterização molecular do câncer (Ogino & Goel, 2008).

Uma explicação plausível para essa aparente discrepância está na dificuldade em se cumprir os requisitos básicos para agregar potenciais biomarcadores aos sistemas formais de estadiamento. Um biomarcador ideal deve ser sensível e específico (*e.g.* discriminar facilmente cânceres com bom e mau prognóstico), ter um bom custo efetivo e ser robusto contra variabilidade operacional (Ludwig et al., 2005).

Tabela 1. Classificação TNM do câncer colón retal[1].

Resumo Esquemático	
T - Tumor Primário	
T0	Não há evidência de tumor primário
Tin	Carcinoma *in situ*: intra-epitelial ou invasão da lâmina própria
T1	Tumor que invade a submucosa
T2	Tumor que invade a muscular própria
T3	Tumor que invade além da muscular própria, alcançando a subserosa ou os tecidos peri-cólicos ou peri-retais, não peritonizados
T4	Tumor que invade diretamente outros órgãos ou estruturas e/ou que perfura o peritônio visceral
N – Linfonodos regionais	
NX	Os linfonodos regionais não podem ser avaliados
N0	Ausência de metástase em linfonodos regionais
N1	Metástase em 1 a 3 linfonodos regionais
N2	Metástase em 4 ou mais linfonodos regionais
M – Metástase à distância	
M0	Ausência de metástase à distância
M1	Metástase à distância

[1]Segundo UICC, 6ª. ed. TNM (Greene et al., 2002).

Tabela 2. Grupamento por estádios do câncer de colón e reto[1].

Estádios	**T**	**N**	**M**
Estádio 0	Tis	N0	M0
Estádio I	T1,T2	N0	M0
Estádio IIA	T3	N0	M0
IIB	T4	N0	M0
Estádio IIIA	T1, T2	N1	M0
IIIB	T3,T4	N1	M0
IIIC	Qualquer T	N2	M0
Estádio IV	Qualquer T	Qualquer N	M1

[1]Segundo UICC, 6ª. ed. TNM (Greene et al., 2002).

Tabela 3. Marcadores séricos de tumores de células germinativas do testículo[1].

S - Marcadores Tumorais Séricos			
SX	Os marcadores tumorais séricos não estão disponíveis ou não foram realizados		
S0	Marcadores tumorais séricos dentro dos limites normais		
	DHL[2]	**HCG (mUI/ml)**	**AFP (ng/ml)**
S1	< 1,5 x *R*	< 5.000	< 1.000
S2	1,5 - 10 x *R*	5.000 - 50.000	1.000 - 10.000
S3	> 10 x *R*	> 50.000	> 10.000

[1] Segundo UICC, 6ª. ed. TNM (Greene et al., 2002).
[2] *R* indica o limite superior do valor normal para a dosagem da DHL.

A grande heterogeneidade intra e intertumoral é uma reconhecida barreira para atingir estes requisitos (Kulasingam & Diamandis, 2008; Ransohoff, 2005). Por exemplo, mesmo mutações no gene *APC* explicando o início de uma seqüência provável de eventos (Figura 5), ainda assim – na ocasião do diagnóstico – não são encontradas mutações *APC* na maior parte dos casos de neoplasias de cólon e reto [dados do projeto COSMIC (Forbes et al., 2008)], o que inviabiliza o uso desse gene como biomarcador. Além disso, somente 7% desses tumores apresentavam mutações nos genes *APC*, *KRAS* e *TP53* combinados (Smith et al., 2002). Estes dados estão em acordo com inúmeros trabalhos mostrando que a progressão das neoplasias sólidas pode seguir diferentes caminhos, resultando em diferentes desfechos (Albertson et al., 2003; Berczi et al., 2002; Flyger et al., 1999; Heng et al., 2006; Hoglund et al., 2002b; Hoglund et al., 2005; Jass et al., 2004; Jin et al., 2002; Lengauer et al., 1998; Osterheld et al., 1999; Shen et al., 2007; Shibata, 2006).

Heterogeneidade das neoplasias sólidas

Tipos de alterações citogenéticas e moleculares

Anormalidades cromossômicas, mutações somáticas e alterações epigenéticas estão entre as principais alterações celulares capazes de refletir o aumento da heterogeneidade das neoplasias sólidas.

As anormalidades cromossômicas nos tumores sólidos são caracterizadas por um marcado grau de aberrações numéricas e estruturais (Duesberg et al., 2005). As aberrações numéricas correspondem a modificações no número padrão de cromossomos da espécie enquanto que as aberrações estruturais correspondem a anomalias na estrutura dos cromossomos. As aberrações numéricas podem derivar de alterações tanto no conjunto haplóide de cromossomos (euploidias) como em partes deste conjunto (aneuploidias) gerando

desbalanços cromossômicos. Por outro lado, as aberrações estruturais podem envolver, por exemplo, trocas balanceadas de material genético entre cromossomos ou trocas não recíprocas, o que resulta em perda ou ganho de material genético (Griffiths et al., 2002).

Segundo o sistema internacional de nomenclatura citogenética para a espécie humana (ISCN, 1995), mais de 30 termos são utilizados para o registro de aberrações citogenéticas. Translocações, deleções e adições de material de origem desconhecida estão entre os tipos mais freqüentes de aberrações estruturais encontradas em cariótipos de neoplasias sólidas (**Tabela 4**). Já o espectro de aberrações numéricas varia de 20 a 300 cromossomos, com grande ocorrência de cariótipos aneuplóides ao redor da diploidia (**Figura 6**). Em função deste marcado grau de anormalidade, aberrações recorrentes, com algum significado clínico, não são encontradas com facilidade nas neoplasias sólidas (Albertson et al., 2003), mas trabalhos recentes mostram que medidas de *diversidade* (ou *entropia*) podem permitir o uso de cariótipos para fins prognósticos (Hoglund et al., 2005; Maley et al., 2006; Merlo et al., 2006).

Assim como as aberrações cromossômicas, as mutações somáticas estão presentes em grande quantidade nas neoplasias sólidas (Venkatesan et al., 2006). Uma estiva recente apontou que, de um total de 18.191 genes investigados em câncer de cólon e mama, 9.4% apresentavam pelo menos uma mutação não silenciosa, sendo que a maioria destas alterações (92.7%) corresponde a mutações pontuais (Wood et al., 2007). Um sumário das mutações somáticas em câncer de cólon está apresentado na **Figura 7**.

As mutações somáticas em câncer podem ser passageiras ou causais. Entende-se por *causais* aquelas mutações cujo papel está diretamente relacionado à progressão tumoral, conferindo alguma vantagem adaptativa à população de células neoplásicas (Futreal et al., 2004).

Tabela 4. Distribuição dos rearranjos cromossômicos estruturais em 14467 cariótipos de biopsias de neoplasias sólidas.[1]

Tipos de rearranjos	***N***	***N*/*C***
Translocações (*t*)	14107	0.9751
Material adicional de origem desconhecida (*add*)	11279	0.7796
Cromossomos derivados (*der*)	10815	0.7476
Deleções (*del*)	10252	0.7086
Cromossomos marcadores (*mar*)	3060	0.2115
Isocromossomos (*i*)	2713	0.1875
Inversões (*inv*)	995	0.0688
Inserções (*ins*)	856	0.0592
Cromossomos dicêntricos (*dic*)	661	0.0457
Duplicações (*dup*)	512	0.0354
Pequenos fragmentos acêntricos (*min*)	499	0.0345
Regiões de coloração homogênea (*hsr*)	434	0.0300
Cromossomos em anel (*r*)	354	0.0245
Associações teloméricas (*tas*)	252	0.0174
Segmentos triplicados (*trp*)	30	0.0021
Cromossomos tricêntricos (*trc*)	4	0.0003
Fissões centroméricas (*fis*)	2	0.0001
Segmentos quadruplicados (*qdp*)	1	0.0001
Total de rearranjos cromossômicos	**56826**	**3.9280**

[1]Dados do banco Mitelman de aberrações citogenéticas (Mitelman Database, 2009) e quantificados pelo software KaryoComplex (Anexo). A relação de rearranjos cromossômicos segue o sistema internacional de nomenclatura citogenética para a espécie humana (ISCN, 1995). Demais rearranjos não estão representados no banco de cariótipos (***N***=soma dos rearranjos cromossômicos presentes nos cariótipos; ***C***= total de cariótipos).

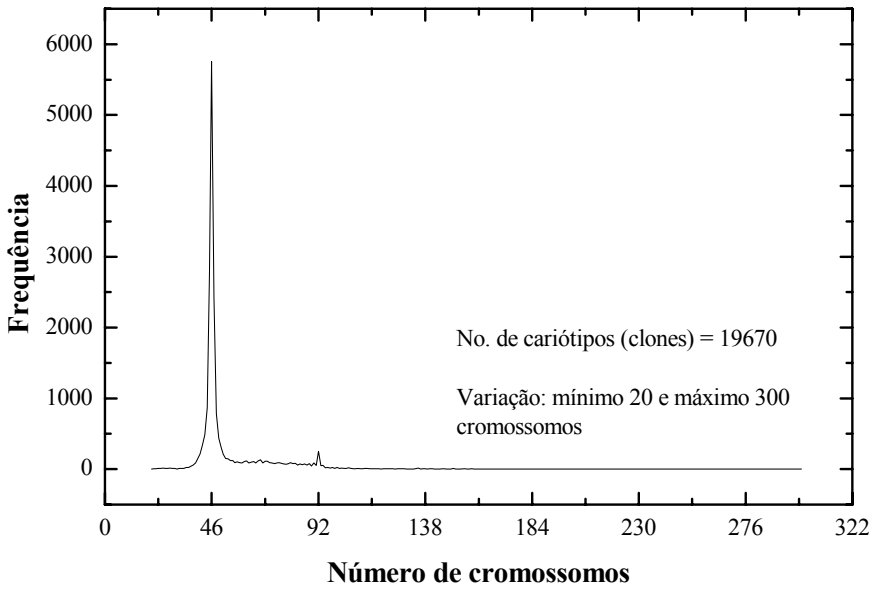


Figura 6. Espectro de aberrações numéricas em neoplasias sólidas. Estatística estimada a partir do banco Mitelman de aberrações citogenéticas (Mitelman Database, 2009).

Segundo estimativas do Censo de Genes de Câncer do Instituto Sanger (Futreal, 2009), 347 genes humanos possuem mutações somáticas catalogadas como causalmente relacionadas ao desenvolvimento do câncer. As mutações passageiras, por outro lado, estão presentes em baixa freqüência e não conferem vantagens adaptativas, mas podem ser distinguidas das variações esperadas para uma célula normal (*e.g.* polimorfismo de nucleotídeos). Estas alterações compõem a maior parte do espectro de alterações somáticas em neoplasias sólidas (Greenman et al., 2007).

Entretanto, a distinção entre mutações *causais* e *passageiras* pode ser bastante superficial quando considerado todo o espectro de alterações de uma amostra de câncer. Por exemplo, em um levantamento global, Wood & colaboradores (Wood et al., 2007) constataram um baixo nível de concordância de mutações causais em 11 amostras de câncer de cólon que tiveram seu genoma seqüenciado (o mesmo ocorrendo com câncer de mama), uma diferença que os autores sugerem ser a base da grande variabilidade tumoral.

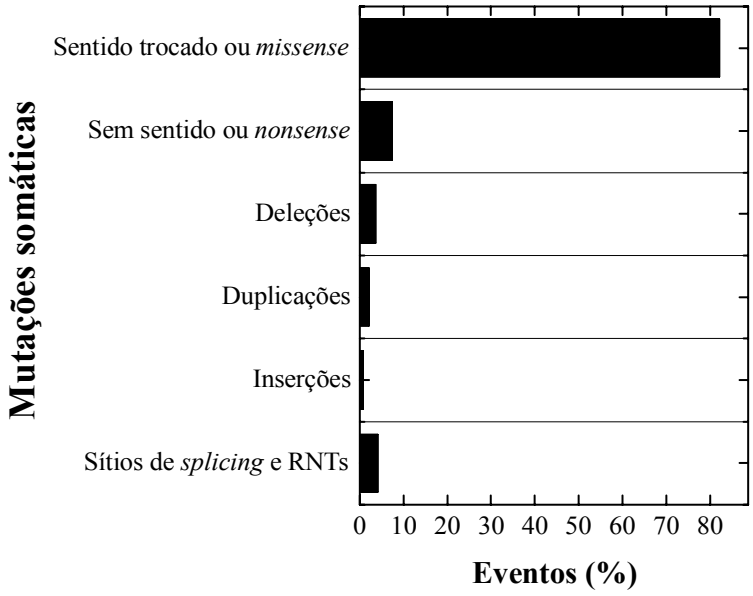


Figura 7. Padrão de mutações somáticas em carcinoma de cólon segundo Wood *et al.* (Wood et al., 2007). Número de genes analisados: 18191; número de genes mutados: 1031; número de mutações (eventos): 1241. RNTs: regiões não traduzidas.

Além disso, modificações epigenéticas podem alterar o funcionamento de genes associados à progressão tumoral sem que haja a necessidade de modificações na seqüência de nucleotídeos (Sawan et al., 2008). Por exemplo, alteração no padrão de metilação do DNA em regiões ricas em dinucleotídeos CG (ilhas CpG) constitui uma das modificações epigenéticas mais bem caracterizadas em câncer. Estas ilhas CpG ficam situadas preferencialmente nas regiões regulatórias dos genes e podem dificultar o acoplamento de fatores de transcrição (ou mesmo inativar genes) quando covalentemente modificadas pela adição de grupos metila (-CH3) (Feinberg et al., 2006). A hipermetilação de genes supressores tumorais parece ser um evento comum em câncer de cólon (Shen et al., 2007).

Padrões de acúmulo das alterações

Muitas vezes a expressão de um dado alelo mutado pode não contribuir para a progressão tumoral quando em heterozigose (*i.e.* em função da presença de um alelo normal). Entretanto, a perda do alelo normal e a conseqüente expressão do fenótipo alterado pode ser favorecida via recombinação mitótica, perda cromossômica ou alterações na estrutura dos cromossomos, um processo também conhecido como perda da heterozigosidade (Rajagopalan et al., 2003). Por exemplo, danos em parte de um cromossomo (ou perda de um cromossomo inteiro) são usualmente corrigidos por mecanismos de recombinação gênica cujo resultado é a duplicação do seguimento homólogo remanescente, criando duas cópias de um dado alelo (Hoeijmakers, 2001). Tal processo de perda da heterozigosidade pode ser acelerado em decorrência de um número maior de eventos recombinatórios quando uma célula se torna instável ao nível dos cromossomos (Rajagopalan et al., 2003).

Esse tipo de instabilidade genômica parece derivar de disfunções na duplicação e segregação cromossômica durante a mitose (Jallepalli & Lengauer, 2001). Embora a detecção da instabilidade cromossômica seja uma tarefa difícil, pois depende da quantificação das taxas

de aberrações cromossômicas (Lengauer et al., 1998), a presença de grande quantidade de aberrações já é um indício de disfunção dos mecanismos de manutenção dos cromossomos, um achado largamente descrito em cariótipos complexos (Mitelman Database, 2009) e uma das lesões mais características e acumuladas em neoplasias sólidas (Ricke et al., 2008). Em câncer colorretal um fator desencadeante pode ser a ocorrência de mutações no gene *APC*, uma vez que este gene está envolvido tanto no controle do fuso mitótico como no ancoramento dos cromossomos (Rusan & Peifer, 2008).

Outro fator desencadeante de instabilidade cromossômica pode ser a perda de telômeros – estruturas formadas por proteínas e seqüências nucleotídicas repetitivas do tipo TTAGGG situadas na extremidade dos cromossomos e que previnem ciclos de quebra, fusão, e rearranjo entre cromossomos (Deng et al., 2008; Murnane, 2006). Alguns trabalhos também sugerem que a aneuploidização já é fator suficiente para provocar a desestabilização dos cromossomas e o acúmulo de aberrações numéricas (Duesberg et al., 2005; Duesberg et al., 2000; Fabarius et al., 2003; Li et al., 2000).

Além das alterações decorrentes da instabilidade cromossômica, um segundo e menos freqüente padrão de acúmulo de lesões está bem descrito no modelo de progressão adenoma-carcinoma: alterações em regiões microssatélites. Microssatélites são seqüências de mono, di, tri, ou tetranucleotídeos repetidos (*i.e.*, CACACACA) encontradas em grandes quantidades espalhadas por todo o genoma e que, devido à sua natureza repetitiva, são propensas a erros de pareamento durante a replicação (Jiricny, 2006). Tumores com disfunção do sistema de reparo de erros de pareamento do ADN (sistema MMR) apresentam regiões microssatélites instáveis, com acúmulo de alterações de natureza nucleotídica (Aquilina & Bignami, 2001) da ordem de seis pares de base (Oda et al., 2005) e aumento na ocorrência de mutações *frameshift* (alteração na fase de leitura) em regiões gênicas repetitivas (Frank, 2007). Esse tipo

instabilidade genômica ocorre em 15% dos carcinomas de cólon e reto e reflete falhas em genes do sistema MMR (Peltomaki, 2001; Soreide et al., 2006).

Outro tipo de instabilidade ao nível dos nucleotídeos ocorre em alguns tipos raros de neoplasias sólidas em conseqüência de mutações nos genes do sistema de reparo por excisão de nucleotídeos (sistema NER) (Friedberg, 2001; Garfinkel & Bailis, 2002; Lengauer et al., 1998). Também referida como instabilidade NER (Lengauer et al., 1998), este tipo de instabilidade deixa como assinatura múltiplas mutações pontuais distribuídas aleatoriamente pelo genoma e serve de modelo para a teoria do *fenótipo mutador* (Loeb, 2001), a qual sustenta que a origem do câncer pode estar no aumento das taxas de mutações pontuais. Segundo esta teoria, danos no sistema de reparo de nucleotídeos seriam os eventos iniciais a partir dos quais todos os demais eventos necessários para a progressão tumoral seriam favorecidos (Loeb et al., 2003).

Entretanto, não existem catalogadas mutações somáticas causalmente associadas ao desenvolvimento do câncer em genes do sistema de reparo de nucleotídeos, um dado que enfraquece essa teoria (Futreal, 2009). Além disso, a quantificação das taxas de mutações pontuais é objeto de intenso debate e não há um consenso se ela está aumentada ou não na maioria das neoplasias sólidas (Klein, 2006; Loeb et al., 2003; Marx, 2002; Sieber et al., 2003; Wang et al., 2002).

Resultados contraditórios decorem em grande parte de diferenças metodológicas. Por exemplo, um levantamento realizado em escala genômica para estimar o perfil de mutações somáticas em câncer de cólon e mama (Sjoblom et al., 2006) mostrou que as taxas de mutação nestas neoplasias são equivalentes às de células normais, embora a técnica usada no estudo seja informativa para mutações submetidas à seleção e/ou expansão clonal. Por outro lado, após desenvolver técnicas focadas na captura do espectro de alterações randômicas (Bielas & Loeb, 2005), Bielas et al. (Bielas et al., 2006) mostraram que em carcinoma de cólon (e outros

quatro tipos de neoplasias sólidas) a freqüência de mutações somáticas está aumentada em até 200X quando comparada ao tecido normal.

Todos estes padrões de acúmulo de alteração ainda são submetidos à seleção/expansão clonal, um processo microevolutivo que tem início no interior do tumor primário e envolve a formação de múltiplas linhagens. Por exemplo, considere a progressão tumoral ilustrada na **Figura 8.** Neste cenário, a diversidade celular flutua (Figura 8A) à medida que a população de células diverge e forma novas linhagens (Figura 8B).

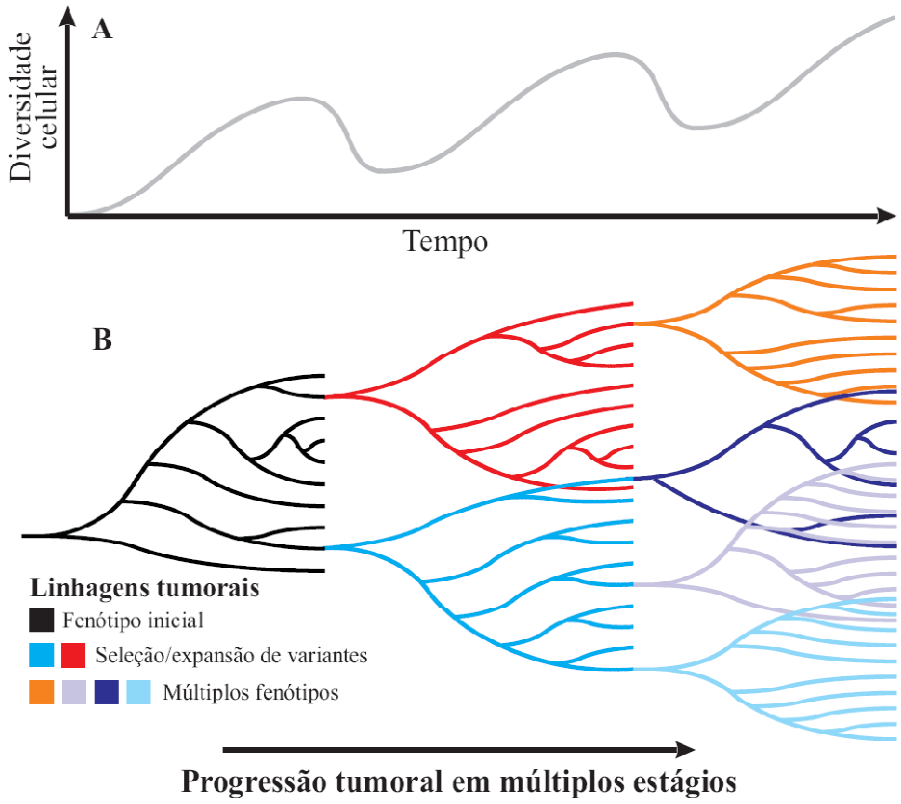


Figura 8. Ondas de divergência, expansão e seleção clonal. Nesta representação, a diversidade celular flutua (**A**) à medida que o tumor progride (**B**) com o surgimento de novos clones e a morte de linhagens menos adaptadas ao microambiente tumoral [Adaptado de (Merlo et al., 2006)].

Tal flutuação decorre de fatores opostos que incidem sobre heterogeneidade tumoral. De um lado, características fenotípicas podem ser fixadas na população de células a partir da expansão de linhagens dominantes, reduzindo a diversidade celular (Shibata, 2006); de outro,

a disfunção dos mecanismos de estabilização do genoma (discutida acima) aumenta a diversidade genética e resulta em novas variantes celulares. Isso estabelece ciclos de divergência, expansão e seleção clonal, o que favorece a formação de múltiplas linhagens, pois estas podem coexistir – sem homogeneização – dentro dos limites espaciais do tumor primário (Gonzalez-Garcia et al., 2002) ou migrar para novos sítios e formar metástases. Assim, novas características fenotípicas e genotípicas podem emergir em diferentes linhagens, em diferentes períodos, originando tumores com subpoluções geneticamente distintas, o que resulta em aumento da heterogeneidade global (Maley et al., 2006; Merlo et al., 2006).

Justificativa

Com a finalidade de sumarizar o exposto acima listamos a seguir algumas questões que motivaram os objetivos delineados nesta tese:

i. Baixa concordância entre eventos enunciados nos modelos de progressão tumoral ***apesar*** da intensa busca por modelos capazes de auxiliar na estratificação de pacientes;
ii. Raridade de biomarcadores moleculares validados para o estadiamento tumoral ***apesar*** do largo espectro de alterações citogenéticas e moleculares;
iii. Achados citogenéticos e moleculares incapazes de refletir padrões epidemiológicos ***apesar de*** diversas anormalidades refletirem o aumento da heterogeneidade das neoplasias sólidas;
iv. Descrição do estadiamento baseada unicamente na extensão atômica da doença para vasta maioria dos tumores ***apesar*** da abundância de métodos para detecção de anomalias citogenéticas e moleculares;
v. Progressão do câncer não universal e seqüência molecular de múltiplos desfechos ***apesar de*** emergir ao nível populacional um padrão previsível;
vi. Concordância de mutações tecido-específicas com o início de uma seqüência provável de eventos, ***apesar de*** – na ocasião do diagnóstico – não serem encontradas mutações na maior parte dos casos;
vii. Abundante fonte de variabilidade genética decorrente da disfunção dos mecanismos de manutenção do genoma ***apesar de*** ocorrer redução de variabilidade via seleção clonal, de um lado, e aumento de variabilidade celular via formação de múltiplas linhagens, de outro.

Objetivo geral

Dada as implicações clínicas da heterogeneidade para o prognóstico de neoplasias sólidas e a reconhecida barreira para o desenvolvimento de potenciais biomarcadores, este trabalho teve como ***objetivo geral*** estudar padrões diversidade tumoral.

Objetivos específicos

Capítulos I e II – explorar padrões diversidade citogenética de cariótipos de neoplasias sólidas usando o formalismo da entropia de Shannon.

Capítulos III e *IV* – estudar a diversidade de expressão gênica em redes de genes potencialmente responsável pelo aumento da heterogeneidade tumoral.

Capítulo V – desenvolver métodos computacionais para mapear a diversidade de expressão gênica dos mecanismos de estabilização do genoma.

Capítulo VI – testar estes métodos computacionais para o desenvolvimento de novos biomarcadores tumorais, bem como avaliar seu desempenho em modelos não correlacionados.

Capítulos VII / anexos – padronizar a metodologia desenvolvida para uso prognóstico e aplicações diversas.

PARTE II

ARTIGOS

Capítulo 1: *Investigação do perfil da diversidade citogenética com base na entropia de Shannon*.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 234 (2005) 487–495
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Profiling cytogenetic diversity with entropy-based karyotypic analysis

Mauro A.A. Castro[a,b,c,*], Tor T.G. Onsten[b,c], Rita M.C. de Almeida[d], José C.F. Moreira[a]

[a]*Departamento de Bioquímica, ICBS, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600–anexo (Lab. 25), Porto Alegre 90035-003, Brazil*
[b]*Departamento de Medicina Interna, Hospital de Clínicas de Porto Alegre, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2350, Porto Alegre 90035-903, Brazil*
[c]*Universidade Luterana do Brasil, Rua Miguel Tostes 101, Canoas 92420-280, Brazil*
[d]*Instituto de Física, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre 91501-970, Brazil*

Received 5 October 2004; received in revised form 29 November 2004; accepted 6 December 2004
Available online 29 January 2005

Abstract

Regardless the high degree of variation observed in solid tumor karyotypes, the use of diversity as a measurable phenomenon remains unexplored. Here we present a systematic cytogenetic analysis using Shannon's entropy as a measure for karyotypic diversity. Karyotypes from 14 epithelial tumor types ($n = 1232$) have the aneuploidy status scored, resulting in highly asymmetrical sample distribution, for which we determine the index of diversity (Shannon H') of structural and numerical chromosomal aberrations. Since karyotypic diversity is tissue-specific, this approach may give rise to new insights into the processes that may account for aneuploidy progression and solid tumor outcomes.

Keywords: Solid tumor; Karyotype; Information theory; Entropy; Cancer; Cytogenetics; Chromosome aberrations

1. Introduction

The majority of human carcinomas exhibit abnormal genetic content, called aneuploidy, for which two main karyotypic abnormalities have been described: structural chromosomal rearrangement (structural aberrations) and abnormal numbers of chromosomes (numerical aberrations) (Rajagopalan et al., 2003).

It has been known for a long time that chromosome abnormalities increase with tumor progression, some times with massive cytogenetic changes (Klein et al., 2002; Matzke et al., 2003). A remarkable spectrum of variability (of complex karyotypes) is observed not only in cancer cells of different tumors but also within the same histopathological type, indicating that abnormality generation is intrinsic to the dynamics underlying tumor growth. As a consequence, all known recurrent cytogenetic aberrations are not of clinical significance, having limited diagnostic and prognostic value (Albertson et al., 2003; Mitelman et al., 1997). This is in deep contrast with hematological neoplasms, in which particular alterations (i.e. translocations and deletions of defined chromosome segment) are clinically significant to predict cancer outcome (Mitelman, 2000).To explain the karyotypic disorder, numerous authors maintain that solid tumors manifest intrinsic genetic instability, characterized by the increase in chromosomal alteration rate per cell division in comparison with normal cells, but it remains a controversial point (Fabarius et al., 2003; Furuya et al., 2000; Hahn and Weinberg, 2002; Marx, 2002; Ribas et al., 2003; Sieber et al., 2003; Stoler et al., 1999).

Whatever the process that may account for the increase in aneuploidy, it is convenient, however, to estimate the complexity of chromosome abnormalities by a measurable, well defined quantity. This is a delicate task; one must first define an order parameter since in works previously reported in the literature, the

*Corresponding author. Tel.: +55 51 3316 5549; fax. +55 51 3316 5540.
E-mail address: mauro@ufrgs.br (M.A.A. Castro).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.12.006

cytogenetic complexity is qualitatively described, but not quantitatively measured (Albertson et al., 2003; Choma et al., 2001; Fabarius et al., 2003; Furuya et al., 2000; Hahn and Weinberg, 2002; Klein et al., 2002; Marx, 2002; Matzke et al., 2003; Mitelman, 2000; Rajagopalan et al., 2003; Ribas et al., 2003; Sieber et al., 2003).

In exploring the possible candidate methods to quantitatively estimate the karyotype degree of complexity, one could naively try the direct examination of a population of karyotypes, by using common methods of statistical analysis. However, this route presents some unavoidable difficulties. For example, the frequency distribution of chromosome rearrangements in breast and colorectal tumors has been shown to acquire extremely asymmetrical and skewed sample outlines (Hoglund et al., 2002b, a), which prohibit the use of standard approaches for statistical analysis.

More than discussing the causes of the aneuploidy, here we focus in the form as it presents itself. Using concepts from information theory, we propose Shannon's entropy (Shannon, 1948) as a measure for karyotypic complexity; adjusted to describe the diversity of chromosome aberrations, this analysis may give rise to new insights into the processes that may account for aneuploidy progression and solid tumor outcomes.

2. Methods

2.1. Data selection

Epithelial tumor types with representative number of karyotypes were retrieved from the Mitelman Database of Chromosome Aberrations in Cancer, unselected series (Mitelman Database, 2004). Forty-five epithelial tumor types were available in Mitelman Database by February of 2004. From these, 31 tumor types were excluded due to small number of cases. Therefore, 14 epithelial tumor types were selected ($n = 3281$ karyotypes): adenocarcinoma of breast ($n = 633$), kidney ($n = 713$), large intestine ($n = 332$), lung ($n = 158$), ovary ($n = 341$), pancreas ($n = 91$), prostate ($n = 190$), stomach ($n = 115$), thyroid ($n = 127$) and uterus, corpus ($n = 88$); basal cell carcinoma, skin ($n = 98$); squamous cell carcinoma of larynx ($n = 90$) and lung ($n = 144$); and transitional cell carcinoma of bladder ($n = 161$). Cases with incomplete karyotypes were excluded from the analysis (i.e. questionable identification of a chromosome—with question mark (?) code). The evaluation of heterogeneity presented in the current paper is best restricted to groups divided into the same number of categories and with equal group size (Feinstein, 2002; Magurran, 1988; Ricotta, 2003). Therefore, tumor types with many cases (outsized) were randomized to produce reduced samples with 88 karyotypes per tumor type (number of karyotypes of adenocarcinoma of the uterus), with a final sample size of 1232 karyotypes; the goal was to scale the larger sample down to the size of the smallest one.

2.2. Karyotypic analysis

The karyotypes were assessed for numerical and structural chromosomal abnormalities: the numerical aberration status of a karyotype is expressed by the ploidy index (PI) (Fabarius et al., 2003), which is the chromosome number divided by 46, and the structural aberration status of a karyotype is expressed by sum of structural aberrations (SSA)—counted in agreement with the ISCN Abbreviated Terms and Symbols (Mitelman, 1995). Therefore, each karyotype received a PI and a SSA score (code-value). The cases with the same code are pooled and the resulting histograms are plotted for analysis of the karyotypic distributions.

2.3. Diversity and statistical analyses

To obtain a quantitative expression of the complexity of the sample distribution, we have measured the information content of the pooled karyotypes using Shannon information or, rather, its positive value, Shannon entropy (Shannon, 1948). Genetic information can be calculated using Shannon's function as demonstrated theoretically (Gatenby and Frieden, 2002; Kendal, 1990). We adjusted the approach to the analysis of karyotypic data as follows:

$$(\text{Shannon}H')\quad H' = -\sum p_i \log(p_i),\quad \text{such that}\quad p_i > 0$$

where p is the probability (frequency) of occurrence of a code-value i (PI or SSA). The probabilities are obtained from the histogram distribution analysis and the resulting Shannon H' value expresses the degree of diversity of the sample distribution. Considering the asymmetric distribution observed for the karyotypic abnormalities, such a method provides an unbiased analysis of diversity when comparing with more standard variation parameters, which are restricted to Gaussian distributions. *Statistical analysis*: to demonstrate stochastic boundaries among abnormalities, 95% confidence intervals are constructed. The tumor types are first pooled into numerical and structural diversity groups, from which random sampling is done with replacement (bootstrapped) 2000 times (Manley, 1998). The pooled Shannon H^* is scored after each resampling and the standard bootstrap confidence limits are then calculated as the estimate pooled $H^* \pm Z_{\alpha/2}$ (bootstrap standard deviation). An approximation of the bias of the estimator is calculated as the difference between the estimator mean from the bootstrapping and the estimate from the pooled sample that is bootstrapped (Manley, 1998), so the bias-corrected mean is obtained. The basic

argument to apply bootstrap method rises from the nature of the diversity estimator under study, given that Shannon H' statistic does not produce the usual unbiased estimator represented by the sample variance (Magurran, 1988). Indeed, bootstrap strategy is valid for any distribution, once the parental sample-size is adequate (Carpenter and Bithell, 2000; Manley, 1998). Although 88 cases could represent a quite small sample, for individual tumor types, in the absence of a better estimator, the standard errors—only—are obtained by bootstrapping; while H' values are maintained as in the original form without adjustment. To identify stochastic inter-group differences (between numerical and structural chromosomal aberrations), we use the Mann–Whitney U test and two-sample randomization test (for combined data) 5000 times (Feinstein, 2002; Manley, 1998). Resampling Stats^{TM} software for Microsoft Excel (Arlington, VA, USA) is used for bootstrap and randomization procedures; and U test is carried out in the non-parametric package of the SPSS software (Chicago, IL, USA) with the two-independent samples method. We considered $P<0.05$ significant.

2.4. Epidemiological data and statistical analyses

Histopathological and topographical nomenclatures of epithelial tumors are primarily reviewed and organized, following the International Classification of Diseases for Oncology (Percy et al., 1990), and the results are used to retrieve epidemiological parameters from the Surveillance, Epidemiology, and End Results Program (SEER, 2004). Relative and absolute 5-year survival rate $\pm$ SEM for each of the 14 tumor types are obtained through the SEER*Stat software (Surveillance Research Program, 2004) at SEER Cancer Statistic Review Database 1992–1999 (SEER, 2004). *Statistical analysis*: the two-tailed t test is used to compare the survival data. Epidemiological parameters are also crossed with Shannon H' estimates for categorical association analysis. To calculate the discrepancy between the two sets of corresponding databases, we applied the Spearman's Rho non-parametric method in the SPSS software. To verify a possible sample bias between the number of cases in the parental database and the parameters that are estimated, we performed Spearman's Rho non-parametric correlation analysis (Fig. 6B and C). Significance is set as specified in diversity and statistical analyses.

3. Results and discussion

3.1. Shannon's entropy as a measure for karyotypic diversity

After assessing chromosome abnormalities in 1232 karyotypes from 14 solid tumor types available in the Mitelman Database of Chromosome Aberrations in Cancer (Mitelman Database, 2004), we obtained the frequency distributions of structural (SSA; Fig. 1A) and numerical aberrations (PI; Fig. 1B). To obtain a quantitative expression of the sample distribution (karyotypic diversity), we have measured the information content of the pooled karyotypes using Shannon information or, rather, its positive value, Shannon entropy (Shannon, 1948).

In colloquial usage, information denotes knowledge; in information theory it denotes a measurable quantity (Shannon H'). For example, an increase in heterogeneity of some chromosome abnormality implies an increase in disorder, or entropy, and loss of information about the cell population (Gatenby and Frieden, 2002). By definition, entropy measures variation or diversity in a distribution of items or events; unchanging patterns—such as cancers with no karyotypic diversity—have karyotypic distribution with zero entropy. Cancers that express different patterns (of karyotypes), on the other hand, have karyotypic distribution entropy greater than zero, indicating higher levels of diversity (Fuhrman et al., 2000; Kendal, 1990). Alternatively, Shannon's function (see methods) gives an estimate on how uncertain we are about a given event in a population, whatever the characteristics of the distribution of cases (symmetrical, skewed, bimodal, etc). To illustrate the information theory analysis, the PI and SSA frequencies

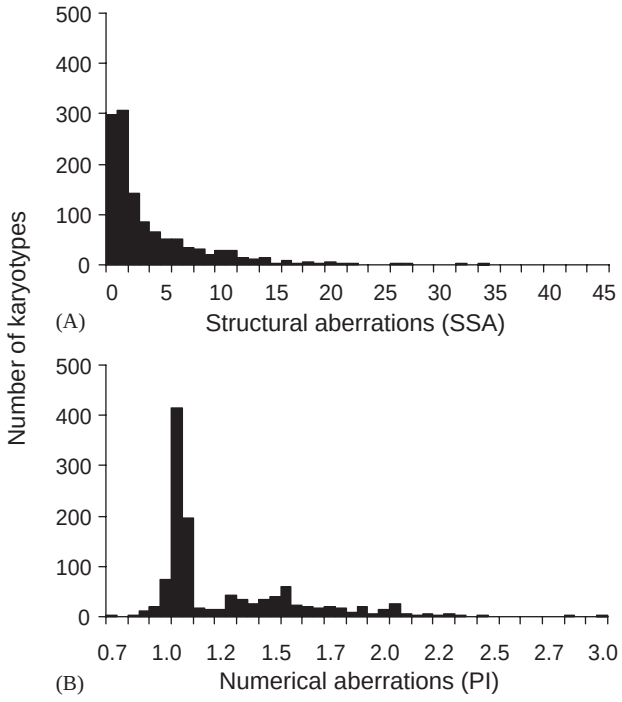


Fig. 1. Frequency distribution of solid tumor karyotypes. Structural (A) and numerical (B) aberrations of 1232 karyotypes from 14 epithelial tumor types ($n=88$ karyotypes per tumor type) were counted and then the karyotypes were pooled following the scores obtained for SSA and PI.

from 14 solid tumor types are pooled in histogram charts (Figs. 2 and 3). For example, the lung adenocarcinoma histograms in Fig. 2 show a more heterogeneous distribution, while prostate adenocarcinoma histograms are limited to a small number of ploidy zones: near-diploid (PI ≈ 1.0), near-triploid (PI ≈ 1.5) and near-tetraploid (PI ≈ 2.0), with similar interpretation for SSA distributions in Fig. 3.

Information theory states that lung adenocarcinoma carries less information or higher entropy, implying

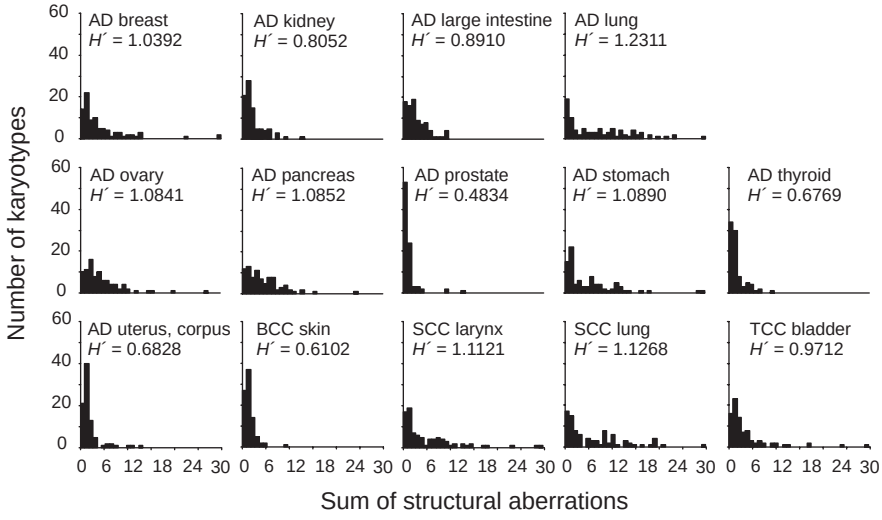


Fig. 2. Distribution of structural aberrations. The karyotypes of 14 epithelial tumor types are pooled according to SSA to obtain the histogram ($n = 88$ karyotypes per tumor type). Epithelial tumor types and corresponding index of diversity (Shannon H') are indicated; the resulting Shannon H' value characterizes the histogram dispersion.

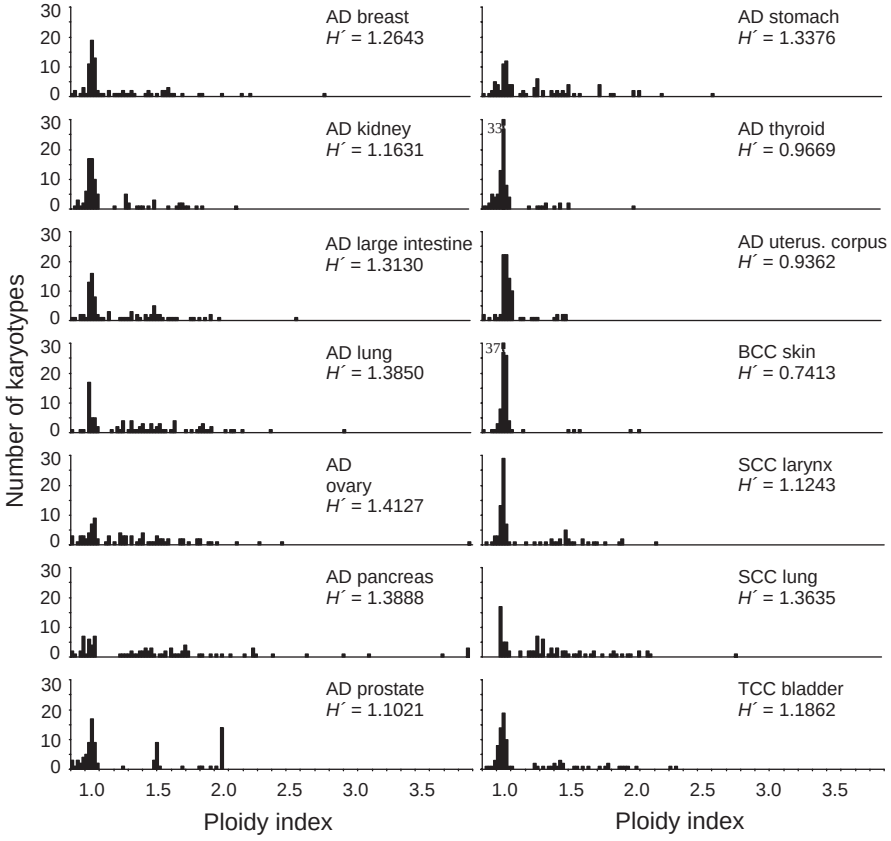


Fig. 3. Distribution of numerical aberrations. The karyotypes of 14 epithelial tumor types are pooled according to PI, which is the chromosome number divided by 46, to obtain the histogram ($n = 88$ karyotypes per tumor type). Epithelial tumor types and corresponding index of diversity (Shannon H') are indicated; the resulting Shannon H' value characterizes the histogram dispersion.

Table 1
Analysis of karyotypic diversity

	n (n)[a]	Numerical aberration[b]	Structural aberration[b]
		$H' \pm$SEM	$H' \pm$SEM
Adenocarcinoma, ovary	341 (88)	1.4127±0.034	1.0841±0.041
Adenocarcinoma, pancreas	90 (88)	1.3888±0.033	1.0852±0.050
Adenocarcinoma, lung	158 (88)	1.3850±0.040	1.2311±0.051
Adenocarcinoma, stomach	119 (88)	1.3376±0.044	1.0890±0.045
Adenocarcinoma, large intestine	344 (88)	1.3130±0.032	0.8910±0.054
Adenocarcinoma, breast	617 (88)	1.2643±0.048	1.0392±0.059
Adenocarcinoma, kidney	713 (88)	1.1631±0.039	0.8052±0.050
Adenocarcinoma, prostate	190 (88)	1.1021±0.048	0.4834±0.044
Adenocarcinoma, thyroid	124 (88)	0.9669±0.046	0.6769±0.057
Adenocarcinoma, uterus, corpus	88 (88)	0.9362±0.045	0.6828±0.052
Squamous cell carcinoma, lung	144 (88)	1.3635±0.041	1.1268±0.049
Squamous cell carcinoma, larynx	90 (88)	1.1243±0.049	1.1121±0.063
Transitional cell carcinoma, bladder	157 (88)	1.1862±0.045	0.9712±0.057
Basal cell carcinoma, skin	98 (88)	0.7413±0.038	0.6102±0.056
	Contrasting aberrations		
	n[c]	$H^* \pm$SEM[d]	95% CI	P[e]
Numerical aberration	1232	1.388±0.015	1.354–1.423	<0.001
Structural aberration	1232	1.021±0.015	0.991–1.052	

[a]Parental sample with n cases in the Mitelman Database of Chromosome Aberrations in Cancer (Mitelman Database, 2004), from which (n) cases are obtained by chance for analysis of diversity.
[b]Only standard errors are estimated by bootstrap procedure.
[c]Overall sample-size.
[d]Estimates (mean and error) obtained by bootstrap procedure.
[e]Significance determined by randomization test (5000×). Equivalent significance levels are obtained by Mann–Whitney U test (Fig. 5A).

higher diversity. It means that knowing that a cell comes from a lung adenocarcinoma we are less informed about its karyotype than in the case of a cell coming from a prostate tumor. Translating this visual inspection into Shannon H' diversity we can compare the histogram distributions and thus heterogeneity through a quantifiable measure. As shown in Table 1, the 14 epithelial tumor types are analysed and have Shannon H' estimated for numerical and structural aberration distributions. Since this analysis does not produce the usual unbiased estimator of sample variance, the standard errors of the diversity indices were estimated by bootstrap procedures (Fig. 4).

3.2. Comparing chromosomal aberrations

Under the null hypothesis, the observed diversity indices are the same between numerical and structural abnormalities. To test this statement, we obtained confidence boundaries for each chromosomal aberration type. We have found that numerical aberration distribution has a larger diversity index as compared with structural aberration ones (Table 1, contrasting aberrations, $P<0.001$; Fig. 5A, $P<0.002$), indicating that the former is the most heterogeneous and disordered karyotypic abnormality.

Chromosome abnormality is a ubiquitous feature in solid tumors and it can be self catalytic in the sense that

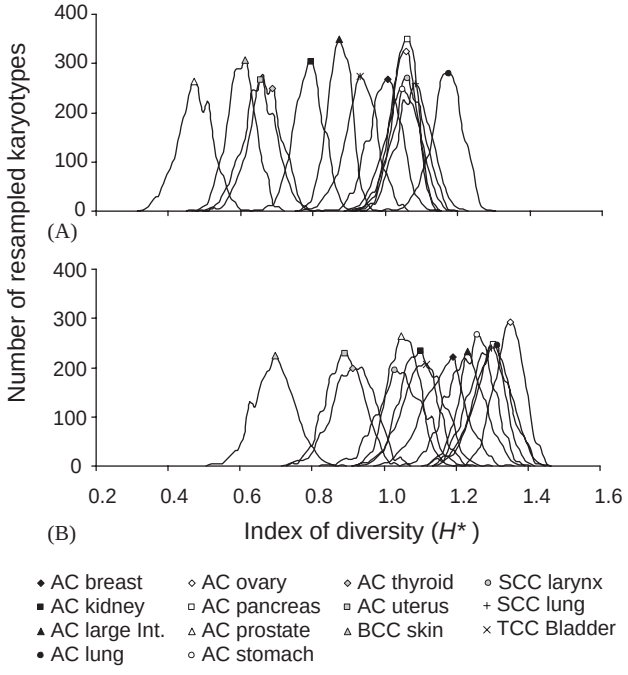
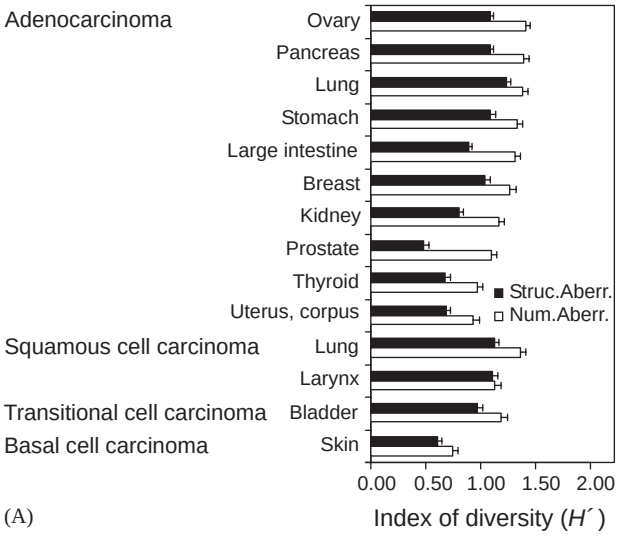


Fig. 4. Results of bootstrap analysis. Frequency plots of numerical (A) and structural aberrations (B) show the distribution of Shannon H^* values. Sampling errors were obtained by random sampling with replacement to estimate the stability (errors) of the diversity indices; AC (adenocarcinoma), BCC (basal cell carcinomas), SCC (squamous cell carcinoma), and TCC (transitional cell carcinomas). The standard error of the estimator H' is available in Table 1 for each distribution.

it begets further genetic abnormality (Li et al., 2000). In order to either avoid tumor emergence or suggest prophylactic procedures it is essential to understand what phenomena may trigger the first genetic abnormalities, probably the ones responsible for cancer inception (Nowak et al., 2002). It is largely accepted that once some genetic abnormality is settled, further genetic abnormalities are favored (Loeb et al., 2003), and the resulting karyotype distribution in tumor types must be deeply dependent of the abnormality generation process inherent of tumor growth.

It is then valid to question whether or not diversity could reveal, in some way, differences between the accumulating processes that preceded and produced the chromosome abnormalities documented in the karyotypes of a tumor type sample. Considering the course of two independent stochastic processes with the same abnormality accumulation rate, it might be expected that the longer (or the earlier) accumulates more changes and, thus, presents higher heterogeneity levels. Nevertheless, numerical and structural aberrations may arise from interdependent processes, since numerical aberrations must lead to an increase in structural aberrations previously formed, when it happens in chromosomes that already present structural changes. This argument indicates that one should expect a correlation between numerical and structural aberration diversity. Indeed, the simple correlation between diversities is clearly observed in Fig. 5B, with a positive and significant correlation between abnormalities (Spearman's Rho $r_s = 0.745$, $P = 0.002$).



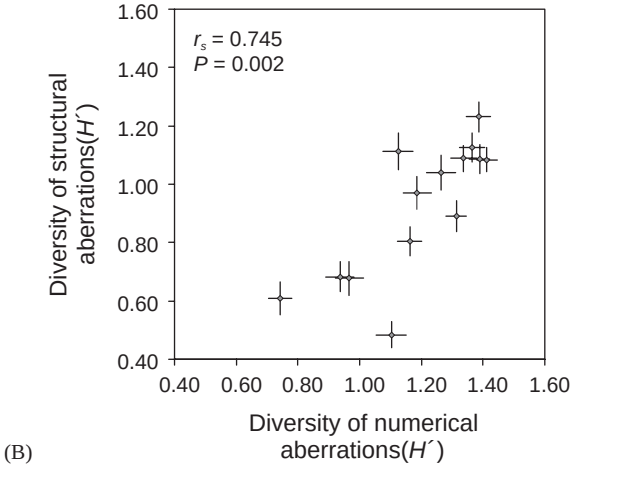


Fig. 5. Comparing chromosomal aberrations. (A) Index of diversity (Shannon H') of numerical (white bars) and structural aberrations (black bars). Numerical aberrations show a higher diversity as compared to structural aberrations ($P = 0.002$; Mann–Whitney U test). (B) Scatter plot of numerical and structural aberrations. The index of categorical association between numerical and structural aberrations (Spearman's Rho) and significance are indicated.

3.3. Biological applicability for karyotypic diversity analysis

Despite the fact that diversity can be a valuable instrument to formulate hypothesis about underlying events, diversity values are merely numbers and their relevance, as with any other statistics, must be judge on the bases of observed correlations with other variables (Ricotta, 2003).

As an example of how information theoretic methods can be used, consider the common knowledge that more aggressive solid tumors tend to have a larger number of chromosomal changes with more complex karyotypes (Mitelman et al., 1997). Hence, tumor types that show a large proportion of cases with highly complex karyotypes (e.g. lung carcinomas) are expected to be more aggressive than tumor types that show frequent simple karyotypes (e.g. bladder cancer) (Ried et al., 1999). To verify if information theory can predict this statement, we compared the diversity indices with survival rates available in epidemiological reports; accepting that there exist unavoidable technical limitations with this sort of analysis, for the reason that survival statistics are not available in the Mitelman database.

With these concepts in mind, we reviewed the histopathological and topographical classifications of the 14 epithelial tumors types to access epidemiological data from the Surveillance, Epidemiology, and End Results Program (SEER, 2004) (Table 2). Examining the scatter plot of diversity indices as a function of the survival data (Fig. 6A) we have found a significant correlation for both numerical and structural chromosome aberrations (Spearman's Rho $r_s = 0.820$, $P = 0.00033$; and $r_s = 0.842$, $P = 0.00016$, respectively). Although further support would require a more stringent analysis, without biased groups, the present data cannot be attributed to a bias in the sample-size presented in the parental databases (Fig. 6B and C), since the trends observed have no correlation with

Table 2
Cancer 5-year relative survival

	5-year survival[a]		
	n[b]	Obs ± SEM	Rel ± SEM
Adenocarcinoma, ovary	6094	0.368 ± 0.015	0.420 ± 0.017
Adenocarcinoma, pancreas	13542	0.035 ± 0.004	0.043 ± 0.005
Adenocarcinoma, lung	43804	0.180 ± 0.005	0.214 ± 0.005
Adenocarcinoma, stomach	14085	0.178 ± 0.008	0.229 ± 0.010
Adenocarcinoma, large intestine	66093	0.502 ± 0.005	0.642 ± 0.006
Adenocarcinoma, breast	6351	0.685 ± 0.014	0.777 ± 0.016
Adenocarcinoma, kidney	18455	0.538 ± 0.009	0.628 ± 0.010
Adenocarcinoma, prostate	165149	0.765 ± 0.003	0.979 ± 0.003
Adenocarcinoma, thyroid	6969	0.924 ± 0.008	0.973 ± 0.008
Adenocarcinoma, uterus, corpus	23392	0.796 ± 0.007	0.899 ± 0.007
Squamous cell carcinoma, lung	24296	0.136 ± 0.005	0.168 ± 0.007
Squamous cell carcinoma, larynx	6974	0.543 ± 0.015	0.644 ± 0.017
Transitional cell carcinoma, bladder	10063	0.456 ± 0.012	0.600 ± 0.016
Basal cell carcinoma, skin	268	0.739 ± 0.068	0.956 ± 0.088

[a]Epidemiological data: Relative (Rel) and absolute (observed—Obs) cancer 5-year survival from Surveillance, Epidemiology, and End Results Program (SEER, 2004). Data are constructed through the merged option available in the SEER*Stat software (Surveillance Research Program, 2004); retrieve parameters are consistent with histopathological and topographical nomenclatures of epithelial tumors specified in Mitelman database and ICD-O-2 (Percy et al., 1990).
[b]Number of cases in SEER database.

neither diversity index nor survival ($-0.15 <$ Spearman's Rho $r_s < 0.28$).

Even though karyotypic diversity does not add more information than the mean survival time—since both measures are highly correlated—karyotypic diversity can be earlier acquired, as karyotypes are investigated on the occasion of the diagnosis, while mean survival time is obtained a posteriori (i.e. 5 year follow-up). Such difference could give a chronological advantage to assess large series of cytogenetic data using Shannon entropy, which may contribute to estimate the likely outcome or course of a solid tumor type. This brief consideration, however, does not pretend to arbitrate against diversity or survival statistics; instead, it suggests a biological applicability for karyotypic diversity analysis.

3.4. Possible implications

The more interesting question is why the tumor diversity described here has shown inverse correlation with survival. Is the correlation due to chromosomal instability? Are unstable phenotypes more aggressive because they are more adaptable and evolve more quickly? An alternative hypothesis might be that tumor types with more heterogeneous phenotypes are simply older and have had a longer time to form metastases.

In attempt to answer these questions, some suggestions could arise from tumor biology. Considering that diversity is also present among tumor cells (Loeb et al., 2003), tumor type diversity could be emerging as an end result of the intratumoral heterogeneity. Indeed, previous works have shown intratumoral heterogeneity (also intrasample) in corectal, breast, prostate, gastric, uterus, and squamous cell carcinomas, particularly at the level of chromosomal imbalances. (Flyger et al., 1999; Fujii et al., 2000; Fujimaki et al., 1996; Furuya et al., 2000; Jin et al., 2002; Katsura et al., 1996; Klein et al., 2002; Osterheld et al., 1999). In this perspective, one is tempted to suggest a sequence of events that begins with chromosomal instability (Rajagopalan et al., 2003), which leads to unbalanced genetic alterations within cell population (via multiple mutations and aneuploidization) and subsequently to adaptive shifts in tumor dynamics with impact on survival rates, i.e. intratumoral clonal selection improving proliferation, invasion and metastases (Albertson et al., 2003; Cahill et al., 1999; Sieber et al., 2003; Storchova and Pellman, 2004). Accordingly, unstable phenotypes could have the potential to increase not only intratumoral heterogeneity but also tumor type diversity. However, it remains to be determined whether diversities within- and between-tumors are actually related to each other and both with survival.

4. Conclusion

Shannon H' index has been used in the environmental sciences and evolution as measure of biological diversity (Adami et al., 2000; Cowell et al., 1998; Magurran, 1988; Ricotta, 2003). In the near past, the major difficulty to use this strategy in cytogenetics was the lack of sufficient

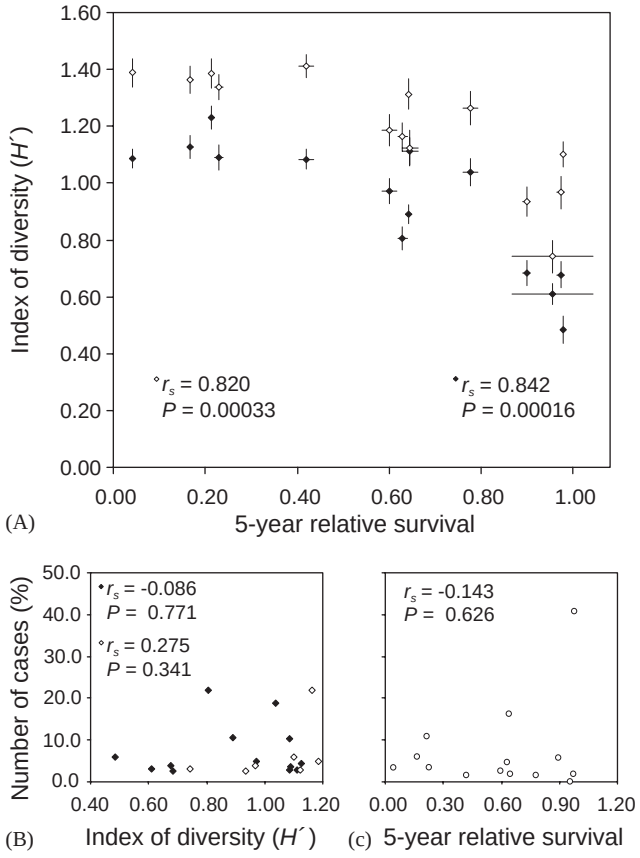


Fig. 6. Categorical association between karyotypic diversity and survival. (A) Scatter plot of numerical (◇) and structural aberration diversity (◆) as a function of cancer 5-year relative survival. (B) Analysis of sample-size bias in karyotypic data: association between number of karyotypes in parental database (Mitelman Database, 2004) and corresponding index of diversity presented in Table 1. (C) Analysis of the sample-size bias in survival data: association between number of cases in the parental database (SEER, 2004) and corresponding cancer 5-year relative survival data presented in Table 2. Spearman's Rho correlation and significance levels are indicated.

data to describe chromosome abnormalities of solid tumors. With the large amount of information available, from now, cytogenetic studies can make use of different methods to estimate diversity of chromosome abnormalities in several other tumor types and deal with karyotypic complexity in a systematic way.

References

Adami, C., Ofria, C., Collier, T.C., 2000. Special Feature: Evolution of biological complexity. Proc. Natl Acad. Sci. USA 97, 4463–4468.

Albertson, D.G., Collins, C., McCormick, F., Gray, J.W., 2003. Chromosome aberrations in solid tumors. Nature Genet. 34, 369–376.

Cahill, D.P., Kinzler, K.W., Vogelstein, B., Lengauer, C., 1999. Genetic instability and darwinian selection in tumours. Trends Cell Biol. 9, M57–M60.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 19, 1141–1164.

Choma, D., Daures, J.P., Quantin, X., Pujol, J.L., 2001. Aneuploidy and prognosis of non-small-cell lung cancer: a meta-analysis of published data. Br. J. Cancer 85, 14–22.

Cowell, L.G., Kepler, T.B., Janitz, M., Lauster, R., Mitchison, N.A., 1998. The distribution of variation in regulatory gene segments, as present in MHC class II promoters. Genome Res. 8, 124–134.

Fabarius, A., Hehlmann, R., Duesberg, P.H., 2003. Instability of chromosome structure in cancer cells increases exponentially with degrees of aneuploidy. Cancer Genet. Cytogenet. 143, 59–72.

Feinstein, A.R., 2002. Principals of Medical Statistics. Chapman & Hall/CRC, Boca Raton.

Flyger, H.L., Larsen, J.K., Nielsen, H.J., Christensen, I.J., 1999. DNA ploidy in colorectal cancer, heterogeneity within and between tumors and relation to survival. Cytometry 38, 293–300.

Fuhrman, S., Cunningham, M.J., Wen, X.L., Zweiger, G., Seilhamer, J.J., Somogyi, R., 2000. The application of Shannon entropy in the identification of putative drug targets. Biosystems 55, 5–14.

Fujii, H., Yoshida, M., Gong, Z.X., Matsumoto, T., Hamano, Y., Fukunaga, M., Hruban, R.H., Gabrielson, E., Shirai, T., 2000. Frequent genetic heterogeneity in the clonal evolution of gynecological carcinosarcoma and its influence on phenotypic diversity. Cancer Res. 60, 114–120.

Fujimaki, E., Sasaki, K., Nakano, O., Chiba, S., Tazawa, H., Yamashiki, H., Orii, S., Sugai, T., 1996. DNA ploidy heterogeneity in early and advanced gastric cancers. Cytometry 26, 131–136.

Furuya, T., Uchiyama, T., Murakami, T., Adachi, A., Kawauchi, S., Oga, A., Hirano, T., Sasaki, K., 2000. Relationship between chromosomal instability and intratumoral regional DNA ploidy heterogeneity in primary gastric cancers. Clin. Cancer Res. 6, 2815–2820.

Gatenby, R.A., Frieden, B.R., 2002. Application of Information Theory and Extreme Physical Information to Carcinogenesis. Cancer Res. 62, 3675–3684.

Hahn, W.C., Weinberg, R.A., 2002. Rules for Making Human Tumor Cells. N. Engl. J. Med. 347, 1593–1603.

Hoglund, M., Gisselsson, D., Hansen, G.B., Sall, T., Mitelman, F., 2002a. Multivariate analysis of chromosomal imbalances in breast cancer delineates cytogenetic pathways and reveals complex relationships among imbalances. Cancer Res. 62, 2675–2680.

Hoglund, M., Gisselsson, D., Hansen, G.B., Sall, T., Mitelman, F., Nilbert, M., 2002b. Dissecting karyotypic patterns in colorectal tumors: two distinct but overlapping pathways in the adenoma-carcinoma transition. Cancer Res. 62, 5939–5946.

Jin, C., Jin, Y., Wennerberg, J., Akervall, J., Dictor, M., Mertens, F., 2002. Karyotypic heterogeneity and clonal evolution in squamous cell carcinomas of the head and neck. Cancer Genet. Cytogenet. 132, 85–96.

Katsura, K., Sugihara, H., Nakai, S., Fujita, S., 1996. Alteration of numerical chromosomal aberrations during progression of colorectal tumors revealed by a combined fluorescence in situ hybridization and DNA ploidy analysis of intratumoral heterogeneity. Cancer Genet. Cytogenet. 90, 146–153.

Kendal, W.S., 1990. The use of information theory to analyze genomic changes in neoplasia. Math. Biosci. 100, 143–159.

Klein, C.A., Blankenstein, T.J.F., Schmidt-Kittler, O., Petronio, M., Polzer, B., Stoecklein, N.H., Riethmuller, G., 2002. Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. Lancet 360, 683–689.

Li, R., Sonik, A., Stindl, R., Rasnick, D., Duesberg, P., 2000. Aneuploidy vs. gene mutation hypothesis of cancer: recent study claims mutation but is found to support aneuploidy. Proc. Natl Acad. Sci. USA 97, 3236–3241.

Loeb, L.A., Loeb, K.R., Anderson, J.P., 2003. Multiple mutations and cancer. Proc. Natl. Acad. Sci. USA 100, 776–781.

Magurran, A., 1988. Ecological Diversity and Its Measurement. Princeton University Press, Princeton.

Manley, B.F.J., 1998. Randomization, Bootstrap and Monte Carlo Methods in Biology, second ed. Chapman and Hall, London.

Marx, J., 2002. Debate surges over the origins of genomic defects in cancer. Science 297, 544–546.

Matzke, M.A., Florian, M.M., Kanno, T., Matzke, A.J.M., 2003. Does the intrinsic instability of aneuploid genomes have a causal role in cancer? Trends Genet. 19, 253–256.

Mitelman Database of Chromosome Aberrations in Cancer, 2004. In: Mitelman, F., Johansson, B., Mertens, F., (Eds.), http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Mitelman, F., 1995. International System for Human Cytogenetic Nomenclature (1995), ISCN (1995), Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature. S Karger, Basel.

Mitelman, F., 2000. Recurrent chromosome aberrations in cancer. Mutat. Res. 462, 247–253.

Mitelman, F., Johansson, B., Mandahl, N., Mertens, F., 1997. Clinical significance of cytogenetic findings in solid tumors. Cancer Genet. Cytogenet. 95, 1–8.

Nowak, M.A., Komarova, N.L., Sengupta, A., Jallepalli, P.V., Shih, I.M., Vogelstein, B., Lengauer, C., 2002. The role of chromosomal instability in tumor initiation. Proc. Natl Acad. Sci. USA 99, 16226–16231.

Osterheld, M.C., Laurini, R., Saraga, E., Bosman, F.T., 1999. Evaluation of heterogeneity of DNA ploidy in early gastric cancers. Anal. Cell. Pathol. 19, 67–72.

Percy, C., Van Holten, V., Muit, C., 1990. International Classification of Diseases for Oncology, second ed. World Health Organization, Geneva.

Rajagopalan, H., Nowak, M.A., Vogelstein, B., Lengauer, C., 2003. The significance of unstable chromosomes in colorectal cancer. Nature Rev. Cancer 3, 695–701.

Ribas, M., Masramon, L., Aiza, G., Capella, G., Miro, R., Peinado, M.A., 2003. The structural nature of chromosomal instability in colon cancer cells. FASEB J. 17, 289–291.

Ricotta, C., 2003. On parametric evenness measures. J. Theor. Biol. 222, 189–197.

Ried, T., Heselmeyer-Haddad, K., Blegen, H., Schrock, E., Auer, G., 1999. Genomic changes defining the genesis, progression, and malignancy potential in solid human tumors: A phenotype genotype correlation. Genes Chromosomes Cancer 25, 195–204.

Surveillance, Epidemiology, and End Results (SEER) Program, 2004. SEER*Stat Database: Incidence—SEER 11 Regs + AK Public-Use for the CSR, Nov 2001 Sub for Exp Races (1992–1999), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, http://www.seer.cancer.gov.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Systems Tech. J. 27, 379–423 623–656.

Sieber, O.M., Heinimann, K., Tomlinson, I.P., 2003. Genomic instability-the engine of tumorigenesis? Nature Rev. Cancer 3, 701–708.

Stoler, D.L., Chen, N., Basik, M., Kahlenberg, M.S., Rodriguez-Bigas, M.A., Petrelli, N.J., Anderson, G.R., 1999. The onset and extent of genomic instability in sporadic colorectal tumor progression. Proc. Natl Acad. Sci. USA 96, 15121–15126.

Storchova, Z., Pellman, D., 2004. From polyploidy to aneuploidy, genome instability and cancer. Nat. Rev. Mol. Cell Biol. 5, 45–54.

Surveillance Research Program, 2004. National Cancer Institute SEER*Stat software. http://www.seer.cancer.gov/seerstat, version 5.0.20.

Capítulo 2: *Natureza estocástica das aberrações cromossômicas em neoplasias sólidas*.

Available online at www.sciencedirect.com
ScienceDirect
Mutation Research 600 (2006) 150–164
Fundamental and Molecular Mechanisms of Mutagenesis
www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Chromosome aberrations in solid tumors have a stochastic nature

Mauro A.A. Castro [a,b,c,d,*], Tor G.H. Onsten [b,d], José C.F. Moreira [a], Rita M.C. de Almeida [c]

[a] *Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003, Brazil*
[b] *Departamento de Medicina Interna, Hospital de Clínicas de Porto Alegre, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2350, Porto Alegre 90035-903, Brazil*
[c] *Instituto de Física, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre 91501-970, Brazil*
[d] *Universidade Luterana do Brasil, Rua Miguel Tostes 101, Canoas 92420-280, Brazil*

Received 27 December 2005; received in revised form 24 March 2006; accepted 13 April 2006
Available online 30 June 2006

Abstract

An important question nowadays is whether chromosome aberrations are random events or arise from an internal deterministic mechanism, which leads to the delicate task of quantifying the degree of randomness. For this purpose, we have defined several Shannon information functions to evaluate disorder inside a tumor and between tumors of the same kind. We have considered 79 different kinds of solid tumors with 30 or more karyotypes retrieved from the Mitelman Database of Chromosome Aberrations in Cancer. The Kaplan–Meier cumulative survival was also obtained for each solid tumor type in order to correlate data with tumor malignance. The results here show that aberration spread is specific for each tumor type, with high degree of diversity for those tumor types with worst survival indices. Those tumor types with preferential variants (e.g. high proportion of a given karyotype) have shown better survival statistics, indicating that aberration recurrence is a good prognosis. Indeed, global spread of both numerical and structural abnormalities demonstrates the stochastic nature of chromosome aberrations by setting a signature of randomness associated to the production of disorder. These results also indicate that tumor malignancy correlates not only with karyotypic diversity taken from different tumor types but also taken from single tumors. Therefore, by quantifying aberration spread, we could confront diverse models and verify which of them points to the most likely outcome. Our results suggest that the generating process of chromosome aberrations is neither deterministic nor totally random, but produces variations that are distributed between these two boundaries.

Keywords: Solid tumor; Karyotype; Information theory; Entropy; Cancer; Cytogenetics; Chromosome aberrations; Diversity

1. Introduction

Several cytogenetic studies have shown that cancer karyotypes have chromosome aberrations that arise with tumor evolution and persist and are propagated in such a manner that the majority of solid tumors not only present great alterations in number but also in chromosome structure [1–4]. Such characteristic is possibly

* Corresponding author at: Departamento de Bioquímica, ICBS, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003, Brazil. Tel.: +55 51 3316 5577; fax: +55 51 3316 5540.
E-mail address: mauro@ufrgs.br (M.A.A. Castro).

0027-5107/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.mrfmmm.2006.04.003

associated to errors in chromosomal segregation [5], whose origin seems to be the instability at chromosome level (chromosome instability—CIN) [6]. However, the mechanism that leads to CIN and to aberrant karyotypes has not been elucidated and remains a point of controversy and discussion [7–9]. Nowadays, one of the major problems to estimate the CIN is the difficulty to obtain a well defined quantity to estimate the rate at which large chromosome portions or even whole chromosomes are gained or lost in cancers [10]. In fact this difficulty arises from the lack of a proper understanding about the exact nature of this instability. Here we provide some clues to deal with this question by using a different approach to identify stochastic boundaries. Indeed, our description of tumor aberrations makes possible to estimate the degree of malignance observed in solid tumors and also to evaluate the contributions coming from numerical and structural chromosome abnormalities for global disorder.

Recently, we have described the complexity of solid tumors by using Shannon's entropy to express karyotypic diversity and uncertainty associated to sample distribution [11]. We have shown that the diversity of chromosome aberrations (of a set of 14 epithelial tumor types) presents a significant correlation with cancer survival observed in population statistics. Now, we expand the conclusions previously published using a larger number of solid tumors of different morphological types and also define complementary Shannon information functions to evaluate the disorder inside a tumor and between tumors of the same kind. The results here show that the sample spread of chromosome aberrations is specific for each tumor type and is correlated with tumor malignance, with lower aberration diversity for those tumor types with better survival statistics. Indeed, the normalized value of global spread of both numerical and structural chromosome abnormalities grows asymptotically with sample number, demonstrating the stochastic nature that underlies chromosome aberrations.

2. Methods

2.1. Karyotype data selection

Solid tumor types with 30 or more karyotypes were retrieved from the Mitelman Database of Chromosome Aberrations in Cancer, unselected series [12]. Therefore, 79 solid tumor types were selected (Table 1). Cases with incomplete karyotypes were excluded from the analysis (e.g. questionable identification of a chromosome—with question mark (?) code). The karyotypes were then assessed for numerical and structural chromosome abnormalities and pooled accordingly to the sum of alterations, counted in agreement with the ISCN Abbreviated Terms and Symbol [13]. Data tables were organized as exemplified in Fig. 2. To obtain a quantitative expression of sample distribution, we have measured the information content of pooled karyotypes using Shannon Information Theory [11,14,15], as characterized in Section 3.

2.2. Survival statistics

Histopathological and topographical nomenclatures of solid tumors were also reviewed and organized, following the International Classification of Diseases for Oncology [16] in order to retrieve epidemiological parameters from the Surveillance, Epidemiology, and End Results Program [17]. Therefore, the cumulative survival was obtained for each solid tumor type through the SEER*Stat software [18] at SEER Statistic Review Database 1973–2001 [17]. The Kaplan–Meier survival curves were then constructed for malignant tumors using the relative survival statistics and, for benignant tumors (which are not available in SEER cancer registries), using the expected survival statistics of a comparable set of cancer free individuals. The cohorts were generated through the merge option of the SEER*Stat software, observing the ICD-O-3 codes described in Table 1 for the selected tumors.

SEER*Stat's relative survival is a net survival measure representing cancer survival in absence of other causes of death, while expected survival is obtained from the expected rate table by matching the cohort cases by race, sex, age, and date at which the age was coded. Observed and expected rates are used to generate relative survival rates [18,19].

3. Mathematical definitions and statistical analysis of karyotypic diversity

3.1. Information content of sample distribution

To obtain a quantitative expression of sample distribution, we have measured the information content of pooled karyotypes using Shannon information [11,14,15,20]. The information content is calculated from probability functions and describes the degree of sample spread. Analogously, entropy measures the degree of order. One is related to the other, since information content increases as entropy decreases [20]. Here, information content also indicates the degree of population diversity. For example, consider the pooled karyotypes presented in Fig. 1. If a karyotype is randomly selected, the probability distribution describes the probability p_i that the observed karyotype will be of a particular type i (e.g. with a particular chromosome number). It is possible to calculate the Shannon information for every probability distribution as

$$H = -\sum_i p_i \ln p_i. \quad (1)$$

Table 1
List of solid tumors

Label	Mitelman Database term[a]		Tissue type	ICD-O-3 code[b]		*n*
	Morphology	Topography		Morphology	Topography	
1	Adenocarcinoma	Breast	Epithelial	8140/3	C50.0–50.9	641
2	Adenocarcinoma	Kidney	Epithelial	8140/3	C64.9	764
3	Adenocarcinoma	Large intestine	Epithelial	8140/3	C18.0–18.9	329
4	Adenocarcinoma	Lung	Epithelial	8140/3	C34.0–34.9	160
5	Adenocarcinoma	Ovary	Epithelial	8140/3	C56.9	339
6	Adenocarcinoma	Pancreas	Epithelial	8140/3	C25.0–25.9	91
7	Adenocarcinoma	Prostate	Epithelial	8140/3	C61.9	189
8	Adenocarcinoma	Stomach	Epithelial	8140/3	C16.0–16.9	114
9	Adenocarcinoma	Thyroid	Epithelial	8140/3	C73.9	126
10	Adenocarcinoma	Uterus, corpus	Epithelial	8140/3	C54.0–54.9	106
11	Adenoma	Breast	Epithelial	8140/0[c]	C50.0–50.9	53
12	Adenoma	Large intestine	Epithelial	8140/0[c]	C18.0–18.9	134
13	Adenoma	Pituitary	Epithelial	8140/0[c]	C75.1	48
14	Adenoma	Salivary gland	Epithelial	8140/0[c]	C07.9–08.9	229
15	Adenoma	Thyroid	Epithelial	8140/0[c]	C73.9	96
16	Alveolar rhabdomyosarcoma	Soft tissue	Mesenchymal	8920/3	C49.0–49.9	54
17	Astrocytoma, grade I–II	Brain	CNS	9400/31 or 9400/32	C71.0–71.5; 71.7–71.9	55
18	Astrocytoma, grade III–IV	Brain	CNS	9400/33 or 9400/34	C71.0–71.5; 71.7–71.9	387
19	Atypical lipomatous tumor/atypical lipoma/well-different liposarcoma	Soft tissue	Mesenchymal	8850/1 or 8851/3	C49.0–49.9	98
20	Basal cell carcinoma	Skin	Epithelial	8090/3	C44.0–44.9; 51.0, 51.9, 60.9, 63.2	98
21	Benign epithelial tumor, NOS	Breast	Epithelial	8010/0[c]	C50.0–50.9	35
22	Carcinoma, NOS	Breast	Epithelial	8010/3	C50.0–50.9	38
23	Carcinoma, NOS	Large intestine	Epithelial	8010/3	C18.0–18.9	47
24	Chondroid hamartoma	Lung	Mesenchymal	–[c]	C34.0–34.9	143
25	Chondrosarcoma, NOS	Skeleton	Mesenchymal	9220/3	C40.0–41.9	93
26	Chordoma	Skeleton	Mesenchymal	9370/3	C40.0–41.9	31
27	Combined germ cell tumors	Testis	Mesenchymal	9064/3	C62.0–62.9	61
28	Desmoid-type fibromatosis	Soft tissue	Mesenchymal	8821/1[c]	C49.0–49.9	39
29	Embryonal rhabdomyosarcoma	Soft tissue	Mesenchymal	8910/3	C49.0–49.9	43
30	Ependymoma	Brain	CNS	9391/3	C71.0–71.5; 71.7–71.9	67
31	Ewing tumor/peripheral primitive neuroectodermal tumor	Skeleton	Mesenchymal	9260/3 or 9364/3	C40.0–41.9	191
32	Ewing tumor/peripheral primitive neuroectodermal tumor	Soft tissue	Mesenchymal	9260/3 or 9364/3	C49.0–49.9	30
33	Gastrointestinal stromal tumor	Stomach	Mesenchymal	8936/0[c] or 8936/1[c]	C16.0–16.9	32
34	Giant cell tumor of bone	Skeleton	Mesenchymal	9250/3	C40.0–41.9	40
35	Glioma, NOS	Brain	CNS	9380/3	C71.0–71.5; 71.7–71.9	39
36	Hepatoblastoma	Liver	Mesenchymal	8970/3	C22.0	42
37	Leiomyoma	Uterus, corpus	Mesenchymal	8890/0[c]	C54.0–54.9	291
38	Leiomyosarcoma	Soft tissue	Mesenchymal	8890/3	C49.0–49.9	56
39	Lipoma	Soft tissue	Mesenchymal	8850/0[c]	C49.0–49.9	171
40	Liposarcoma, myxoid/round cell	Soft tissue	Mesenchymal	8852/3 or 8853/3	C49.0–49.9	76
41	Malignant fibrous histiocytoma	Soft tissue	Mesenchymal	8830/3	C49.0–49.9	54
42	Malignant melanoma	Eye	Epithelial	8720/3	C69.0–69.9	145
43	Malignant melanoma	Skin	Epithelial	8720/3	C44.0–44.9; 51.0, 51.9, 60.9, 63.2	94
44	Malignant melanoma	Unknown site	Epithelial	8720/3	C80.9	59
45	Malignant peripheral nerve sheath tumor/Triton	Soft tissue	CNS	9561/3	C49.0–49.9	61
46	Meningioma	Brain	CNS	9530/0[c]	C71.0–71.5; 71.7–71.9	666
47	Meningioma	Cerebellum	CNS	9530/0[c]	C71.6	32
48	Meningioma	Spinal cord	CNS	9530/0[c]	C72.0	45
49	Mesothelioma	Pleura	Mesenchymal	9050/3	C38.4	96
50	Myxofibrosarcoma	Soft tissue	Mesenchymal	8811/3	C49.0–49.9	30

Table 1 (*Continued*)

Label	Mitelman Database term[a]		Tissue type	ICD-O-3 code[b]		*n*
	Morphology	Topography		Morphology	Topography	
51	Neuroblastoma	Adrenal	CNS	9500/3	C74.0–74.9	123
52	Neuroblastoma	Soft tissue	CNS	9500/3	C49.0–49.9	41
53	Neuroglial neoplasm, special type	Brain	CNS	9505/1[c]	C71.0–71.5; 71.7–71.9	36
54	Nonneoplastic epithelial disorder/lesion	Breast	Epithelial	–[c]	C50.0–50.9	35
55	Nonneoplastic epithelial disorder/lesion	Thyroid	Epithelial	–[c]	C73.9	44
56	Nonneoplastic mesenchymal disorder/lesion	Skeleton	Mesenchymal	–[c]	C40.0–41.9	36
57	Nonneoplastic mesenchymal disorder/lesion	Soft tissue	Mesenchymal	–[c]	C49.0–49.9	49
58	Oligodendroglioma	Brain	CNS	9450/3	C71.0–71.5; 71.7–71.9	45
59	Oncocytoma	Kidney	Epithelial	8290/0[c]	C64.9	46
60	Osteosarcoma, NOS	Skeleton	Mesenchymal	9180/3	C40.0–41.9	140
61	Primitive neuroectodermal tumor/Medulloblastoma	Brain	CNS	9470/3 or 9473/3	C71.0–71.5; 71.7–71.9	56
62	Primitive neuroectodermal tumor/medulloblastoma	Cerebellum	CNS	9470/3 or 9473/3	C71.6	112
63	Retinoblastoma	Eye	CNS	9510/3	C69.0–69.9	124
64	Seminoma/dysgerminoma	Testis	Mesenchymal	9060/3 or 9061/3	C62.0–62.9	74
65	Squamous cell carcinoma	Larynx	Epithelial	8070/3	C32.0–32.9	94
66	Squamous cell carcinoma	Lung	Epithelial	8070/3	C34.0–34.9	146
67	Squamous cell carcinoma	Oral cavity	Epithelial	8070/3	C04.0–06.9	60
68	Squamous cell carcinoma	Oro- and hypopharynx	Epithelial	8070/3	C10.0–10.9 or C13.0–13.9	35
69	Squamous cell carcinoma	Tongue	Epithelial	8070/3	C01.9–02.9	53
70	Squamous cell carcinoma	Uterus, cervix	Epithelial	8070/3	C53.0–53.9	38
71	Squamous cell carcinoma	Vagina	Epithelial	8070/3	C52.9	39
72	Superficial fibromatosis	Soft tissue	Mesenchymal	8813/0[c]	C49.0–49.9	44
73	Synovial sarcoma	Soft tissue	Mesenchymal	9040/3	C49.0–49.9	124
74	Teratoma (mature and immature)	Ovary	Mesenchymal	9080/0 or 9080/3	C56.9	36
75	Teratoma (mature and immature)	Testis	Mesenchymal	9080/0 or 9080/3	C62.0–62.9	158
76	Transitional cell carcinoma	Bladder	Epithelial	8120/3	C67.0–67.9	159
77	Undifferentiated carcinoma, large cell	Lung	Epithelial	8012/33	C34.0–34.9	38
78	Undifferentiated carcinoma, small cell	Lung	Epithelial	8041/33	C34.0–34.9	45
79	Wilms tumor	Kidney	Mesenchymal	8960/3	C64.9	389

[a] Solid tumors with (*n*) cases ≥ 30 listed in Mitelman Database of Chromosome Aberrations in Cancer [12].

[b] Tumor morphology/topography ICD-O-3 codes [16]: corresponding codes of tumor terms catalogued in Mitelman Database used to retrieve epidemiological data (i.e. relative survival) from Surveillance, Epidemiology, and End Results Program [18].

[c] Benign tumors, uncertain whether benign or malignant tumors and nonneoplastic disorders (/0;/1; and –) are not available in SEER cancer registries [17]. For these 21 tumor types, are used the ICD-O-3 code of the corresponding tissue type described in solid tumor list: epithelial tumors (8010–8720), mesenchymal tumors (8811–9370) and Central Nervous System tumors—CNS (9380–9561) in order to retrieve the expected survival of a comparable set of cancer free individuals.

The quantity H indicates how diverse a distribution is, giving an estimate of karyotypic diversity. Thus an increase in heterogeneity of some chromosome abnormality implies higher disorder, higher H and loss of information.

In Fig. 1A, the histograms of numerical chromosome abnormalities show a more heterogeneous distribution when comparing to structural chromosome abnormalities presented in Fig. 1B. Information theory analysis yields that numerical chromosome abnormalities carry less information or higher entropy, implying higher diversity. It means that knowing that a karyotype comes from distribution represented in Fig. 1A we are less informed about its type than in the case of a karyotype coming from distribution of Fig. 1B.

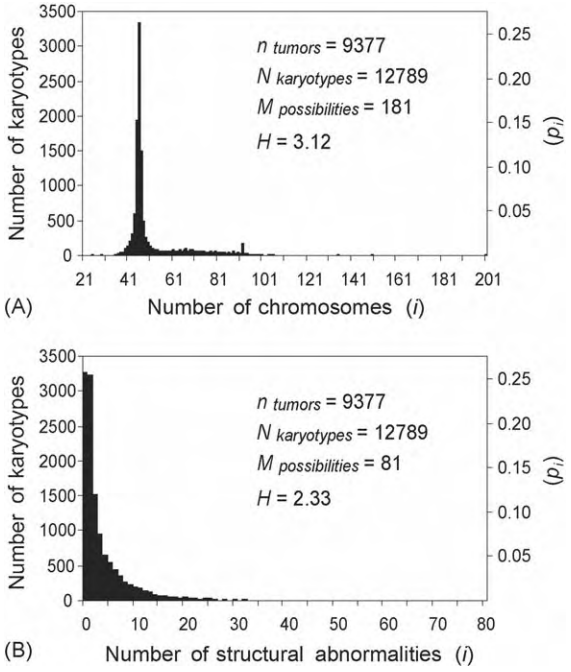


Fig. 1. Information content of sample distribution. Solid tumor karyotypes from Mitelman Database are pooled accordingly to the sum of chromosome abnormalities. Data tables were organized as exemplified in Fig. 2. (A) Karyotypes with i chromosomes are pooled to construct the histogram distribution, which represents the total number of karyotypes (left axis) or the proportion of each type i in the probability density function (right axis). (B) Karyotypes with i structural abnormalities are pooled, as in (A). The sum over $-p_i \ln(p_i)$ measure the degree of sample spread, which is expressed by H value; M is number of possibilities for each karyotype; N is the number of karyotypes counted in the sample of n tumors (considering that tumors are characterized by clones, N also represents the total number of clones in the sample). Indeed, here H corresponds to the Shannon information function described in Eq. (11e), but to reach the same result, it must be divided by the factor $\ln(Mn)$ in order to normalize the quantities. For further details, please see Section 3.2.

3.2. Characterization of information functions to assess karyotypic diversity

To define different Shannon information functions, consider a sample of karyotypes of n tumors of a given type. Each tumor α of the sample ($\alpha = 1, \ldots, n$) has K_α different clones, among M possible clones. The clones are characterized by the kind of alterations they present. Here we shall separately consider only the number of numerical and structural chromosome aberrations. Take, for the moment, the label i as indicating the number of chromosomes, that could vary from i_0 to i_f, for example. In this case the number of possible clones is $M = i_f - i_0 + 1$. For a given tumor α, we can define $s(i, \alpha)$ as being the number of clones with i chromosomes. K_α is given by

$$K_\alpha = \sum_{i=i_0}^{i=i_f} s(i, \alpha). \tag{2}$$

We can also define other quantities from the values of $s(i, \alpha)$ for the entire sample. We begin by the number L_i of karyotypes of type i throughout the entire sample of n tumors, that is

$$L_i = \sum_{\alpha=1}^{n} s(i, \alpha). \tag{3}$$

We now can obtain from either K_α or L_i the total number N of karyotypes:

$$N = \sum_{\alpha=1}^{n} K_\alpha = \sum_{i=i_0}^{i=i_f} L_i. \tag{4}$$

From Eqs. (2)–(4) we may define different, normalized probability functions. We begin by the most general one, $z(i, \alpha)$, that is the probability that among all karyotypes we randomly choose one that is of the i type and belongs to tumor α:

$$z(i, \alpha) = \frac{s(i, \alpha)}{N}. \tag{5}$$

We can also define for each tumor α the probability $p(i, \alpha)$ that a randomly chosen karyotype is of the ith type:

$$p(i, \alpha) = \frac{s(i, \alpha)}{K_\alpha}. \tag{6}$$

and, for each type i of karyotype we may define the probability $\eta(i, \alpha)$ that a randomly chosen karyotype belongs to the αth tumor,

$$\eta(i, \alpha) = \frac{s(i, \alpha)}{L_i}. \tag{7}$$

Finally we can also define P_α as the probability that, randomly choosing among the N karyotypes, the chosen one belongs to the αth tumor. It expresses the relative number of karyotypes in each tumor. It may be written as

$$P_\alpha = \frac{K_\alpha}{N}, \tag{8}$$

and analogously, we may define the normalized probability Q_i that a randomly chosen karyotype is of the ith type. It expresses the relative number of i-karyotypes:

$$Q_i = \frac{L_i}{N}. \tag{9}$$

All five probability functions defined above are normalized. The normalization conditions are written as

$$\sum_i \sum_\alpha z(i,\alpha) = 1. \tag{10a}$$

$$\sum_i p(i,\alpha) = 1. \tag{10b}$$

$$\sum_\alpha \eta(i,\alpha) = 1. \tag{10c}$$

$$\sum_\alpha P_\alpha = 1. \tag{10d}$$

$$\sum_i Q_i = 1. \tag{10e}$$

We associate a normalized Shannon information function to each one of the above probability functions, as follows:

$$H_{\text{global}} = -\frac{1}{\ln(Mn)} \sum_i \sum_\alpha z(i,\alpha) \ln z(i,\alpha). \tag{11a}$$

$$H_\alpha = -\frac{1}{\ln(Mn)} \sum_i p(i,\alpha) \ln p(i,\alpha). \tag{11b}$$

$$H_i = -\frac{1}{\ln(Mn)} \sum_\alpha \eta(i,\alpha) \ln \eta(i,\alpha). \tag{11c}$$

$$H^* = -\frac{1}{\ln(Mn)} \sum_\alpha P_\alpha \ln P_\alpha. \tag{11d}$$

$$H' = -\frac{1}{\ln(Mn)} \sum_i Q_i \ln Q_i. \tag{11e}$$

where we have divided all terms by the factor $\ln(Mn)$ in order to normalize the quantities. The idea is to compare among samples with different numbers of n tumors and between different karyotypic characterizations (numerical or structural abnormalities) with different numbers of M possibilities, such that H will always vary between 0 and 1. H_{global} reflects the spread of the global distribution $z(i,\alpha)$, while H_α is the intra tumoral spread of the αth tumor, that is, it measures the diversity that exists in the αth tumor. Analogously, H_i is the spread associated to the i-karyotype, or how many different tumors present this karyotype. It is zero if only one tumor present that karyotype and has its largest value in the case that all tumors present the i karyotype in equal amounts. H^* gives a measure of the spread due to the differences in number of karyotypes among tumors (due to different clones), that is, it has its largest value when all tumors present the same number of karyotypes. Finally, H' is related to the spread among karyotypes, meaning that it has its largest value when there is the same number of each karyotype in the whole sample. The relation between these five information functions is given by the following relations:

$$H_{\text{global}} = H^* + \sum_\alpha P_\alpha H_\alpha = H^* + H_{\text{tumor}}. \tag{12a}$$

$$H_{\text{global}} = H' + \sum_i Q_i H_i = H' + H_{\text{karyo}} \tag{12b}$$

The information function H_{global}, related to the global probability function $z(i,\alpha)$, carries information about the global spread of the whole sample of N karyotypes distributed among n tumors. However, this spread may be concentrated in the differences among tumors or in the typical spread of the tumors individually. Eq. (12a) expresses the fact that the spread H_{global} may be calculated as the sum of the average intratumoral spread $H_{\text{tumor}} \equiv \sum_\alpha P_\alpha H_\alpha$ and H^*, which gives the spread of number of karyotypes per tumor, a possible measure of the difference between tumors. On the other hand, Eq. (12b) gives the spread H_{global} as the sum of the average intertumoral or intrakaryotype spread $H_{\text{karyo}} \equiv \sum_i Q_i H_i$ and H', the spread of number of each karyotype in the sample, a possible measure of the differences between karyotypes. In fact, depending on whether the relevant contribution to H_{global} comes from differences between tumors, or differences in the probability of different karyotypes may, for example, bring some light into the role played by error randomness or, when a given karyotype is significantly more frequent, indicate a relevant correlation with that kind of tumor.

3.3. Characterization of sample spread types

When every tumor has only one karyotype, say karyotype k, there is a maximum value of H_{global} for n tumors, $H_{\max} = \ln(n)/\ln(Mn)$, corresponding to two different situations. We consider first the situation when all tumors are equal for a given chromosome aberration number (numerical or structural). For each tumor α, $H_\alpha = 0$, yielding $H_{\text{tumor}} = \sum_\alpha P_\alpha H_\alpha = 0$, meaning that there is only one karyotype in that tumor. For karyotype k, present in all n tumors, there is a non-zero value of information function, specifically $H_{i=k} = \ln(n)/\ln(Mn)$, while for the other karyotypes $H_i = 0$, yielding $H_{\text{karyo}} = \sum_i Q_i H_i = \ln(n)/\ln(Mn)$. On the other hand, all tumors being equal, $H^* = \ln(n)/\ln(Mn)$ while $H' = 0$, as there is only one karyotype. Consequently $H_{\text{global}} = H_{\max} = \ln(n)/\ln(Mn)$. We named this spreading as *type A* (exemplified in Fig. 2A for numerical chromosome aberrations).

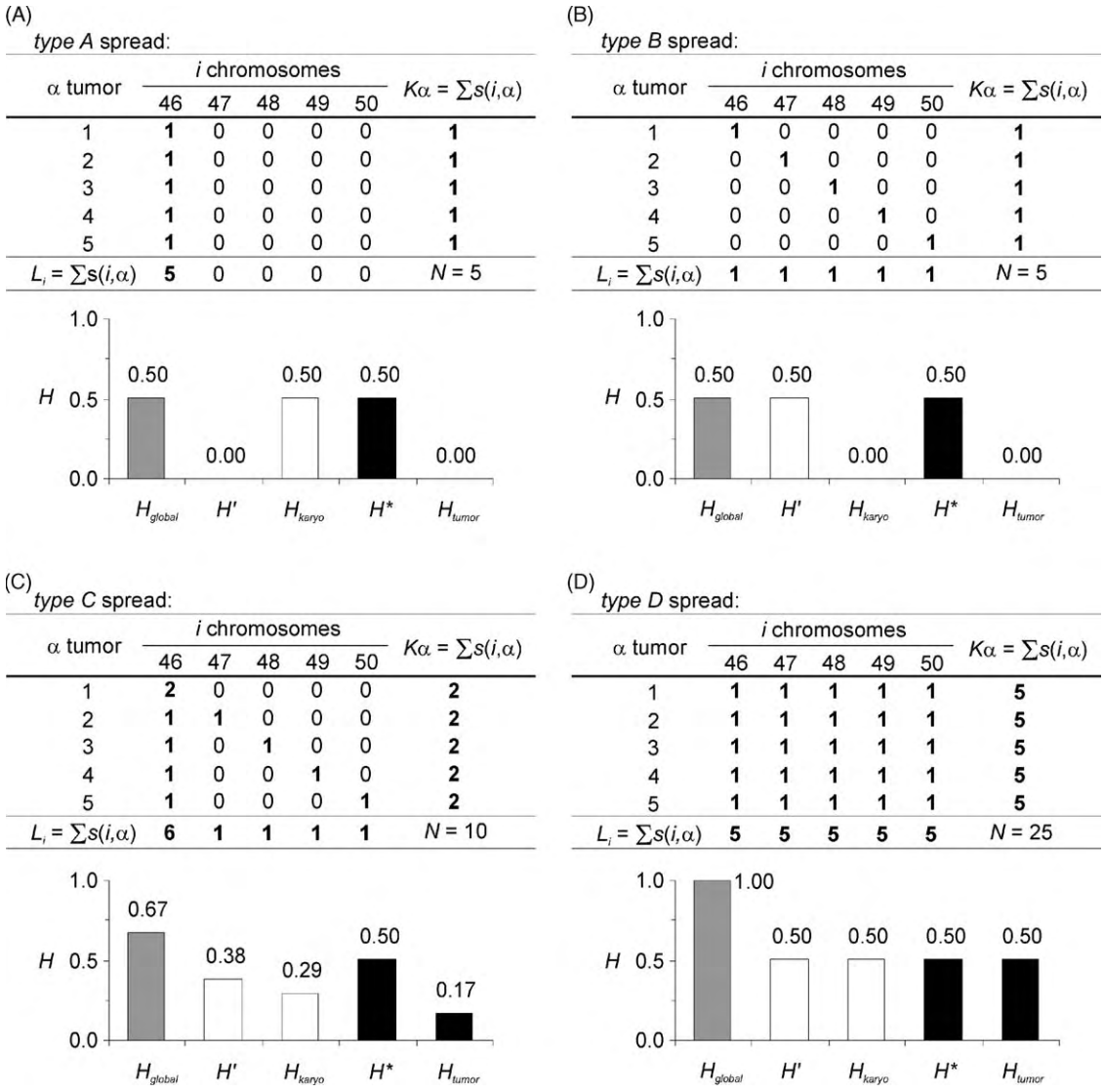


Fig. 2. Illustration of four possible situations of sample spread (e.g. numerical chromosome aberrations). The tables present five tumors ($n = 5.0$) with five possible karyotypes ($M = 5.0$). For a given tumor α, $s(i, \alpha)$ is the number of clones with i chromosomes; K_α is the number of different clones for each tumor α and L_i is the number of karyotypes of type i. Below each sample spread table are presented the values of global spread (H_{global}) with the contributions H', H_{karyo}, H^* and H_{tumor}. (A) *type A* spread: when all tumors are equal for a given chromosome aberration number with equal karyotypes; (B) *type B* spread: when all tumors present different karyotypes; (C) *type C* spread: when tumors can have more than one karyotype (e.g. presenting more than one clone) combined as *type A*/*B* spread; (D) *type D* spread: when tumors have the largest possible karyotype spread in the sample.

The second situation we consider is the case when all tumors are different (*type B* spread, exemplified in Fig. 2B). Assume $k(\alpha)$ as the karyotype belonging to tumor α. Here H_α and H_{tumor} are zero, since as before there is only one karyotype in each tumor. H_i and H_{karyo} are also zero, because each karyotype is present in only one tumor. On the other hand, $H' = \ln(n)/\ln(Mn)$, because there are n different karyotypes. Also, $H^* = \ln(n)/\ln(Mn)$, yielding $H_{\text{global}} = H_{\max} = \ln(n)/\ln(Mn)$.

Comparing the two situations considered above, we see that in spite the fact of H_{global} having the same value for both cases, the partial information functions H' and H_{karyo} present different values. In the first situation (*type A* spread) there is a lack of information because, once given the karyotype, one cannot say from which tumor it comes. In the second case (*type B* spread), the lack of information arises because there are n different karyotypes. In *type A* spread, there is a strong correlation

between the disease and the *k* karyotype, while in *type B* spread, the correlation between karyotypes and disease is zero.

We illustrate another two possible situations when tumors can have more than one karyotype (e.g. presenting more than one clone), with $H_{max} > \ln(n)/\ln(Mn)$. In Fig. 2C, we present the *type C* spread, which combines *type A* and *type B*. As expected, this spread has values of Shannon information functions that follow *type A* and *type B* profiles, but with higher values of information functions, especially for global spread. Finely, in Fig. 2D, we present the *type D* spread, when tumors have the largest possible spread, with $H_{global} = 1.0$.

In real cases we expect some correlation between karyotypes and malignancy. Also there is dispersion in each tumor type. When this dispersion represents large differences between tumor types, a detailed investigation on the frequency Q_i of karyotypes in different samples could, perhaps, indicate which of them correlate more strongly with disease.

4. Results

4.1. Applying Shannon information functions to evaluate sample spread of solid tumor karyotypes

We now present the results of measuring the above-defined Shannon information functions in solid tumor karyotypes, which were retrieved from the Mitelman Database of Chromosome Aberrations in Cancer [12]. We have considered both numerical and structural abnormalities. We have considered 79 different kinds of tumors with different numbers in each sample (Table 1). For structural chromosome aberrations, we considered $i = 0, \ldots, 80$, such that $M = 81$. For numerical chromosome aberrations, $i = 21, \ldots, 201$, such that $M = 181$.

In Fig. 3A, we present the results for numerical chromosome aberrations. In both plots we present the results for H_{global}, together with the partial contributions. We first observe that H_{global} presents a fairly constant value, oscillating around 0.5. In the upper graph (a) we plotted separately the contributions coming from H' and H_{karyo}. We can see that both contributions are relevant to H_{global}, and the relative contribution may vary from one kind of tumor to another. The lower plot in Fig. 3A (b), on the other hand, indicates that the main contribution to H_{global} comes from H^*, meaning that each tumor in a given sample presents a similar number of different karyotypes, while in each tumor there is a small number of karyotypes in comparison to the number of possibilities. Together, both graphs indicate that (1) each sample present a small number of karyotypes (in comparison to the possibilities) (2) this small number is similar to all tumors in a sample, (3) the karyotypes presented may be dispersed or not, depending on the sample. Therefore, the majority of tumor samples varies through *type A* and *type B* spread—together with its combinations. In these spread types, the dispersion appears for H' and H_{karyo}, indicating that we should look for tumor-specific differences in these two Shannon information functions. The same results are presented in Fig. 3B, but now for structural abnormalities. The results vary quantitatively, but qualitatively they are the same.

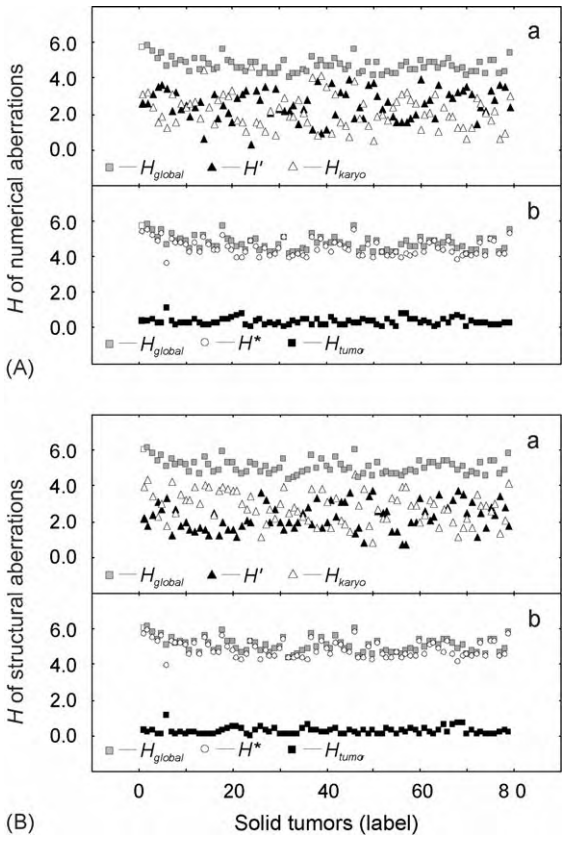


Fig. 3. Measure of sample spread of chromosome aberrations. The results of Shannon information functions are presented for numerical (A) and structural aberrations (B). The solid tumors follow the order presented in Table 1 (label). The global spread (H_{global}) is presented together with partial contributions H' and H_{karyo} (a) or H^* and H_{tumor} (b).

To compare the data obtained from numerical and structural chromosome aberrations, we plotted the results from one versus the other, in Fig. 4A. The results for H_{global}, H^* and H_{tumor} correlate almost perfectly and, again, the dispersion appears in H' and H_{karyo}. Indeed, Fig. 4B shows that the relative contribution of H' and H_{karyo} vary through the global sample, contrasting with

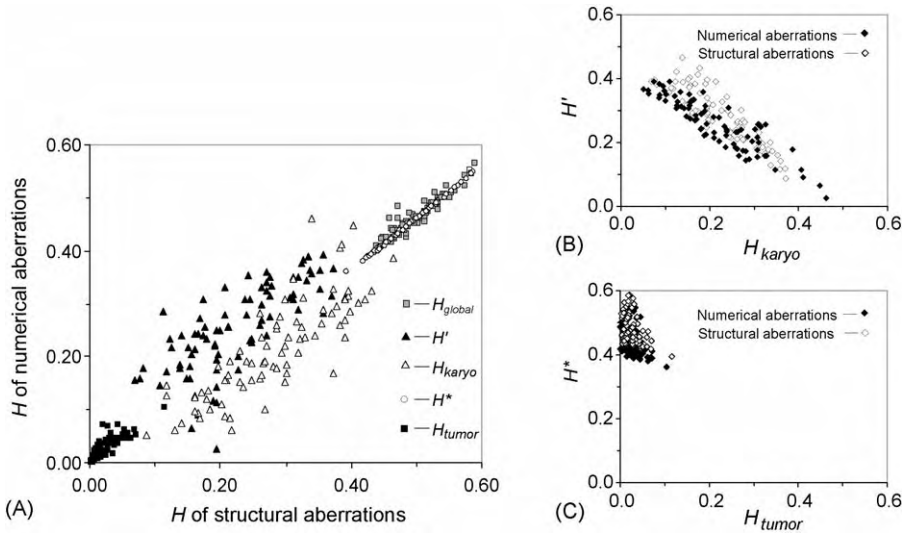


Fig. 4. Comparing chromosome aberrations. (A) Scatter plot of spread data of numerical and structural aberrations. The results for H_{global}, H^* and H_{tumor} correlate almost perfectly and the dispersion appears in H' and H_{karyo}. (B) Scatter plot of partial contributions H' and H_{karyo}. (C) Scatter plot of partial contributions H^* and H_{tumor}.

H^* and H_{tumor} (Fig. 4C), which emphasizes the contribution from H^*.

Furthermore, considering current cytogenetic recommendations [13], it is also important to realize that the contribution from H_{tumor} (intratumoral spread) could be higher, since the number of clones in some karyotypes may be underestimate. For example, the composite karyotypes (which can be created to describe great karyotypic heterogeneity within tumors) combine all chromosome abnormalities of a sample into one representative karyotype [13].

Accordingly, the results presented above indicate that, for some tumor types, the production of aberrations is not specific of a given tumor, in the sense that it is not the same aberration that is produced in different tumors of the same kind (e.g. *type B* spread). For other tumor types, a given karyotype is significantly more frequent (e.g. *type A* spread), indicating a relevant correlation with that kind of tumor. Therefore, the aberration spread, say spread profile, may be specific for each type of tumor, in the sense that it is not the same spread that is produced among tumor samples. It is then interesting to verify how aberration spread correlates with malignancy.

4.2. Association between sample spread and survival statistics

In order to assess tumor malignance, we retrieved survival data from population-based registries in Surveillance, Epidemiology and End Results SEER Program database for the 79 solid tumor types previously selected in Mitelman Database (Table 1). Therefore, the Kaplan–Meier distributions of cumulative survival estimates of 60 months cohorts are obtained for each corresponding solid tumor type, which are presented in Fig. 5A.

As time goes by, the cohort composition varies, due to the different survival rates of different cohorts. In order to measure the time evolution of average values of information functions, we define a time average $\bar{f}(t)$ of any Shannon information function f as

$$\bar{f}(t) = \frac{\sum_{\beta=1}^{n} f\sigma(\beta, t)}{\sum_{\beta=1}^{n} \sigma(\beta, t)}. \tag{13}$$

where $\sigma(\beta, t)$ is the survival fraction at time t of the cohort β ($\beta = 1, \ldots, 79$).

The idea is to estimate the time evolution of such an average for each Shannon information function, when these averages are taken considering the surviving portion of the initial cohort. Two situations may indicate an association between survival and $\bar{H}(t)$: (1) if the average $\bar{H}(t)$ progresses to its minimum value from $t = 0$ to $t = 60$, it indicates a relative increase of those cohorts with low values of information functions, and (2) if the average $\bar{H}(t)$ progresses to its maximum value from $t = 0$ to $t = 60$, it indicates a relative increase of those cohorts with high values of information functions. (The situation that the average $\bar{H}(t)$ stays constant from $t = 0$ to $t = 60$ indicates no correlation.)

In Fig. 5B we present the results of $\bar{H}(t)$ functions for H' and H_{karyo} of numerical chromosome aberrations. Thus, we can see that the contributions $\bar{H}'$ and $\bar{H}_{\text{karyo}}$

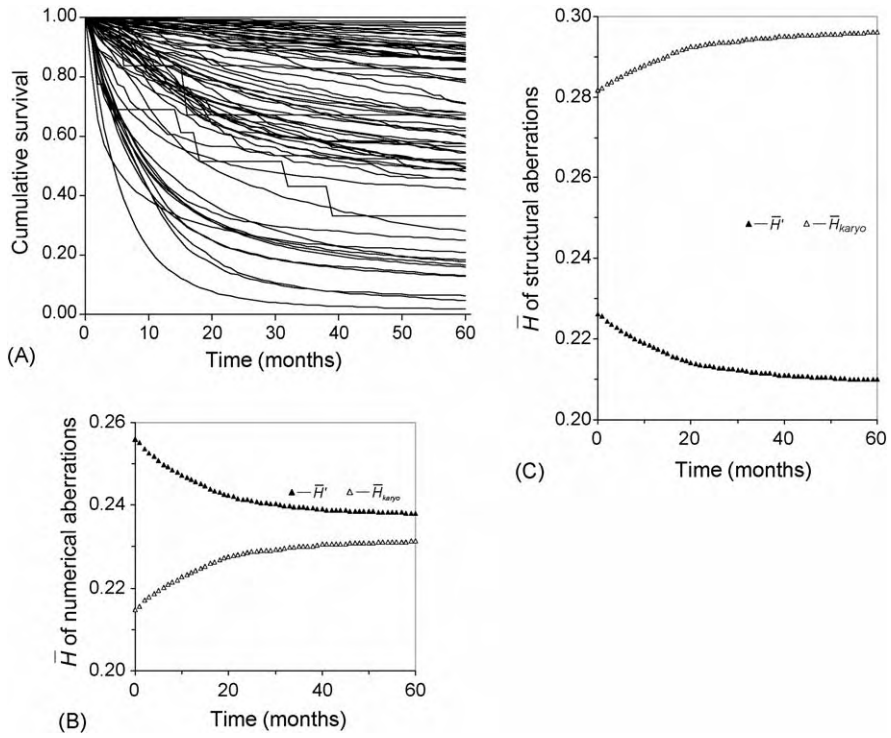


Fig. 5. Association between sample spread of chromosome aberrations and survival statistics. (A) Kaplan–Meier distributions of cumulative survival estimates of 60 months cohorts obtained for each corresponding solid tumor type presented in Table 1. (B) Time evolution of average values of Shannon information functions H' and H_{karyo} of numerical aberrations, obtained by the average functions $\bar{H}'$ and $\bar{H}_{\text{karyo}}$ that estimate the surviving portion of each solid tumor type through the 60 months cohorts (i.e. Eq. (13)). (C) Time evolution of average values of H' and H_{karyo} of structural aberration, as obtained in B. The others $\bar{H}(t)$ functions are presented in Fig. 6.

vary in time, which indicates they correlate with survival. However, the manner they vary differs, with $\bar{H}'$ progressing to its minimum value with time, and $\bar{H}_{\text{karyo}}$ progressing to its maximum value. This result shows a relative increase of those tumor types with low H' and high H_{karyo}, which indicates a higher proportion of *type A* spread in the end of the cohort. Similar results are presented in Fig. 5C, but now for structural abnormalities.

Considering the *type A* description presented in Fig. 2A, it is interesting to note that the relative increase of this spread type along time may be interpret as a larger survival of those cohorts that present tumors with fewer dispersion of its chromosome abnormalities; in other terms, it indicates that recurrence of chromosome aberrations is a good prognosis. Therefore, this result shows that karyotypic diversity correlates with malignancy.

The other $\bar{H}(t)$ functions, although less critical to the analysis, also point to the same time dependent transition, from *type B* to *type A* spread, with the spread degree decreased, indicating that tumor malignancy correlates not only with karyotypic diversity taken from different tumor types but also taken from single tumors (Fig. 6).

4.3. Association between spread and sample size

Now, we would like to address the possible states and dynamics that may underlie the generation process of chromosome aberrations. An important question nowadays is whether chromosome aberrations are random events or arise from an internal deterministic mechanism, which leads to the delicate task of how to quantify the degree of randomness. Among many possible states, the higher degree of randomness refers to a process that can produce independent and identically distributed (i.i.d.) variables [21]. If an observed value in the sequence is influenced by its position in the sequence or biased distributed, the process is not truly random and we could be dealing with other types of stochastic process (e.g. periodic) or even a deterministic one.

Monitoring disorder production in a process may bring information on the process dynamics since – in general – it is different for different types of processes. In this case, a natural question to consider is the manner disorder grows with n, the number of tumors of a given type. Therefore, to decide whether a given sequence is

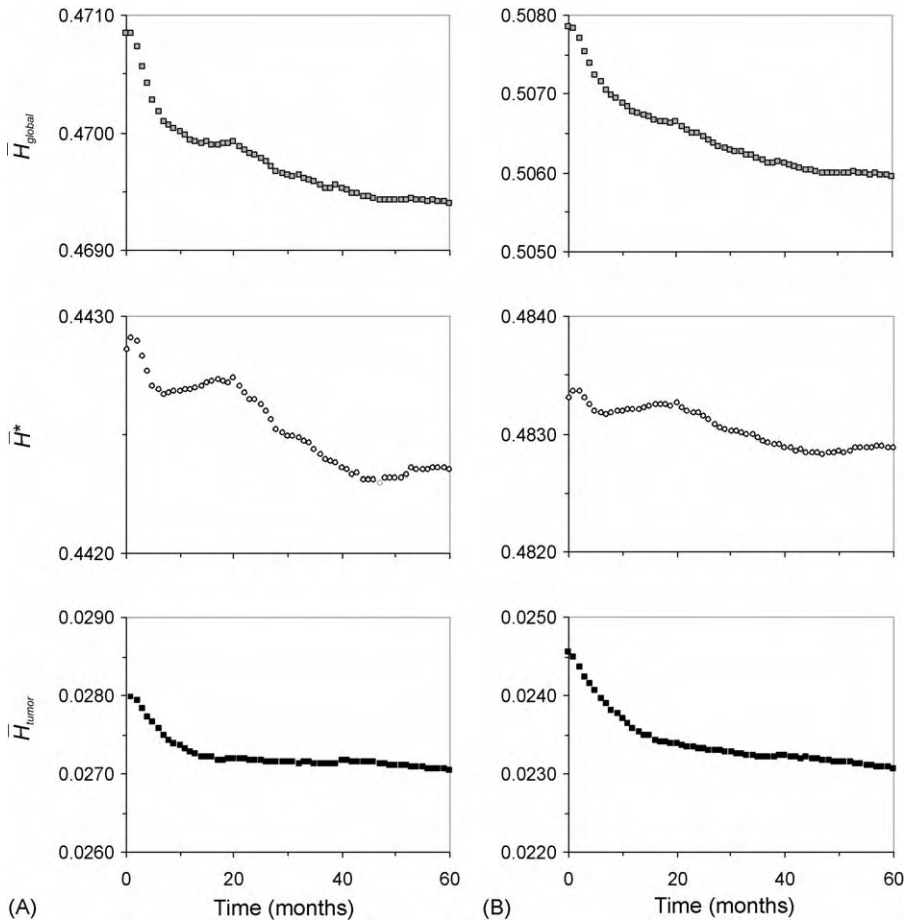


Fig. 6. Time evolution of average values of Shannon information functions H_{global} and the partial contributions H^* and H_{tumor} of numerical (A) and structural chromosome aberrations (B). The values are obtained by the average function $\bar{H}(t)$ that estimates the surviving portion of each solid tumor type through 60 months cohorts (please, see Eq. (13) in the text).

i.i.d., under the null hypothesis of randomness, the values of Shannon information functions of the sequence should follows *type D* spread (as exemplified in Fig. 2D). If there is any dependence between variables, *type A–C* spreads should be considered (Fig. 2A–C).

In order to set theoretical limits, consider first the curves presented in Fig. 7. When the global information function of *type D* spread is plotted as a function of n, we have a constant value, with $H_{global} = 1.0$ (Fig. 7, dashed line). On the other hand, when *type A* and *type B* spreads are plotted in the same way, $H_{global} < 1.0$ for initial n and increase asymptotically to 1.0 (Fig. 7, solid lines), while *type C* spread should be located between these two boundaries (Fig. 7, grey area), presenting a progressive value towards 1.0 as the sample size increases.

The observed values of Shannon information functions of solid tumors are now presented in Fig. 8. As we can observe, the global spread of both numerical and structural chromosome aberrations correlate almost perfectly with n and seems to growth asymptotically, yielding $0.0 < H_{global} < 1.0$. Although H_{global} seems to follow theoretical spread *type* A and/or *type* B (solid line), it presents fluctuations that best locate it in *type C* spread. Indeed, when analyzing other Shannon information functions presented in Fig. 8, it became clear that *type C* spread is the best profile to fit the observed data. Therefore, these results exclude the possibility of both deterministic and i.i.d. processes (i.e. zero value of global information function and *type D* spread, respectively) and show a signature of the randomness degree associated to the production of chromosome aberrations.

5. Discussion

Contrasting with hematological neoplasms, whose clinical value of various cytogenetic abnormalities are

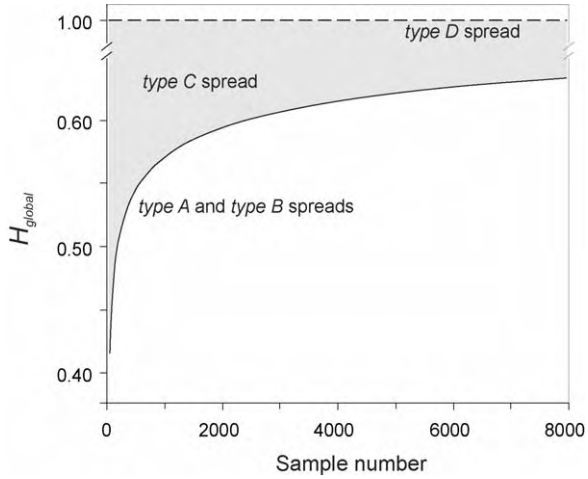


Fig. 7. Theoretical curves of global spread as a function of sample number. When *type D* spread is plotted as a function of n, we have a constant value with $H_{global} = 1.0$ (dashed line). On the other hand, when *type A* and *type B* spreads are plotted in the same way, $H_{global} < 1.0$ for initial n and increase asymptotically to 1.0 (solid lines), while *type C* spread is located between these two boundaries (grey area).

being increasingly appreciated as diagnostic and prognostic tools, the high degree of chromosomal aberrations in solid tumors has limited the use of cytogenetic data as an information source on tumor aggressiveness, patient's diagnosis/prognosis and clinical outcomes [1,6,22]. In solid tumors, the cytogenetic characteristic with some prognostic value is a simple account of an aberrant karyotype (i.e. with many karyotypic alterations) [23–25]. Part of this limitation is attributed to sporadic identification of tumor-specific aberrations [26,27], which suggests that the formation of tumor variants makes each patient a unique case [4]. Unfortunately, this degree of heterogeneity has obscured the identification of a possible set of recurrent chromosome abnormalities that would be specific for each solid tumor type (for statistical review of aberrant cancer karyotypes, see [28]). Two strong and controversial trends to approach this issue are (i) studies attempting to identify specific and nonrandom karyotypic alterations in solid tumors [29–42] and (ii) cancer models focusing on the stochastic nature of random events associated to carcinogenesis [43–49].

For example, although tumor evolution dynamics is frequently described as a successive acquisition of adaptive capabilities [50], these capabilities may be acquired under different scenarios (due to different causes), as genetic mutation, clonal selection, etc. Many of them are complementary, but some authors point to important divergences about the randomness of tumor aberrations and about the effect of genomic instability on tumor initiation and progression [51].

One of these scenarios considers the occurrence of a mutator phenotype, which tends to acquire and accumulate random genetic mutations with higher frequency when comparing to normal cells [6]. The initial mutation is a rare stochastic event, but initiates a cascade of mutations that could leave as signature multiples point mutations randomly distributed through the genome.

A different scenario considers the instability at chromosome level as a necessary condition to form the majority of solid tumors. Accordingly, large chromosomal alterations (i.e. numerical and structural chromosome aberrations) arise due to the collapse of the mechanisms responsible for stability and correct chromosome segregation. In this scenario, random aberration production rate is increased, and when this is simultaneous to either an increase in the mitotic rate or a reduction in cell death rate, successive waves of clonal expansion of adapted variants [4,10] could be generated. Alternatively to chromosome instability theory, some authors sustain that cancer could be a result of chromosome unbalance, which leads to malignant phenotype due to successive asymmetrical mitosis, producing specific chromosome combinations [52,53].

On its turn, clonal selection theory describes the dynamics of a population of cells that, after undergoing random mutations, is submitted to selection, favoring only some of the mutant genotypes. In this scenario, selection footprint would be a nonrandom mutation pattern [8], assuming that the viable mutant genotypes are rare. Other authors further emphasize the mutagenic effects on somatic evolution process, suggesting that the very chromosome errors are far from being random and, in fact, emerge from selective pressure of surrounding environment [7,54].

Concerning these interesting ideas and given several discrepancies among different scenarios, our work may contribute as a measure of randomness. The statistical analysis presented here could help to improve tumor models, exploring possible statements that do not correspond to the type of sample spread profile under observation. Alternatively, we could confront different tumor models, verifying which of them point to the most likely outcome.

For instance, we have shown that karyotypic diversity increases with tumor number and follows theoretical curves of global spread, which is a signature of the randomness associated to chromosome aberration production. This result indicates that, in agreement with karyotypic diversity analysis, the generating process

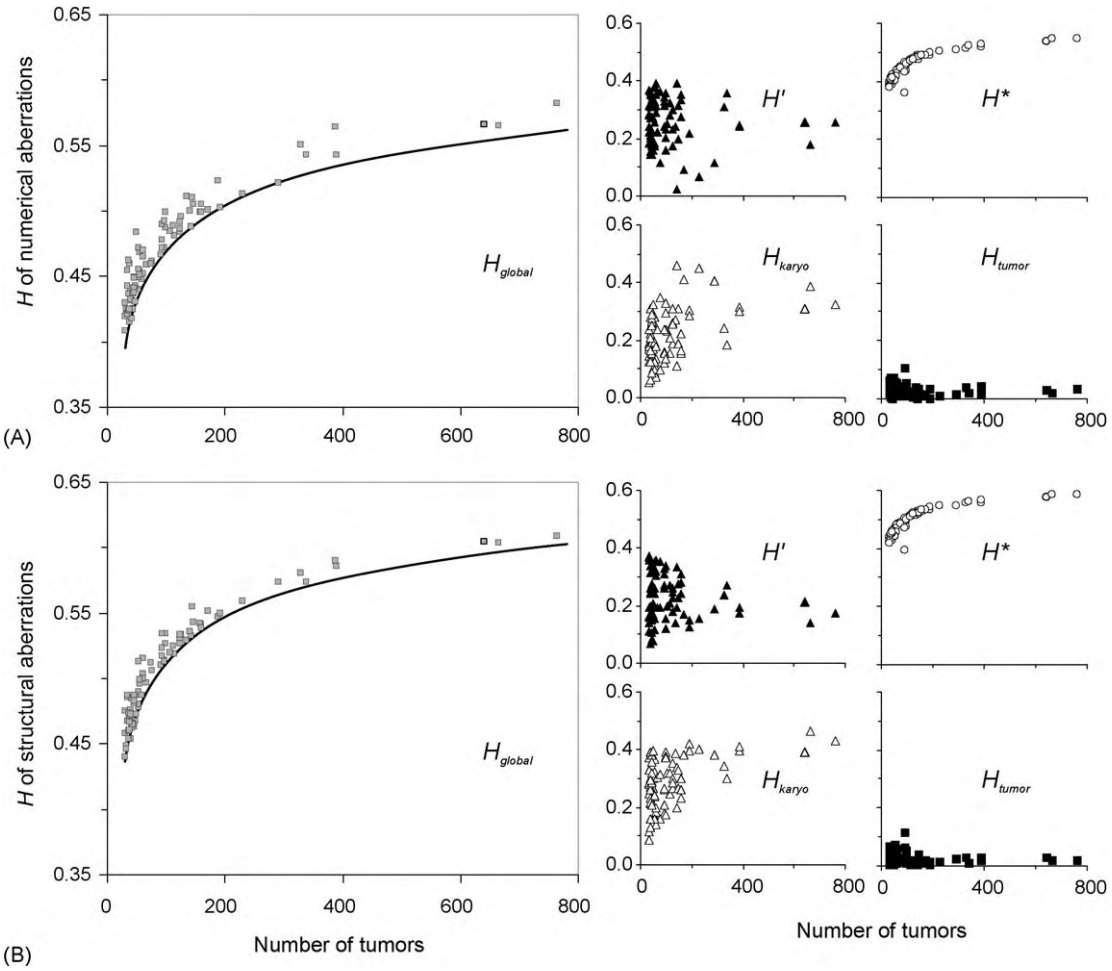


Fig. 8. Values of Shannon information of chromosome aberrations as a function of sample number. Numerical (A) and structural aberrations (B) have global spread (H_{global}) and the partial contributions (H', H_{karyo}, H^* and H_{tumor}) plotted as a function of tumor number. The theoretical curves of *type A* and *type B* spreads (solid lines) are plotted together with the observed H_{global}.

of chromosome aberration is neither deterministic nor totally random, suggesting that it might be interesting to assume a deterministic and a random term complementarily acting in tumor evolution theories.

Finally, we have shown that equally classified tumors differ between themselves concerning karyotypic diversity. Furthermore, we have shown that a classification of diversity levels of tumors may have a diagnostic value. In some tumor types, we can observe processes that tend to form preferential variants, which correspond to those tumors types with better survival indices. In other tumor types, we can observe a high degree of diversity, corresponding to those tumor types with worst survival indices. Therefore, tumor malignancy is correlated with karyotypic diversity estimated by Shannon information, what could yield diagnostic and prognostic tools.

Acknowledgment

This study was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, DF, Brazil).

References

[1] F. Mitelman, B. Johansson, N. Mandahl, F. Mertens, Clinical significance of cytogenetic findings in solid tumors, Cancer Genet. Cytogenet. 95 (1997) 1–8.
[2] O. Fernandez-Capetillo, A. Nussenzweig, Aging counts on chromosomes, Nat. Genet. 36 (2004) 672–674.
[3] D.A. Sinclair, Cell biology. An age of instability, Science 301 (2003) 1859–1860.
[4] C. Lengauer, K.W. Kinzler, B. Vogelstein, Genetic instabilities in human cancers, Nature 396 (1998) 643–649.

[5] R.B. Nicklas, How cells get the right chromosomes, Science 275 (1997) 632–637.
[6] L.A. Loeb, K.R. Loeb, J.P. Anderson, Multiple mutations and cancer, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 776–781.
[7] J. Breivik, The evolutionary origin of genetic instability in cancer development, Semin. Cancer Biol. 15 (2005) 51–60.
[8] O.M. Sieber, K. Heinimann, I.P. Tomlinson, Genomic instability—the engine of tumorigenesis? Nat. Rev. Cancer 3 (2003) 701–708.
[9] A. Fabarius, R. Hehlmann, P.H. Duesberg, Instability of chromosome structure in cancer cells increases exponentially with degrees of aneuploidy, Cancer Genet. Cytogenet. 143 (2003) 59–72.
[10] H. Rajagopalan, M.A. Nowak, B. Vogelstein, C. Lengauer, The significance of unstable chromosomes in colorectal cancer, Nat. Rev. Cancer 3 (2003) 695–701.
[11] M.A.A. Castro, T.T.G. Onsten, R.M.C. de Almeida, J.C.F. Moreira, Profiling cytogenetic diversity with entropy-based karyotypic analysis, J. Theor. Biol. 234 (2005) 487–495.
[12] F. Mitelman, B. Johansson, F. Mertens, Mitelman Database, Mitelman Database of Chromosome Aberrations in Cancer, 2005, http://cgap.nci.nih.gov/Chromosomes/Mitelman.
[13] ISCN (1995): An international system for human cytogenetic nomenclature, S. Karger, Basel, 1995.
[14] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (379–423) (1948) 623–656.
[15] W.S. Kendal, The use of information theory to analyze genomic changes in neoplasia, Math. Biosc. 100 (1990) 143–159.
[16] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D. Parkin, S. Whelan, International Classification of Diseases for Oncology, World Health Organization, Geneva (2000).
[17] SEER, Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence—SEER 11 Regs + AK Public-Use, Nov 2003 Sub (1973-2001 varying), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2004, based on the November 2003 submission, 2005, http://www.seer.cancer.gov.
[18] SEER, Surveillance Research Program, National Cancer Institute SEER*Stat software, version 6.1.4, http://www.seer.cancer.gov/seerstat, 2005.
[19] K.A. Cronin, E.J. Feuer, Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival, Stat. Med. 19 (2000) 1729–1740.
[20] R.A. Gatenby, B.R. Frieden, Information dynamics in carcinogenesis and tumor growth, Mutat. Res. 568 (2004) 259–273.
[21] J.P. Crutchfield, D.P. Feldman, Regularities unseen, randomness observed: levels of entropy convergence, Chaos 13 (2003) 25–54.
[22] D.G. Albertson, C. Collins, F. McCormick, J.W. Gray, Chromosome aberrations in solid tumors, Nat. Genet. 34 (2003) 369–376.
[23] H.L. Flyger, J.K. Larsen, H.J. Nielsen, I.J. Christensen, DNA ploidy in colorectal cancer, heterogeneity within and between tumors and relation to survival, Cytometry 38 (1999) 293–300.
[24] J.H. Lee, S.H. Noh, K.Y. Lee, S.H. Choi, J.S. Min, DNA ploidy patterns in advanced gastric carcinoma; is it a clinically applicable prognosticator? Hepatogastroenterology 48 (2001) 1793–1796.
[25] A. Russo, M. Migliavacca, I. Zanna, M.R. Valerio, M.A. Latteri, N. Grassi, G. Pantuso, S. Salerno, G. Dardanoni, I. Albanese, M. La Farina, R.M. Tomasino, N. Gebbia, V. Bazan, p53 mutations in L3-loop zinc-binding domain, DNA-ploidy, and S phase fraction are independent prognostic indicators in colorectal cancer: a prospective study with a five-year follow-up, Cancer Epidemiol. Biomarkers Prev. 11 (2002) 1322–1331.
[26] F. Mitelman, Recurrent chromosome aberrations in cancer, Mutat. Res. 462 (2000) 247–253.
[27] M. Hoglund, D. Gisselsson, T. Sall, F. Mitelman, Coping with complexity. Multivariate analysis of tumor karyotypes, Cancer Genet. Cytogenet. 135 (2002) 103–109.
[28] M. Hoglund, A. Frigyesi, T. Sall, D. Gisselsson, F. Mitelman, Statistical behavior of complex cancer karyotypes, Genes Chromosomes Cancer 42 (2005) 327–341.
[29] Y. Kitayama, H. Igarashi, F. Watanabe, Y. Maruyama, M. Kanamori, H. Sugimura, Nonrandom chromosomal numerical abnormality predicting prognosis of gastric cancer: a retrospective study of 51 cases using pathology archives, Lab. Invest. 83 (2003) 1311–1320.
[30] F. Micci, C.U. Walter, M.R. Teixeira, I. Panagopoulos, B. Bjerkehagen, G. Saeter, S. Heim, Cytogenetic and molecular genetic analyses of endometrial stromal sarcoma: nonrandom involvement of chromosome arms 6p and 7p and confirmation of JAZF1/JJAZ1 gene fusion in t(7;17), Cancer Genet. Cytogenet. 144 (2003) 119–124.
[31] Y. Jin, C. Martins, L. Salemark, B. Persson, C. Jin, J. Miranda, I. Fonseca, N. Jonsson, Nonrandom karyotypic features in basal cell carcinomas of the skin, Cancer Genet. Cytogenet. 131 (2001) 109–119.
[32] M. Kirchhoff, H. Rose, B.L. Petersen, J. Maahr, T. Gerdes, J. Philip, C. Lundsteen, Comparative genomic hybridization reveals non-random chromosomal aberrations in early preinvasive cervical lesions, Cancer Genet. Cytogenet. 129 (2001) 47–51.
[33] M.R. Mendlick, M. Nelson, D. Pickering, S.L. Johansson, T.A. Seemayer, J.R. Neff, G. Vergara, H. Rosenthal, J.A. Bridge, Translocation t(1;3)(p36.3;q25) is a nonrandom aberration in epithelioid hemangioendothelioma, Am. J. Surg. Pathol. 25 (2001) 684–687.
[34] V. Sirivatanauksorn, Y. Sirivatanauksorn, P.A. Gorman, J.M. Davidson, D. Sheer, P.S. Moore, A. Scarpa, P.A. Edwards, N.R. Lemoine, Non-random chromosomal rearrangements in pancreatic cancer cell lines identified by spectral karyotyping, Int. J. Cancer 91 (2001) 350–358.
[35] C. Jin, Y. Jin, J. Wennerberg, M. Dictor, F. Mertens, Nonrandom pattern of cytogenetic abnormalities in squamous cell carcinoma of the larynx, Genes Chromosomes Cancer 28 (2000) 66–76.
[36] M.A. Nelson, M.D. Radmacher, R. Simon, M. Aickin, J. Yang, L. Panda, J. Emerson, D. Roe, L. Adair, F. Thompson, J. Bangert, S.P. Leong, R. Taetle, S. Salmon, J. Trent, Chromosome abnormalities in malignant melanoma: clinical significance of nonrandom chromosome abnormalities in 206 cases, Cancer Genet. Cytogenet. 122 (2000) 101–109.
[37] A. Pimkhaokham, Y. Shimada, Y. Fukuda, N. Kurihara, I. Imoto, Z.Q. Yang, M. Imamura, Y. Nakamura, T. Amagasa, J. Inazawa, Nonrandom chromosomal imbalances in esophageal squamous cell carcinoma cell lines: possible involvement of the ATF3 and CENPF genes in the 1q32 amplicon, Jpn. J. Cancer Res. 91 (2000) 1126–1133.
[38] L. Gorunova, L.A. Parada, J. Limon, Y. Jin, M. Hallen, I. Hagerstrand, M. Iliszko, Z. Wajda, B. Johansson, Nonrandom chromosomal aberrations and cytogenetic heterogeneity in gallbladder carcinomas, Genes Chromosomes Cancer 26 (1999) 312–321.
[39] Y. Jin, C. Martins, C. Jin, L. Salemark, N. Jonsson, B. Persson, L. Roque, I. Fonseca, J. Wennerberg, Nonrandom karyotypic features in squamous cell carcinomas of the skin, Genes Chromosomes Cancer 26 (1999) 295–303.

[40] R. Taetle, M. Aickin, J.M. Yang, L. Panda, J. Emerson, D. Roe, L. Adair, F. Thompson, Y. Liu, L. Wisner, J.R. Davis, J. Trent, D.S. Alberts, Chromosome abnormalities in ovarian adenocarcinoma: I. Nonrandom chromosome abnormalities from 244 cases, Genes Chromosomes Cancer 25 (1999) 290–300.

[41] L. Gorunova, M. Hoglund, A. Andren-Sandberg, S. Dawiskiba, Y. Jin, F. Mitelman, B. Johansson, Cytogenetic analysis of pancreatic carcinomas: intratumor heterogeneity and nonrandom pattern of chromosome aberrations, Genes Chromosomes Cancer 23 (1998) 81–99.

[42] Y. Jin, F. Mertens, B. Persson, T. Warloe, H.P. Gullestad, L. Salemark, C. Jin, N. Jonsson, B. Risberg, N. Mandahl, F. Mitelman, S. Heim, Nonrandom numerical chromosome abnormalities in basal cell carcinomas, Cancer Genet. Cytogenet. 103 (1998) 35–42.

[43] R.A. Gatenby, T.L. Vincent, An evolutionary model of carcinogenesis, Cancer Res. 63 (2003) 6212–6220.

[44] G. Gregori, L. Hanin, G. Luebeck, S. Moolgavkar, A. Yakovlev, Testing goodness of fit for stochastic models of carcinogenesis, Math. Biosc. 175 (2002) 13–29.

[45] Y. Iwasa, F. Michor, M.A. Nowak, Stochastic tunnels in evolutionary dynamics, Genetics 166 (2004) 1571–1579.

[46] M.P. Little, E.G. Wright, A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data, Math. Biosci. 183 (2003) 111–134.

[47] F. Michor, M.A. Nowak, S.A. Frank, Y. Iwasa, Stochastic elimination of cancer cells, Proc. R. Soc. Lond. B: Biol. Sci. 270 (2003) 2017–2024.

[48] R.S. Root-Bernstein, M.I. Bernstein, A simple stochastic model of development and carcinogenesis, Anticancer Res. 19 (1999) 4869–4876.

[49] R.R. Sarkar, S. Banerjee, Cancer self remission and tumor stability—a stochastic approach, Math. Biosci. 196 (2005) 65–81.

[50] W.C. Hahn, R.A. Weinberg, Rules for making human tumor cells, N. Engl. J. Med. 347 (2002) 1593–1603.

[51] J. Marx, Debate surges over the origins of genomic defects in cancer, Science 297 (2002) 544–546.

[52] R. Li, A. Sonik, R. Stindl, D. Rasnick, P. Duesberg, Aneuploidy vs. gene mutation hypothesis of cancer: recent study claims mutation but is found to support aneuploidy, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 3236–3241.

[53] P. Duesberg, R. Li, A. Fabarius, R. Hehlmann, The chromosomal basis of cancer, Cell. Oncol. 27 (2005) 293–318.

[54] A. Bardelli, D.P. Cahill, G. Lederer, M.R. Speicher, K.W. Kinzler, B. Vogelstein, C. Lengauer, Carcinogen-specific induction of genetic instability, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 5770–5775.

Capítulo 3: *Alteração funcional da rede expressão gênica do sistema de reparo por excisão de nucleotídeos em neoplasias sólidas esporádicas*.

Published online 1 March 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 6* ***1859–1867***
doi:10.1093/nar/gkm061

Impaired expression of NER gene network in sporadic solid tumors

Mauro A. A. Castro[1,2,3,*], José C. M. Mombach[2,4], Rita M. C. de Almeida[2] and José C. F. Moreira[1]

[1]Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003, Brazil, [2]Instituto de Física, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre 91501-970, Caixa Postal 15051, Brazil, [3]Universidade Luterana do Brasil, Av. Itacolomi 3600, Gravataí 94170-240, Brazil and [4]Centro de Ciências Rurais, Unipampa/São Gabriel - Pós-Graduação em Física, Prédio 13, Universidade Federal de Santa Maria, Santa Maria 97105-900, Brazil

Received November 13, 2006; Revised December 28, 2006; Accepted January 19, 2007

ABSTRACT

Nucleotide repair genes are not generally altered in sporadic solid tumors. However, point mutations are found scattered throughout the genome of cancer cells indicating that the repair pathways are dysfunctional. To address this point, in this work we focus on the expression pathways rather than in the DNA structure of repair genes related to either genome stability or essential metabolic functions. We present here a novel statistical analysis comparing ten gene expression pathways in human normal and cancer cells using serial analysis of gene expression (SAGE) data. We find that in cancer cells nucleotide-excision repair (NER) and apoptosis are the most impaired pathways and have a highly altered diversity of gene expression profile when compared to normal cells. We propose that genome point mutations in sporadic tumors can be explained by a structurally conserved NER with a functional disorder generated from its entanglement with the apoptosis gene network.

INTRODUCTION

Cancer cells have large and small abnormalities in their genetic material: additional or missing chromosomes, mutated genes and other types of alterations. The lost of genome stability pathways is associated with genetic deterioration of cancer cells and is one of the most important aspects of carcinogenesis. In fact, mutations in mismatch repair (MMR), nucleotide-excision repair (NER), base-excision repair (BER) and recombinational repair genes have been causally implicated in the acquisition of a genome instability phenotype (1).

Genome instability in solid tumors originates from either somatic mutations (observed in the majority of sporadic cancers) or germline mutations (associated to rare hereditary cancer syndromes). Considering the list of repair genes presented in Cancer Gene Census (2), germline mutations can be observed in NER, BER and MMR, while somatic mutations are described only in recombinational repair (homologous recombination and non-homologous end joining). On the other hand, mutations in apoptotic genes are recurrently observed in both types of solid tumors as listed in the census.

The genotype signature of the malfunctioning of these stability gene networks is 2-fold: aneuploidy (e.g. translocations, gain or loss of entire or large parts of chromosomes) and/or random point mutations (e.g. nucleotide changes randomly distributed throughout the genome) (3). The omnipresence of random point mutations in sporadic solid tumors (4) and the recurrent absence of mutations in nucleotide repair genes (2) suggest a functional deficiency in these stability pathways without structural alterations in the related DNA sequence.

There are different views explaining how a cell loses genome stability and acquires a cancerous phenotype (5,6). In one proposed scenario, large chromosomal changes are required for triggering the onset of cancer, such as varying the number of whole chromosomes or cutting and/or pasting their fragments among different chromosomes. Then either the expression of unbalanced gene dosage (7) and/or alterations in mitotic check points (8) can, under adequate conditions, give place to a cancer. An alternative idea proposes that cancer cells

*To whom correspondence should be addressed. Tel: +55 51 33085577; Fax: +55 51 33085540; Email: mauro@ufrgs.br
Correspondence may also be addressed to Rita de Almeida. Tel: +55 51 33086521; Fax: +55 51 33086286; Email: rita@if.ufrgs.br

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

have a 'mutator phenotype' that favors the acquisition of point mutations, which eventually affect tumor suppressors or oncogenes yielding to cancer (9). Supporting this idea, a list of mutated genes found in human colorectal and breast cancer covering several gene functions shows that point mutations are the most common alterations found throughout the genome of cancer cells (over 87%) (10).

The two scenarios are qualitatively possible, since both offer explanations to the typical chromosome configurations and nucleotide alterations of a cancer cell. In order to discriminate between different scenarios, studies of chromosome and gene structures should be complemented by quantitative analysis of gene expression of cancer cells. As a contribution in this direction, we present here a pioneer, comprehensive statistical analysis of 10 gene expression pathways in normal and cancer cells using serial analysis of gene expression (SAGE) data from the public gene expression resource (SAGE Genie) (11) available at Cancer Genome Anatomy Project (CGAP) (12).

MATERIALS AND METHODS

Data selection

Human cancer and normal tissue SAGE libraries are retrieved using SAGE Library Finder tool at SAGE Genie website (http://cgap.nci.nih.gov/SAGE) based on the search criteria: tag length (short 10 bp), tissue preparation [bulk, short-term culture (STC), antibody purified (ABP), microscope dissected (MCD) or cell line] and tissue histology (cancer or normal). The final list is presented in Supplementary Tables S1 to S4 and contains only cancer libraries that had at least one normal equivalent tissue library, and vice-versa, matching both the search criteria. The list of SAGE libraries is also retrieved for tag-to-gene corresponding libraries using SAGE Absolute Level Lister (SALL) tool at SAGE Genie website. This tool links SAGE unique tags to genes via UniGene cluster IDs (e.g. it packs into one file *Tag Sequence*, *Tag Frequency*, *UniGene Cluster ID* and *Gene Symbol*). SALL database retrieval was conducted in June 2006 (UniGene Build #191 and #192). Therefore, according to the search criteria, we retrieved the largest human SAGE collection up to date at NCI's Cancer Genome Anatomy Project for the analysis presented here.

Mathematical definitions and analysis of SAGE libraries

In the SAGE database, a SAGE library corresponds to one tumor sample exam, which is made from mRNA extracts from different tissue preparations (bulk, short-term culture, antibody purified, microscope dissected or cell line) and histology (cancer or normal), as described in detail by Lash *et al.* (13). One such library gives the amount of every detected transcript in the sample, each one being labeled by a 10-letter tag, corresponding to 10 bases close to the poly-A tail, whose length is long enough to discriminate every possible transcript. Transcripts related to different gene networks may be grouped and used to quantify and characterize their expression activity. Here we analyze both the amount of transcript production and its diversity in ten gene pathways, chosen due to either their recognized relation with genome stability (apoptosis, chromosome stability, mismatch repair, nucleotide-excision repair, base-excision repair and recombinational repair) or, as a control group, due to their essential life-supporting activities (ribosome, ATP synthase, electron transport chain and glycolysis). The tumor types were selected such that they present a library of normal cells, to be used as control. The complete list of SAGE libraries and details about database search are available in the Supporting Online Material.

To obtain a quantitative expression of sample distribution of SAGE tags, we have measured the information content of SAGE libraries using Shannon Information Theory (14–18) defined as follows. Consider n as the number of all selected SAGE libraries of a given tumor type. Each library of this set is labeled by $\alpha(\alpha = 1, \ldots, n)$ and has N_α tags, among M_α possible ones, that is, possible transcripts. For a given SAGE library in this set, we can define $s(i,\alpha)$ as being the number of transcripts (tags) of a given type i, $(i = 1, \ldots, M_\alpha)$, whose sum for a given α adds up to N_α. The probability $p(i,\alpha)$ that, among the N_α tags of the α-library, a randomly chosen transcript is of the type i is written as

$$p(i,\alpha) = \frac{s(i,\alpha)}{N_\alpha}, \quad \mathbf{1}$$

such that $\sum_i p(i,\alpha) = 1$. The normalized Shannon information function H_α is defined as

$$H_\alpha = -\frac{1}{\ln(M_\alpha)} \sum_i^{M\alpha} p(i,\alpha) \ln p(i,\alpha), \quad \mathbf{2}$$

where we have divided all terms by the factor $\ln(M_\alpha)$ in order to normalize the quantities, guaranteeing that $0 \leq H_\alpha \leq 1$. The idea is to compare among samples of different tissues that may present different numbers of M_α possibilities (e.g. different numbers of possible transcripts). While N_α reflects gene expression activity (the amount of tags in the αth library), H_α reflects the spread of the distribution $s(i,\alpha)$, i.e. it measures the diversity that exists in the αth library.

Finally, in order to normalize the quantities by sets of tags, taking as reference normal tissue histology, we define the relative diversity h_α for any given set of genes as

$$h_\alpha = \frac{H_\alpha^c}{H_\alpha^c + H_\alpha^r}, \quad \mathbf{3}$$

where H_α^c and H_α^r are, respectively, the diversity of cancer and normal SAGE libraries. Observe that $0 \leq h_\alpha \leq 1$, and $h_\alpha < ½$ implies $H_\alpha^c < H_\alpha^r$, that is, the transcript distribution in the αth library is narrower in cancer cells than in the normal tissue, while $h_\alpha > ½$ represents the inverse case. In analogy, the relative gene expression activity n_α of the α library is defined as

$$n_\alpha = \frac{N_\alpha^c}{N_\alpha^c + N_\alpha^r}, \quad \mathbf{4}$$

where N_α^c and N_α^r are, respectively, the gene expression activity of cancer and normal tissue (i.e. number of

SAGE tags). Again, $0 \leq n_\alpha \leq 1$, and $n_\alpha < ½$ implies $N_\alpha^c < N_\alpha^r$, that is, in this library the cancer cells have lower gene activity, producing less transcripts than the normal case (e.g. Supplementary Figures S1 and S2).

Diversity of gene expression pathways

To estimate the diversity of expression pathways related to genome stability, we carry out the following steps: (i) define gene expression pathways of interest; (ii) identify groups of genes that best represents each pathway; (iii) identify the best SAGE tags of these genes—among all possible tags—presented in the collection of tags of each SAGE library; (iv) arrange the SAGE tags into a separate file—one for each pathway; (v) verify the agreement of the original database with subset files; (vi) build a curated database; and (vii) estimate the degree of diversity of pooled SAGE tags, as defined in mathematical definitions section.

We focused this study in six genomic stability pathways (apoptosis, chromosome stability, mismatch repair, nucleotide-excision repair, base-excision repair and recombinational repair). Here, the group of genes representing each gene expression pathway is considered as a group of *UniGene Cluster IDs*. The lists of selected genes and pathways are presented in Supplementary Tables S9–S14, including references used for selection. In order to link genes and SAGE tags, we used the UniGene number as common identifier. Next, we checked libraries looking for UniGene number duplication. This process reveals that, in the original database retrieved from the SAGE Genie website, several pooled tags (Unique Tags) present the same UniGene cluster ID and, therefore, they are pooled as single UniGene number. The curated libraries are then used to build the subset files of tags—one for each gene expression pathway. These files are used in the final step to estimate the diversity of gene expression. The curated database and a Microsoft Excel™ spreadsheet that automatically calculates diversity scores for multiple pathways are available upon request.

We have considered several internal controls in order to use as invariant references among cancer and normal SAGE libraries. The idea is to estimate the diversity of a gene expression pathway that produces co-expressed genes, ideally always in the same proportion, independently of tissue type or histology. For this purpose, we consider the following criteria for selection: (i) gene products should be present in stoichiometric amounts because they are part of the same stable complex and/or are functionally associated at the molecular level (19); (ii) the candidate pathway must occur in all cell types because they are necessary for the cell survival and/or are implicated in basal cell metabolism (20); (iii) the pathway must be involved in core, conserved biological functions (21). Among likely candidates, we evaluated four co-expressed gene groups (named here by its final products): (i) ribosome (e.g. ribosomal proteins); (ii) ATP synthase; (iii) electron transport chain; and (iv) glycolysis. The lists of selected control pathways are presented in Supplementary Tables S5–S8, including references used for selection.

Pairwise data of cancer versus normal

After estimating SAGE library parameters for each expression pathway, as defined above, the values of cancer libraries are compared to normal ones. The pairwise comparisons produce two distribution types of matched pairs. One related to SAGE library properly (overall SAGE tags); other related to gene expression pathways (a subset of SAGE tags). The pairwise comparisons are then plotted in order to examine either the entire data distributions or by tissue types individually. The number of pairwise libraries—cancer versus normal—is limited by the number of SAGE libraries available in SAGE Genie website up to date (http://cgap.nci.nih.gov/SAGE). Therefore, each cancer library is paired with each normal library of the same tissue type, as presented in Supplementary Tables S15–S16, providing 493 pairwise comparisons—368 for solid tumors (ST) and 125 for cell lines(CL): brain (ST = 192; CL = 20); breast (ST = 75; CL = 45); cerebellum (ST = 40; CL = 8); stomach (ST = 28); prostate (ST = 12; CL = 32); colon (ST = 4; CL = 8); pancreas (ST = 4; CL = 8); skin (ST = 3); ovary (ST = 3; CL = 2); liver (ST = 2; CL = 2); lung (ST = 2); kidney (ST = 1); peritoneum (ST = 1); thyroid (ST = 1). Indeed, there are only 492 pairwise libraries in SAGE tag-to-gene analysis because the skin library 'SAGE_Skin_melanoma_B_DB3' was not integrated with SAGE Absolute Level Lister (SALL) tool at the period of our study.

Analysis of protein/gene interaction networks

The protein–protein interaction network associating genes of the six genome stability pathways is generated using the database STRING ('search tool for the retrieval of interacting genes/proteins') (22,23) with input options 'databases', 'experiments' and 70% confidence level. In order to identify each gene in the database, we used both HUGO ID (24) and Ensembl Peptide ID (25) (Supplementary Table S17). Alternatively, the amino acid sequence of a given protein is supplied to identify the corresponding entry. The results from the search are saved in data files 'tab-delimited text fields' describing edge relationships and then handled in Medusa application (26) (i.e. optimized software for accessing protein interaction data from STRING). Pathways are discriminated by different colors and data are crossed with Cancer Gene Census (2) in order to indicate genes whose somatic mutations have been reported to be causally implicated in human cancer. The complete file matching entry IDs, data interactions and mutated genes are available upon request. Finally, graphs are exported to postscript files to have figure quality improved and edited in CorelDraw® graphic design tools (Corel Corp., Ottawa, Canada).

Statistical analysis

Under the null hypothesis, the stochastic contrast among k expression pathways is given by $h_{\alpha A} = h_{\alpha B} = \cdots = h_{\alpha K}$. Although the distributions do not seriously deviate from

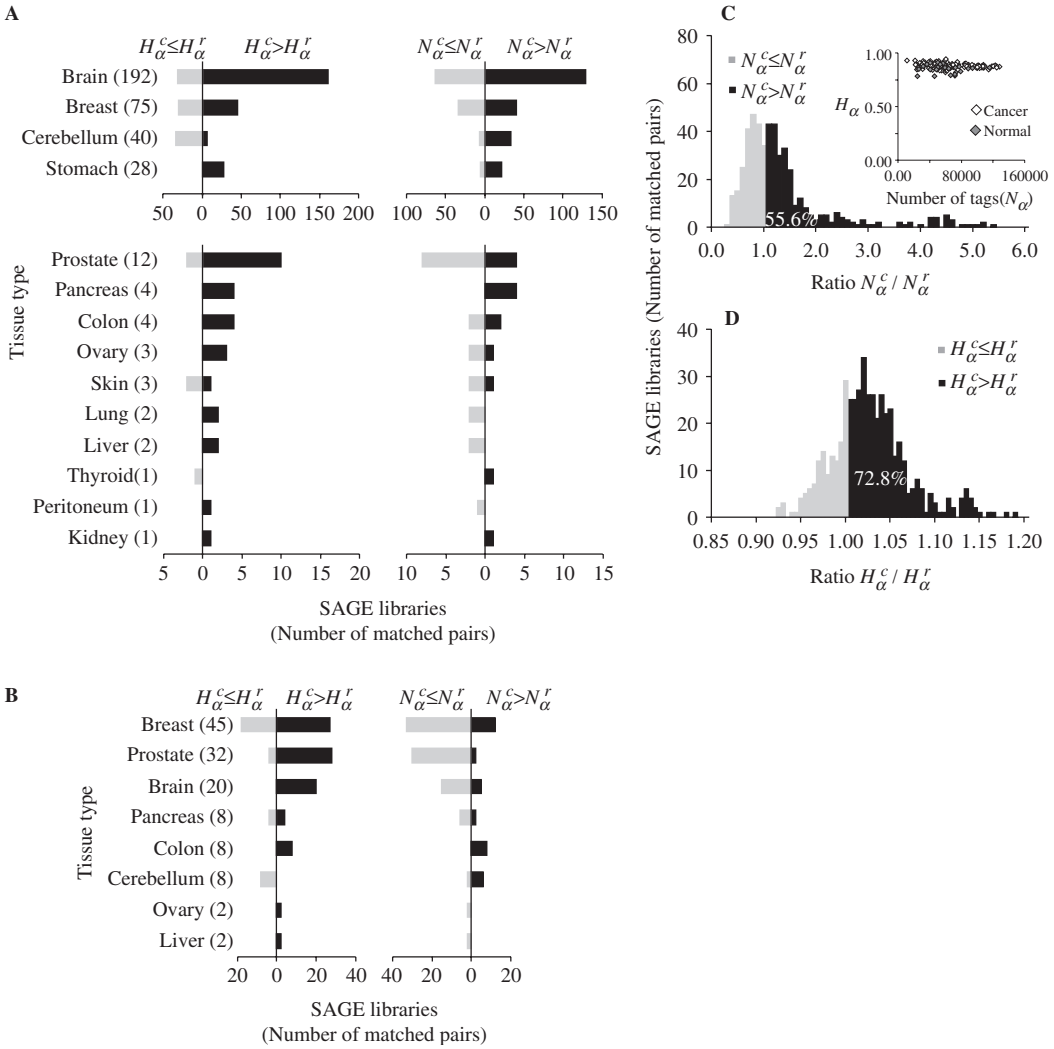


Figure 1. Distributions H_α^c/H_α^r and N_α^c/N_α^r for different tissue types. (**A**) Nonparametric distribution of matched pairs of solid tumor libraries. The length of the black horizontal stripes correspond to the number of libraries presenting ratios of H_α^c/H_α^r and N_α^c/N_α^r that are larger than one, while the gray stripes to the left correspond to the number of libraries with ratios less than one. The number of matched pairs is indicated for each tissue type. (**B**) Cell lines, as described in A. (**C**) Histogram distribution of matched pairs of libraries compared by the number of SAGE tags. Inset shows diversity H_α as a function of the number N_α for cancer and normal libraries. (**D**) Histogram distribution of matched pairs of libraries compared to diversity of SAGE tags. The percentage of matched pairs with $N_\alpha^c/N_\alpha^r > 1$ and $H_\alpha^c/H_\alpha^r > 1$ are indicated. The list of SAGE libraries is available in Supplementary Tables S1–S4.

normality, as observed in Figures 1 and 2, with equal sample sizes among pathways, the data failed to meet the assumptions of ANOVA for normality and homogeneity of variance and, thus, we used Kruskal–Wallis one-way analysis of variance followed by Mann–Whitney test for comparisons. The tests were performed in SPSS nonparametric statistical package (SPSS for Windows, release 14.0.0. SPSS Inc., Chicago, IL). Values are expressed as mean ± SEM. Significance is considered at $P < 0.05$.

RESULTS AND DISCUSSION

In what follows we present the results concerning the above defined quantities for different tumor types. First, we compare the gene expression activity N_α^c and diversity H_α^c of each cancer SAGE library from several tissue types with its respective normal case (Figure 1A and B). The great majority of cancer tissues showed an increased gene expression activity and diversity, since in almost all cases the majority of libraries present $N_\alpha^c/N_\alpha^r > 1$ and $H_\alpha^c/H_\alpha^r > 1$. The resulting global histogram distributions of cancer-normal pairwise libraries are then plotted in Figure 1C and D, showing that the number and diversity of tags in cancer libraries are higher than normal tissue in 56.6 and 72.8% of the cases.

To evince the contribution from the different gene networks, which is not shown in the global histograms, in Figure 2 we present the same matched pairs of SAGE libraries, assessed for subsets of tags corresponding

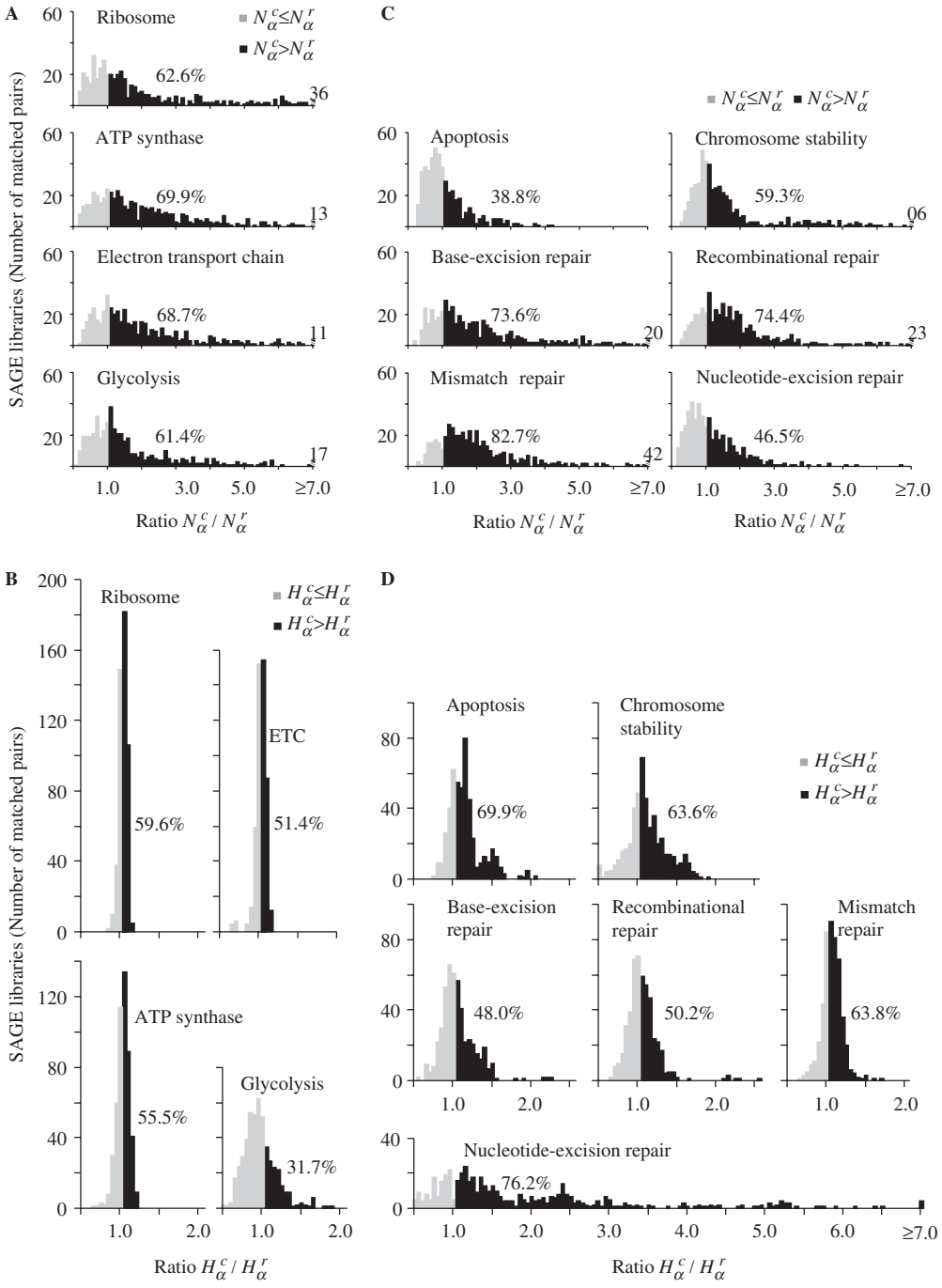


Figure 2. Distributions of H_α^c/H_α^r and N_α^c/N_α^r for different gene expression pathways. (**A**) and (**B**) Gene expression pathways involved in core cell functions. The group of genes of each pathway is presented in Supplementary Tables S5–S8, defined accordingly to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (35). (**C**) and (**D**) Gene expression pathways involved in genome stability functions. The group of genes of each pathway is presented in Supplementary Tables S9–S14, defined according to several references (35–38). The percentage of matched pairs with $N_\alpha^c/N_\alpha^r > 1$ or $H_\alpha^c/H_\alpha^r > 1$ is indicated in all distributions. Another supplemental file listing all genes involved in core biological function and genome stability pathways are provided in spreadsheet format (Supplementary Tables S17 and S18). These data are also presented in log-log scatter plots (Supplementary Figures S8 and S9) and a correlation analysis between SAGE tag and SAGE tag-to-gene data is presented in Supplementary Figure S10.

to different gene expression pathways. Comparing the histograms presented in Figure 2A and C, we conclude that apoptosis and NER pathways present a smaller number of cancer libraries with $N_\alpha^c > N_\alpha^r$, indicating reduced gene expression in these pathways. When considering the diversity of tags (Figure 2B and D) the results are opposite, since apoptosis and NER present more cancer libraries with $H_\alpha^c > H_\alpha^r$, indicating increased diversity in these gene expression pathways. Furthermore, observe the contrast between the diversity distributions

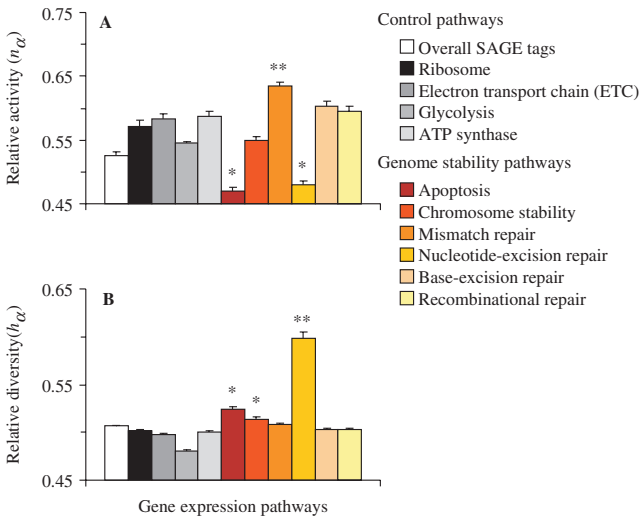


Figure 3. Stochastic contrasts among gene expression pathways according to diversity and number of SAGE tags. (**A**) Relative activity n_α as defined in Equation (4). (**B**) Relative diversity h_α as defined in Equation (3). The values are expressed as mean ± SEM ($n = 492$). Statistical analyses are carried out by Kruskal–Wallis one-way analysis of variance followed by Mann–Whitney test for comparisons. *Different from controls with $P < 0.001$; **different from others with $P < 0.001$.

for Ribosome and NER pathways: Ribosome histogram is narrowly peaked around 1, indicating that cancer and normal cells have almost identical transcript profiles for this highly conserved pathway, while NER diversity histogram presents a broad distribution, biased to $H_\alpha^c/H_\alpha^r > 1$. In fact, except for NER, all other pathways presented a uniform gene activity change in the sense that the relative profiles are conserved. Moreover, NER presents a negative correlation between gene activity and diversity (log–log scatter plots of core cell functions and genomic stability pathways are presented in Supplementary Figures S3 and S4). This result indicates that NER gene activity decreases due to the reduction of normal gene expression peaks in cancer cells. In this way, the increased relative contribution of a broad profile background enhances diversity.

In order to consolidate these results and simultaneously compare all gene expression pathways, we present in Figure 3A the average values of the relative activity n_α for each gene network. As we can observe, NER and apoptosis present the lowest amount of relative activity ($P < 0.001$), indicating again an altered state of gene expression. In contrast, NER has the highest relative diversity ($P < 0.001$) (Figure 3B), which corroborates that the low level of gene expression occurs together with changes on gene expression profile of this repair pathway. Since gene expression in cell lines could not reliably reflect the gene expression in bulk tissues, we also present in Supplementary Figures S5 and S6 an individual analysis, by tissue type and preparation. Overall, the results indicate that the conclusions drawn on these observations follow the same outcome of the pooled analysis, especially considering the most representative solid tumors in the sample (i.e. brain, breast, cerebellum and stomach).

In cancer cells programmed cell death mechanism is in general structurally impaired (27), what is coherent with the observed gene expression profile of apoptosis transcripts. However, NER is in general structurally intact in sporadic solid tumors, since no somatic mutations in NER genes have been reported to be causally implicated in oncogenesis (2). The observed transcript profile then suggests that NER-transactivation-dependent functions are affected in cancer cells.

As both apoptosis and NER networks are simultaneously affected, a causal correlation is plausible, considering that both networks are entangled. One scenario is suppression of NER transcription activity due to global alterations in cell-death control. A second alternative would be apoptosis and NER impairment caused by the malfunctioning of a gene, either due to failures in activation-dependent functions or gene mutations. A natural candidate in this last case is *TP53* based on the wealth of experimental evidence that this gene plays a role in both apoptosis and NER networks (28). As an illustration of these two possibilities we present in Figure 4A a protein–protein interaction network associating genes of the six genome stability pathways investigated here. The graph is generated using database STRING (22) with input options 'Experimental/Biochemical Data' and 'Association in Curated Databases,' with 70% confidence interval, meaning in this graph that genes are linked whenever direct (physical) or indirect (functional) protein interaction is reported in curated databases. Figure 4A suggests that either scenario is possible. This graph indicates a strong interaction among all pathways with significant overlapping among different genes. Concerning apoptosis and NER, *TP53* plays a key role, connecting both networks (Supplementary Figure S7). In fact, there are many reports in the literature pointing *p53* affecting both dependent and independent transactivation NER functions, as well as affecting apoptosis (28–31). Also, it is reasonable to assume that damage in a specific gene function may affect its neighbors in the network, causing perturbations that may disrupt the whole network. The implication of these observations is that the vulnerability of NER and apoptosis could reside in the same core 'node.' Other scenarios are also possible, as defective genes independently acting on both networks, but then more genes should be simultaneously impaired in order to account for the results presented here.

Furthermore, observing the network architecture and the organization of interactions in Figure 4A, one can see that NER topology suggests the existence of a functional module overlapping three pathways, i.e. NER, MMR and chromosome stability (module ρ). To quantify the interaction pattern among genes in the network we calculated the connectivity k, defined as the number of links that a given node has with other nodes (32). Figure 4B presents the nodes by increasing connectivity. There are two striking features in this figure. First, all ρ-nodes present high connectivity. Second, there are no mutations in high connectivity nodes. This may be indicating that the joint functions of the nodes in ρ are essential to turn the cell viable, what would play the role of a protection mechanism for the *organism* against proliferation of

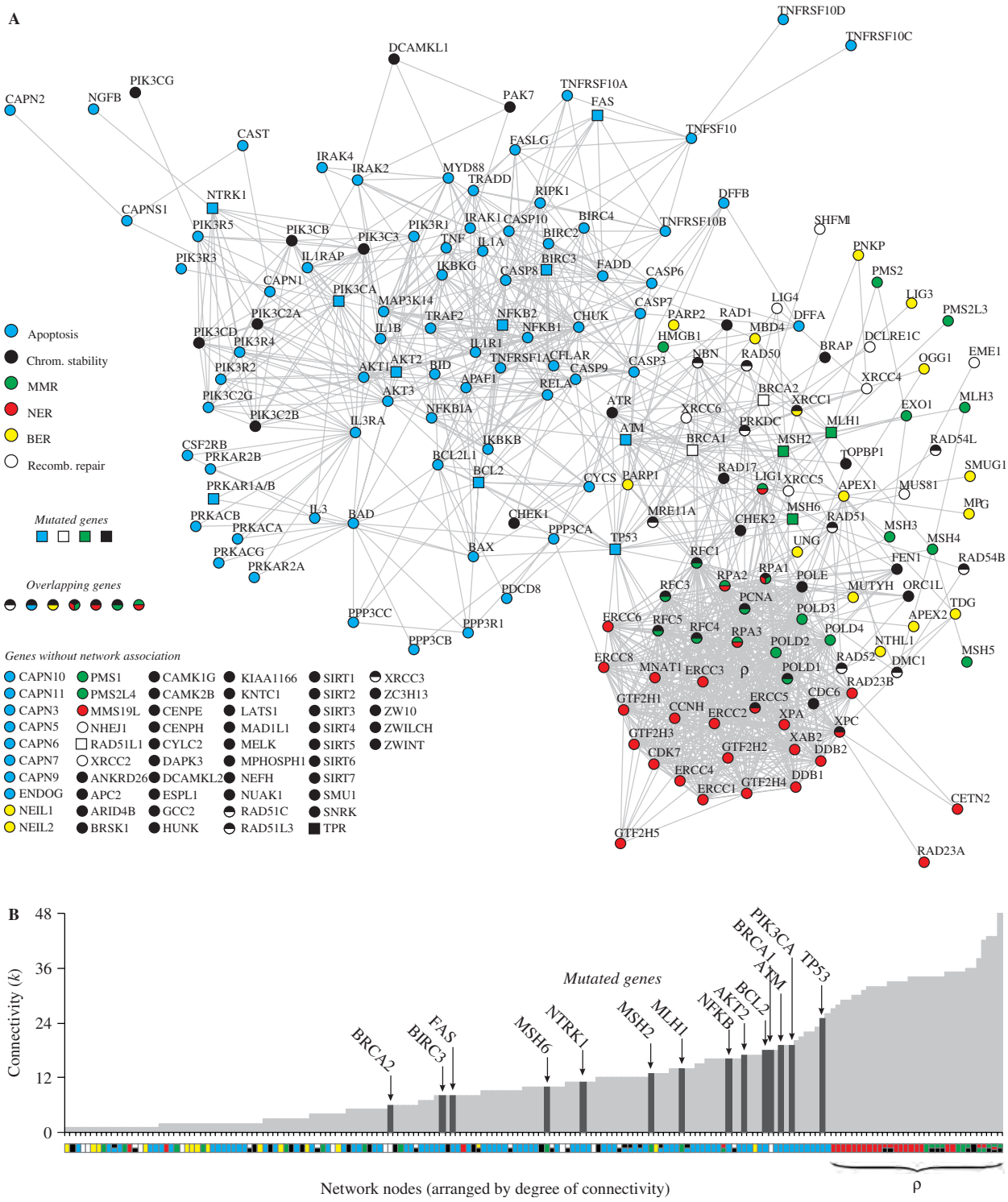


Figure 4. Graph of interactions among genes involved in apoptosis and DNA repair pathways generated using database STRING (22) with input options 'Experimental/Biochemical Data,' 'Association in Curated Databases,' and 70% confidence level. (**A**) Different pathways are represented in different colors. Nodes with more than one color represent genes participating in more than one pathway. Square nodes represent genes whose somatic mutations have been reported to be causally implicated in oncogenesis (2). The group of genes of each pathway is presented in details in Supplementary Tables S9–S14, defined according to several references (35–38). Genes without known interactions with other genes are listed in the bottom left of the figure. (**B**) Connectivity k of interacting nodes, which shows the number of links that a given node has with other nodes. The color of a node indicates its pathway, as in (A); mutated genes are also indicated. Mismatch repair (MMR), nucleotide-excision repair (NER), base-excision repair (BER).

mutation-prone clones. Furthermore, *TP53* appears as the mutated gene with highest k degree, what could be interpreted as a gene with high enough connectivity such that a mutation has a great effect in disrupting the cell apoptosis and repair system, but low enough connectivity such that the cell is still viable. Following this speculative point of view, mutations in more than one gene with lower connectivity should be required to disrupt both apoptosis and repair systems, that is, to cause cancer. This would hence explain why mutations in other genes than *TP53* are less probable and why *TP53* is not mutated in all tumors (10). In other words, it may happen that higher connectivity and higher mutation probability in cancer cells are correlated up to a connectivity threshold, when mutations render the cell unviable. This possibility is consistent with the growth failure and premature death described in at least three NER-deficient mouse models (33) and with the correlation between protein connectivity and indispensability described in yeast proteome (34).

In summary, the above statistical analysis indicates that, relative to normal tissues, cancer cells present (i) enhanced overall gene expression, indicating a higher transcriptional activity, (ii) decreased apoptosis and NER gene expression activity, (iii) conserved expression profiles for control gene pathways, (iv) high diversity in transcript profiles for NER, suggesting suppression of expression peaks, enhancing the relative background contribution. It is then possible that conditions that disable apoptosis, probably due to mutations in apoptotic genes, also affects NER-transactivation-dependent functions via *p53*. NER malfunctioning could then account for random point mutations scattered throughout the cancer cell genome. Furthermore, the analysis of network connectivity points to a highly connected module (ρ) involving genes from NER, MMR and chromosome stability where mutations are recurrently absent in cancer cells, but whose functions could be impaired by mutations in peripheral genes, linking this module with apoptosis.

A natural perspective of these findings is to extend the same approach for diagnostic purpose, for example, and thus test the robustness of our conclusions, which could also indicate whether this model can be applied to identification of other pathways involved in cancer progression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work has been partially supported by Brazilian Agencies FAPERGS, CNPq and CAPES. We acknowledge KEGG, STRING and CGAP/SAGE databases for providing public access to its data. Funding to pay the Open Access publication charge was provided by CNPq (grant 140947/2006-0).

Conflict of interest statement. None declared.

REFERENCES

1. Hoeijmakers,J.H.J. (2001) Genome maintenance mechanisms for preventing cancer. *Nature*, **411**, 366–374.
2. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
3. Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.
4. Venkatesan,R.N., Bielas,J.H. and Loeb,L.A. (2006) Generation of mutator mutants during carcinogenesis. *DNA Repair*, **5**, 294–302.
5. Marx,J. (2002) Debate surges over the origins of genomic defects in cancer. *Science*, **297**, 544–546.
6. Merlo,L.M.F., Pepper,J.W., Reid,B.J. and Maley,C.C. (2006) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924–935.
7. Duesberg,P., Li,R., Fabarius,A. and Hehlmann,R. (2005) The chromosomal basis of cancer. *Cell. Oncol.*, **27**, 293–318.
8. Rajagopalan,H., Nowak,M.A., Vogelstein,B. and Lengauer,C. (2003) The significance of unstable chromosomes in colorectal cancer. *Nat. Rev. Cancer*, **3**, 695–701.
9. Loeb,L.A., Loeb,K.R. and Anderson,J.P. (2003) Multiple mutations and cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 776–781.
10. Sjoblom,T., Jones,S., Wood,L.D., Parsons,D.W., Lin,J., Barber,T.D., Mandelker,D., Leary,R.J., Ptak,J. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
11. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11287–11292.
12. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
13. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
14. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
15. Kendal,W.S. (1990) The use of information theory to analyze genomic changes in neoplasia. *Math. Biosc.*, **100**, 143–159.
16. Castro,M.A.A., Onsten,T.T.G., de Almeida,R.M.C. and Moreira,J.C.F. (2005) Profiling cytogenetic diversity with entropy-based karyotypic analysis. *J. Theor. Biol.*, **234**, 487–495.
17. Gatenby,R.A. and Frieden,B.R. (2004) Information dynamics in carcinogenesis and tumor growth. *Mutat. Res.*, **568**, 259–273.
18. Castro,M.A.A., Onsten,T.G.H., Moreira,J.C.F. and de Almeida,R.M.C. (2006) Chromosome aberrations in solid tumors have a stochastic nature. *Mutat. Res.*, **600**, 150–164.
19. Teichmann,S.A. and Babu,M.M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.*, **20**, 407–410.
20. Thellin,O., Zorzi,W., Lakaye,B., De Borman,B., Coumans,B., Hennen,G., Grisar,T., Igout,A. and Heinen,E. (1999) Housekeeping genes as internal standards: use and limits. *J. Biotechnol.*, **75**, 291–295.
21. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
22. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
23. Mering,C.V., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional\ associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
24. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the human gene nomenclature database, 2004 updates. *Nucleic Acids Res.*, **32**, D255–D257.

25. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
26. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
27. Zhivotovsky,B. and Kroemer,G. (2004) Apoptosis and genomic instability. *Nat. Rev. Mol. Cell Biol.*, **5**, 752–762.
28. Sengupta,S. and Harris,C.C. (2005) p53: Traffic cop at the crossroads of DNA repair and recombination. *Nat. Rev. Mol. Cell Biol.*, **6**, 44–55.
29. Hwang,B.J., Ford,J.M., Hanawalt,P.C. and Chu,G. (1999) Expression of the p48 xeroderma pigmentosum gene is p53-dependent and is involved in global genomic repair. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 424–428.
30. Adimoolam,S. and Ford,J.M. (2002) p53 and DNA damage-inducible expression of the xeroderma pigmentosum group C gene. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 12985–12990.
31. Rubbi,C.P. and Milner,J. (2003) p53 is a chromatin accessibility factor for nucleotide excision repair of DNA damage. *EMBO J.*, **22**, 975–986.
32. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
33. Hoogervorst,E.M., van Steeg,H. and de Vries,A. (2005) Nucleotide excision repair and p53-deficient mouse models in cancer research. *Mutat. Res.*, **574**, 3–21.
34. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
35. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
36. Wang,Z.H., Cummins,J.M., Shen,D., Cahill,D.P., Jallepalli,P.V., Wang,T.L., Parsons,D.W., Traverso,G., Awad,M. *et al.* (2004) Three classes of genes mutated in colorectal cancers with chromosomal instability. *Cancer Res.*, **64**, 2998–3001.
37. Wood,R.D., Mitchell,M. and Lindahl,T. (2005) Human DNA repair genes. *Mutat. Res.*, **577**, 275–283.
38. Jiricny,J. (2006) The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.*, **7**, 335–346.

Capítulo 4: *Sobre a ausência de mutações em genes do sistema de reparo por excisão de nucleotídeos em neoplasias sólidas esporádicas*.

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

On the absence of mutations in nucleotide excision repair genes in sporadic solid tumors

J.C.M. Mombach[1,2], M.A.A. Castro[3,4], J.C.F. Moreira[4] and R.M.C. de Almeida[5]

[1]Centro de Ciências Rurais, UNIPAMPA, São Gabriel
[2]Universidade Federal de Santa Maria, Santa Maria, RS, Brasil
[3]Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil
[4]Universidade Luterana do Brasil, Gravataí, RS, Brasil
[5]Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

Corresponding author: R.M.C. de Almeida
E-mail: rita@if.ufrgs.br

Genet. Mol. Res. 7 (1): 152-160 (2008)
Received November 12, 2007
Accepted January 15, 2008
Published February 19, 2008

ABSTRACT. In general, stochastic tumors show genomic instability associated with the proliferation of DNA point mutations, that is, a mutator phenotype. This feature cannot be explained by a dysfunctional mismatch repair alone, and indicates that nucleotide excision repair (NER) and/or base excision repair should be suppressed. However, mutations in NER genes are not causally implicated in the oncogenesis of sporadic solid tumors, according to the Cancer Gene Census at http://www.sanger.ac.uk/genetics/CGP/Census/. This brings up an apparent paradox: how to explain the recurrent non-existence in NER genes of somatic mutations causally related to cancer? In a recent study, we have shown that the origin of point mutations in cancer cell genomes can be explained by a structurally conserved NER with a functional disorder generated from its

entanglement with a disabled apoptosis gene network. In the present study, we further characterize NER gene network properties and show that it has a highly connected architecture. This feature suggests that the absence of mutations in NER genes in sporadic solid tumors is a result of their participation in many essential cellular functions.

Key words: Nucleotide excision repair; Cancer; Gene network; Sporadic solid tumors

INTRODUCTION

Genome instability in solid tumors originates from either somatic mutations (observed in the majority of sporadic cancers) or germline mutations (associated with rare hereditary cancer syndromes) (Hoeijmakers, 2001). Germline mutations are present in all cells of an individual while somatic mutations occur in one or few cells. On the other hand, somatic mutations are favored by genomic instability, which increases the probability that a cell turns cancerous due to some insult such as, for example, exposure to a carcinogenic agent (Breivik and Gaudernack, 2004). We can then expect the genome instability per cell to be greater in the few mutated cells that may originate sporadic tumors than in the cells of individuals showing germline syndromes prone to cancer development. In fact, germline syndromes make an individual vulnerable to some specific cancer, characteristic of each syndrome: depending on the locus of the germline mutation of a given syndrome, different DNA repair mechanisms are affected, generating genome instability as a result of different kinds of insults (Chao and Lipkin, 2006).

Genomic stability is guaranteed by i) DNA repair mechanisms, ii) chromosome stability mechanisms (such as chromosome segregation control), and iii) as an extreme resource, apoptosis, which in fact plays the role of a tissue repair mechanism, since it guarantees tissue homeostasis (Hoeijmakers, 2001; Castro et al., 2006). Consequently, since cancer cells display genome instability, we can conclude that in neoplasms one or more of these mechanisms are dysfunctional.

Here, we consider genomic instability at the nucleotide level. There are three repair systems responsible for correcting DNA point damage: mismatch repair (MMR), base excision repair (BER) and nucleotide excision repair (NER). Germline mutations, associated with syndromes prone to cancer development, are found at genes related to these three systems, for example, adenomatosis polyploidy coli at genes related to MMR and BER and xeroderma pigmentosum at genes related to NER (Friedberg, 2003). As expected, these mutations yield a mutator phenotype. However, somatic mutations causally related to cancer are never found in NER or BER (Futreal et al., 2004), which could be thought as an alternative way to produce a mutator phenotype.

Nevertheless, stochastic tumors show, in general, genomic instability associated with the proliferation of DNA point mutations, that is, a mutator phenotype. As this feature cannot be explained by a dysfunctional MMR alone, it indicates that NER and/or BER should be suppressed. In fact, in a previous study, we have found that the activity and diversity profiles of the gene transcripts related to the NER pathway are altered in the majority of tumors listed in NIH/CGAP database (Wheeler et al., 2003), while BER is functionally intact. In Figure 1 we reproduce our previous finding, showing that NER and apoptosis are the most impaired pathways of all tested pathways involved in genome stability (Castro et al., 2007).

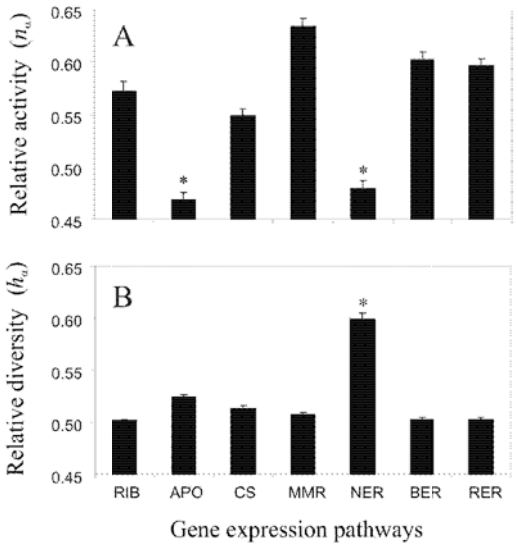


Figure 1. Analysis of gene expression of genome stability pathways. Pathways were evaluated according to two different expression features: gene activity level (i.e., number of transcripts in the pathway) and the distribution of transcripts among the set of genes that characterizes each pathway (i.e., diversity of transcripts). **A.** Relative activity n_α, as defined by Castro et al. (2007), expresses the gene activity in cancer cells in relation to normal cells. **B.** Relative diversity h_α, as defined by Castro et al. (2007), expresses the spread in the distribution of transcripts in cancer cells in relation to normal cells. The results were also compared with conserved pathways used as invariant expression controls (i.e., ribosome, ATP synthase, electron transport chain, and glycolysis). Here, we present the most conserved one: ribosome. RIB = ribosome; APO = apoptosis; CS = chromosome stability; MMR = mismatch repair; NER = nucleotide excision repair; BER = base excision repair; RER = recombinational repair. *Different from controls with $P < 0.001$.

As apoptosis and NER pathways are functionally entangled, it is possible that NER is dysfunctional due to apoptosis suppression. Another way would be through a somatic mutation. As somatic mutations in NER genes causally linked to cancer apparently do not exist, according to Cancer Gene Census (Futreal et al., 2004), how do we explain the recurrent non-existence in NER genes of somatic mutations causally related to cancer?

In this paper aimed at explaining the recurrent absence of somatic mutations causally linked to cancer, we compare the properties of the interaction network of NER genes with those of genes that may show these mutations.

Although cancer reduces an organism's life span, it enhances the fitness of the cancer cell. Consequently, stochastic mutations that reduce cell fitness are not likely to originate cancer. This implies that highly connected genes should not give rise to cancer when randomly mutated. Cell fitness should be taken here in a broad sense: not only should the mutated cell win the race for nutrients and show an enhanced proliferation rate to overcome a clonal selection, it must also evade internal tissue control mechanisms, such as apoptosis, that would remove transformed cells as compared to the original tissue.

In this study, we were interested in determining whether the recurrent absence in NER genes of somatic mutations causally linked to cancer is related to the fact that they are highly functionally connected. To verify this statement, we assessed the connectivity of NER genes and compared it to that of all genes whose somatic mutations have been identified as being causally implicated in cancer (Futreal et al., 2004). We emphasize that for all calculations, we considered

not only physical but also predicted functional interactions, as explained in the Material and Methods section. We found that the NER networks are much more connected than cancer genes. We propose that this feature may explain why stochastic mutations in NER genes, by affecting more neighbors in the network as compared to cancer genes, may reduce cell fitness.

MATERIAL AND METHODS

Data selection

The list of 79 human cancer genes used in this study was obtained from the Cancer Gene Census of the Sanger Institute. This census is an up-to-date catalog of genes causally implicated in cancer and is available at http://www.sanger.ac.uk/genetics/CGP/Census/. Here, we considered only somatically mutated genes described in epithelial tumors (i.e., we focused on sporadic solid tumors) and that are also annotated in STRING database (von Mering et al., 2007). The list of 28 genes representing the nucleotide-excision repair pathway follows the catalog of Human DNA repair genes (Wood et al., 2005) and is available at http://www.cgal.icnet.uk/DNA_Repair_Genes.html. The complete file showing Entry IDs of NER and cancer genes are listed in Table 1.

Analysis of protein/gene interaction networks

The protein-protein interaction networks and functional interactions associating genes are generated using the STRING database ('search tool for the retrieval of interacting genes/proteins') (von Mering et al., 2007) with input option "databases" (functional interactions), "experiments" (physical interactions), and 90% confidence level. STRING is a database of known and predicted protein-protein and functional interactions and it allows data mining for individual genes (e.g., a 'target gene'). In this way, all annotations in the database are searched concerning this target gene, which is the central node of the predicted network of interactions. In order to identify genes in the database, we used both HUGO ID (Wain et al., 2004) and Ensembl Peptide ID (Birney et al., 2006). Genes were also checked in NCBI Entrez (Wheeler et al., 2003) to link the approved gene name with respective STRING ID. Alternatively, the amino acid sequence of a given protein was supplied to identify the corresponding gene. The results from the search were saved in data files ('tab-delimited text fields') describing edge relationships. Graphs were built with the Medusa program (Hooper and Bork, 2005) and exported to postscript files to have figure quality improved and edited. Network parameters were obtained as proposed by Barabási and Oltvai (2004). Briefly, we calculated the connectivity for all genes specified in the Data selection section. The connectivity k_i of a given gene i is defined as the number of links that a given node has with other nodes. Genes that are not annotated in STRING database were not included in our analysis. We obtained a total of 107 protein-protein interaction networks (28 for NER and 79 for cancer genes). The complete file showing the network parameters is shown in Table 1.

Statistical analysis

For statistical comparison of the gene networks, we used the Mann-Whitney test. Tests were performed with the SPSS nonparametric statistical package (SPSS for Windows, release 14.0.0., SPSS Inc., Chicago, IL). Results are reported as means ± SEM with $P < 0.05$ for significance.

Table 1. Network parameters of nucleotide excision repair and cancer genes.

Network[a,b]	Gene symbol	Entrez ID	Ensembl ID	k_i[c]
CAN	AKAP9	10142	ENSP00000339128	10
CAN	AKT2	208	ENSP00000309428	19
CAN	APC	324	ENSP00000257430	11
CAN	ATF1	466	ENSP00000262053	3
CAN	BRAF	673	ENSP00000288602	10
CAN	BRCA1	672	ENSP00000337814	41
CAN	BRCA2	675	ENSP00000267071	5
CAN	BRD4	23476	ENSP00000263377	0
CAN	CCND1	595	ENSP00000227507	88
CAN	CDH1	999	ENSP00000261769	14
CAN	CDKN2A	1029	ENSP00000354638	12
CAN	COPEB	1316	ENSP00000173785	3
CAN	CTNNB1	1499	ENSP00000344456	69
CAN	CYLD	1540	ENSP00000308928	1
CAN	D10S170	8030	ENSP00000263102	0
CAN	EGFR	1956	ENSP00000275493	48
CAN	ELKS	23085	ENSP00000354158	2
CAN	EP300	2033	ENSP00000263253	65
CAN	ERBB2	2064	ENSP00000269571	14
CAN	ERG	2078	ENSP00000288319	0
CAN	ETV1	2115	ENSP00000242066	0
CAN	ETV6	2120	ENSP00000266427	0
CAN	FBXW7	55294	ENSP00000281708	4
CAN	FGFR2	2263	ENSP00000350166	8
CAN	FGFR3	2261	ENSP00000339824	10
CAN	GNAS	2778	ENSP00000302237	47
CAN	GOLGA5	9950	ENSP00000163416	1
CAN	HMGA1	3159	ENSP00000318322	1
CAN	HRAS	3265	ENSP00000309845	49
CAN	HRPT2	3279	ENSP00000256767	0
CAN	KRAS	3845	ENSP00000256078	16
CAN	KTN1	3895	ENSP00000348562	1
CAN	LIFR	3977	ENSP00000263409	6
CAN	MADH4	4089	ENSP00000341551	35
CAN	MAML2	84441	ENSP00000327563	1
CAN	MAP2K4	6416	ENSP00000262445	28
CAN	MECT1	94159	ENSP00000345001	0
CAN	MEN1	4221	ENSP00000337088	1
CAN	MET	4233	ENSP00000317272	31
CAN	MLH1	4292	ENSP00000231790	9
CAN	MSH2	4436	ENSP00000233146	8
CAN	MSH6	2956	ENSP00000234420	4
CAN	MYC	4609	ENSP00000259523	25
CAN	MYCL1	4610	ENSP00000335376	0
CAN	NCOA4	8031	ENSP00000351902	1
CAN	NONO	4841	ENSP00000298087	4
CAN	NRAS	4893	ENSP00000261444	15
CAN	NTRK1	4914	ENSP00000351486	11
CAN	NTRK3	4916	ENSP00000354207	1
CAN	NUT	256646	ENSP00000329448	0
CAN	PAX8	7849	ENSP00000263334	0
CAN	PCM1	5108	ENSP00000327077	0
CAN	PIK3CA	5290	ENSP00000263967	21
CAN	PLAG1	5324	ENSP00000325546	0
CAN	PPARG	5468	ENSP00000287820	9
CAN	PRCC	5546	ENSP00000271526	0

Continued on next page

Table 1. Continued.

Network[a,b]	Gene symbol	Entrez ID	Ensembl ID	k_i[c]
CAN	PRKAR1A	5573	ENSP00000226090	21
CAN	PRO1073	29005	ENSP00000332769	0
CAN	PTCH	5727	ENSP00000332353	6
CAN	PTEN	5728	ENSP00000304973	39
CAN	RB1	5925	ENSP00000267163	80
CAN	RET	5979	ENSP00000344798	8
CAN	SFPQ	6421	ENSP00000263532	5
CAN	SMO	6608	ENSP00000249373	2
CAN	STK11	6794	ENSP00000324856	7
CAN	TCEA1	6917	ENSP00000353558	56
CAN	TCF1	6927	ENSP00000257555	2
CAN	TFE3	7030	ENSP00000338360	2
CAN	TFEB	7942	ENSP00000351685	2
CAN	TFG	10342	ENSP00000240851	0
CAN	TMPRSS2	7113	ENSP00000330330	0
CAN	TNFRSF6	355	ENSP00000347979	16
CAN	TP53	7157	ENSP00000269305	93
CAN	TPM3	7170	ENSP00000271850	2
CAN	TPR	7175	ENSP00000264142	25
CAN	TRIM33	51592	ENSP00000351250	0
CAN	TSHR	7253	ENSP00000298171	4
CAN	VHL	7428	ENSP00000256474	11
CAN	ZNF331	55422	ENSP00000345514	0
NER	CCNH	902	ENSP00000256897	98
NER	CDK7	1022	ENSP00000256443	98
NER	CETN2	1069	ENSP00000276365	2
NER	DDB1	1642	ENSP00000301764	20
NER	DDB2	1643	ENSP00000256996	19
NER	ERCC1	2067	ENSP00000013807	46
NER	ERCC2	2068	ENSP00000221481	85
NER	ERCC3	2071	ENSP00000285398	87
NER	ERCC4	2072	ENSP00000310520	46
NER	ERCC8	1161	ENSP00000265038	41
NER	GTF2H1	2965	ENSP00000265963	85
NER	GTF2H2	2966	ENSP00000274400	85
NER	GTF2H3	2967	ENSP00000228955	84
NER	GTF2H4	2968	ENSP00000259895	85
NER	GTF2H5	404672	ENSP00000333219	7
NER	LIG1	3978	ENSP00000263274	28
NER	MMS19L	64210	ENSP00000307263	0
NER	MNAT1	4331	ENSP00000261245	97
NER	RAD23A	5886	ENSP00000321365	1
NER	RAD23B	5887	ENSP00000350708	22
NER	RPA1	6117	ENSP00000254719	70
NER	RPA2	6118	ENSP00000263698	69
NER	RPA3	6119	ENSP00000223129	66
NER	XAB2	56949	ENSP00000351137	42
NER	XPA	7507	ENSP00000259463	19
NER	XPC	7508	ENSP00000285021	19

[a]Nucleotide excision repair (NER) genes according to Wood (2005). The gene list was checked in HGNC Database (Wain et al., 2004) and NCBI Entrez (Wheeler et al., 2003) to link the approved gene name with respective STRING ID.
[b]List of somatically mutated genes (CAN) causally implicated in cancer, described for epithelial tissue in Cancer Gene Census (Futreal et al., 2004), and also annotated in STRING database (von Mering et al., 2007). The gene list was obtained from the Wellcome Trust Sanger Institute Cancer Genome Project web site, http://www.sanger.ac.uk/genetics/CGP. Genes were checked in HGNC Database (Wain et al., 2004) and NCBI Entrez (Wheeler et al., 2003) to link the approved gene name with respective STRING ID.
[c]The connectivity k_i of each gene i was estimated as described in Material and Methods.

RESULTS AND DISCUSSION

Four graphs of protein-protein interactions exemplifying gene network analysis are presented in Figure 2. The central node corresponds to the target gene for which the search is made. The predicted associations start from this central node and include all associations found in the STRING database (von Mering et al., 2007). Figure 2A and B represent the networks of NER target genes, while Figure 2C and D are two examples of networks generated from cancer genes as targets. It is clear that NER genes are more connected. Note that these networks show all genes connected with the target genes, including genes belonging to other functions or pathways. We generated 107 such networks (28 for NER and 79 for cancer genes), one for each target gene.

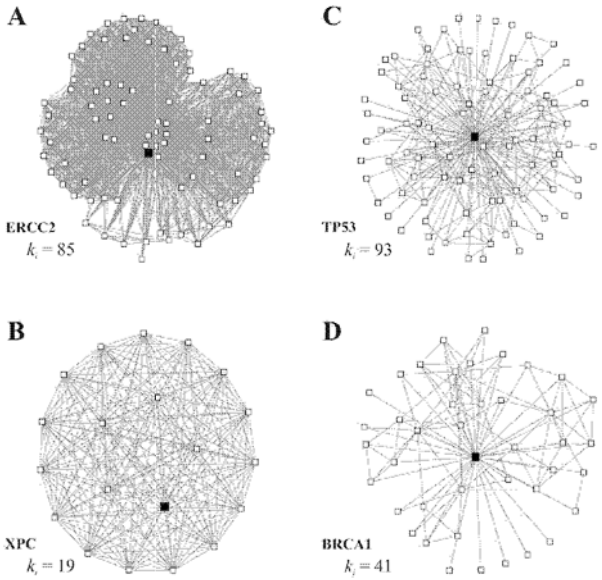


Figure 2. Graph of interactions illustrating nucleotide excision repair (NER) and cancer gene networks. The predicted associations are generated using database STRING (von Mering et al., 2007) with input options "Experimental/Biochemical Data", "Association in Curated Databases" and 90% confidence level. **A** and **B.** ERCC2 and XPC genes, illustrating NER gene networks. **C** and **D.** TP53 and BRCA1 genes, illustrating somatically mutated gene networks. Black node represents the central gene for which the network is built (i.e., 'target gene'). White nodes represent all genes connected to the central node. The connectivity k_i of the central node is indicated. Network parameters are listed in Table 1 for all NER genes and all cancer genes for which somatic mutations have been causally implicated in epithelial tumors (i.e., sporadic solid tumors).

The connectivity calculated for all NER and cancer genes is presented in Figure 3A and is provided in Table 1. As one can observe, the great majority of NER genes have connectivity higher than do cancer genes. These latter networks are significantly more sparse (Figure 3B; $P < 0.05$). In fact, this feature has been partially observed in a previous study for a much smaller set of cancer genes (17 genes), those restricted to genome stability functions (Castro et al., 2007). Here, we extended the comparison to all 79 cancer genes associated with sporadic solid tumors found in the Cancer Genes Census (Futreal et al., 2004), considering the 79 networks produced by STRING database for each cancer gene as a target gene. In Figure 4, we present the interaction network, considering only NER genes. As we can see, the network is highly connected, which confirms a cooperative functioning (Christensen et al., 2007).

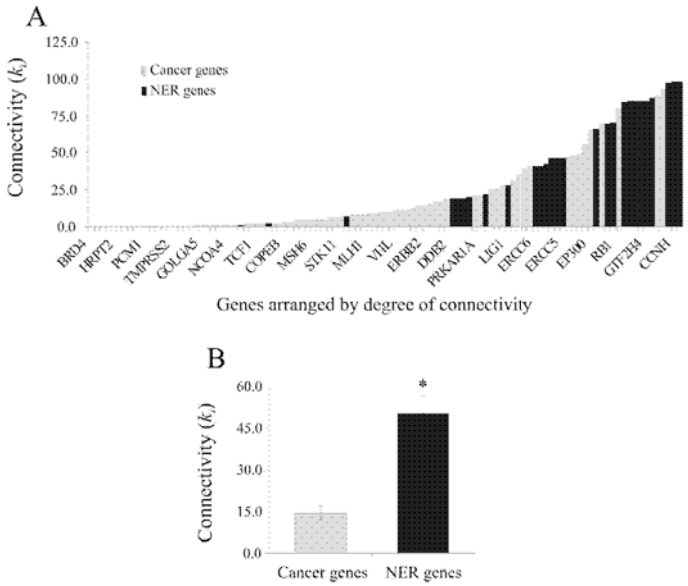


Figure 3. Contrasting nucleotide excision repair (NER) and cancer gene networks. Analysis of protein-protein interaction networks of NER and cancer genes for which somatic mutations have been causally implicated in epithelial tumors. **A.** Connectivity k_i of a given gene i, which shows the number of links that a given node has with other nodes as defined in Material and Methods. Genes are ranked by degree of connectivity. NER genes are indicated in black columns. **B.** Average connectivity of NER and cancer genes. Asterisk indicates significant difference between groups, with $P < 0.05$.

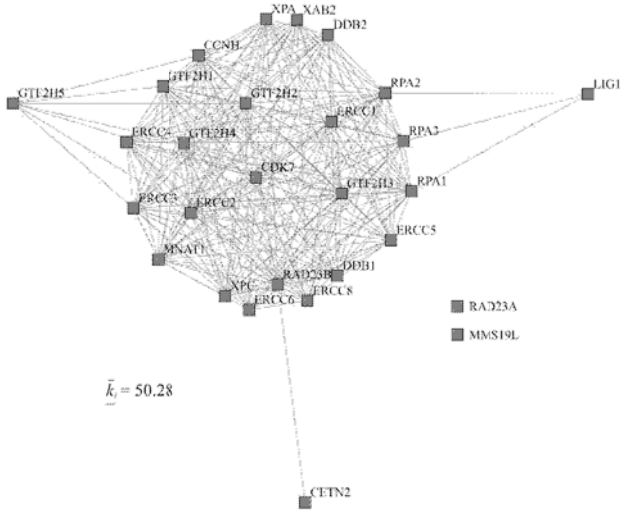


Figure 4. Graph of the nucleotide excision repair (NER) gene network built using software MEDUSA (Hooper and Bork, 2005). The network is highly connected. The network shows the known and predicted protein-protein and functional interactions from the database STRING (von Mering et al., 2007) with input options "databases", "experiments" and 90% confidence level. Nodes without predicted associations at this level are indicated. Average connectivity k_i of NER genes is indicated (considering all interactions of NER genes - including external interactions). The list of NER gene identifiers (i.e., Entry IDs) is provided as a supplementary spreadsheet (Table 1).

We can explain these multiple interactions of NER genes by their participation in a variety of cellular processes. Some of the NER gene products are pre-associated tightly in complexes as heteromultimers. These include the 10 subunits of TFIIH, the ERCC1-XPF complex, the XPC-RAD23B complex, and the heterotrimeric RPA1-RPA2-RPA3 complex. Additionally, some known biochemical facts about NER gene products are unusual in comparison to other DNA repair pathways. Many NER genes have essential functions in the cell other than DNA repair. This is true for all 10 subunits of TFIIH (as an essential factor for RNA transcription), for ERCC1 and XPF (which also act in a recombination pathway), for XPG (which works with TFIIH as a transcription activator), for the RPA subunits (which form the major single-stranded binding protein, also necessary for semi-conservative DNA replication), and for LIG1, and XAB2.

An additional support to our proposal comes from experiments of gene silencing. A recent study has shown that silencing of some NER genes by means of RNA interference in cultures of HeLa and MCF-7 cells causes major growth disadvantages (Biard, 2007). Consequently, the contrast among NER and cancer genes in regard to their properties in the network suggests that, in general, random point mutations directly in NER genes may reduce cell fitness in sporadic solid tumors, explaining the absence of somatic mutations in these genes.

ACKNOWLEDGMENTS

Research partially supported by the Brazilian agencies FAPERGS, CNPq, and CAPES. We acknowledge STRING database and Wellcome Trust Sanger Institute Cancer Genome Project for providing public access to their data. We thank Marialva Sinigaglia for useful discussions.

REFERENCES

Barabási AL and Oltvai ZN (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101-113.

Biard DS (2007). Untangling the relationships between DNA repair pathways by silencing more than 20 DNA repair genes in human stable clones. *Nucleic Acids Res.* 35: 3535-3550.

Birney E, Andrews D, Caccamo M, Chen Y, et al. (2006). Ensembl 2006. *Nucleic Acids Res.* 34: D556-D561.

Breivik J and Gaudernack G (2004). Resolving the evolutionary paradox of genetic instability: a cost-benefit analysis of DNA repair in changing environments. *FEBS Lett.* 563: 7-12.

Castro MA, Onsten TG, Moreira JC and de Almeida RM (2006). Chromosome aberrations in solid tumors have a stochastic nature. *Mutat. Res.* 600: 150-164.

Castro MA, Mombach JC, de Almeida RM and Moreira JC (2007). Impaired expression of NER gene network in sporadic solid tumors. *Nucleic Acids Res.* 35: 1859-1867.

Chao EC and Lipkin SM (2006). Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. *Nucleic Acids Res.* 34: 840-852.

Christensen C, Thakar J and Albert R (2007). Systems - level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET. Syst. Biol.* 1: 61-77.

Friedberg EC (2003). DNA damage and repair. *Nature* 421: 436-440.

Futreal PA, Coin L, Marshall M, Down T, et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4: 177-183.

Hoeijmakers JH (2001). Genome maintenance mechanisms for preventing cancer. *Nature* 411: 366-374.

Hooper SD and Bork P (2005). Medusa: a simple tool for interaction graph analysis. *Bioinformatics* 21: 4432-4433.

von Mering C, Jensen LJ, Kuhn M, Chaffron S, et al. (2007). STRING 7 - recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* 35: D358-D362.

Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, et al. (2004). Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.* 32: D255-D257.

Wheeler DL, Church DM, Federhen S, Lash AE, et al. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31: 28-33.

Wood RD, Mitchell M and Lindahl T (2005). Human DNA repair genes, 2005. *Mutat. Res.* 577: 275-283.

Capítulo 5: *ViaComplex: ferramenta de bioinformática para análise de redes de expressão gênica em contexto genômico*.

Artigo aceito para publicação - Bioinformatics

Systems biology

ViaComplex: software for landscape analysis of gene expression networks in genomic context

Mauro A. A. Castro[1,*], José L. Rybarczyk Filho[2], Rodrigo J. S. Dalmolin[1], Marialva Sinigaglia[3], José C. F. Moreira[1], José C. M. Mombach[4] & Rita M. C. de Almeida[2]

[1]Unidade de Bioinformática, Departamento de Bioquímica, [2]Instituto de Física and [3]Núcleo de Bioinformática, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003, Brazil. [4]Departamento de Física, Universidade Federal de Santa Maria, Santa Maria 97105-900, Brazil.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: ViaComplex is an open-source application that builds landscape maps of gene expression networks. The motivation for this software comes from two previous publications (Nucleic Acids Res, 35(6):1859-67, 2007; Nucleic Acids Res, 36(19):6269-83, 2008). The first paper presents a network-based model of genome stability pathways where we defined a set of genes that characterizes each genetic system. In the second paper we analyzed this model by projecting functional information from several experiments onto the gene network topology. In order to systematize the methods developed in these papers, ViaComplex provides tools that may help potential users to assess different high-throughput experiments in the context of six core genome maintenance mechanisms. This model illustrates how different gene networks can be analyzed by the same algorithm.

Availability: http://lief.if.ufrgs.br/pub/biosoftwares/viacomplex.

Contact: M.A.A.C. (mauro@ufrgs.br) or R.M.C.A. (rita@if.ufrgs.br).

1 INTRODUCTION

Genome maintenance mechanisms (GMM) are critical for cell homeostasis. Evolution has shaped sophisticated repair systems that cover most of the insults that can cause genome damages. Defects in any of these systems can predispose to cancer (Castro et al., 2008). At least four DNA damage repair pathways operate in mammals that, together with apoptosis and chromosome stability pathways, comprise the basis of GMM (Zhivotovsky and Kroemer, 2004; Hoeijmakers, 2001).

We have previously constructed a network-based model of human GMM (Castro et al., 2007) in which different gene activity data were projected onto the interaction map (Castro et al., 2008). Here we extend these previous studies and develop a new software which could serve as a generalized tool to evaluate gene expression networks. With a graphical user interface, ViaComplex can either compute gene activity data for the internal model, or import customized models of gene/protein interaction networks. In this case, the GMM network model illustrates the type of problem that can be deal with the software for different gene networks. It can be used to produce publication quality images where data are visualized as functional landscapes projected onto gene network maps. ViaComplex also provides a statistical module based on the concept of information theory where multiple hypotheses are controlled by the false discovery rate (FDR) approach.

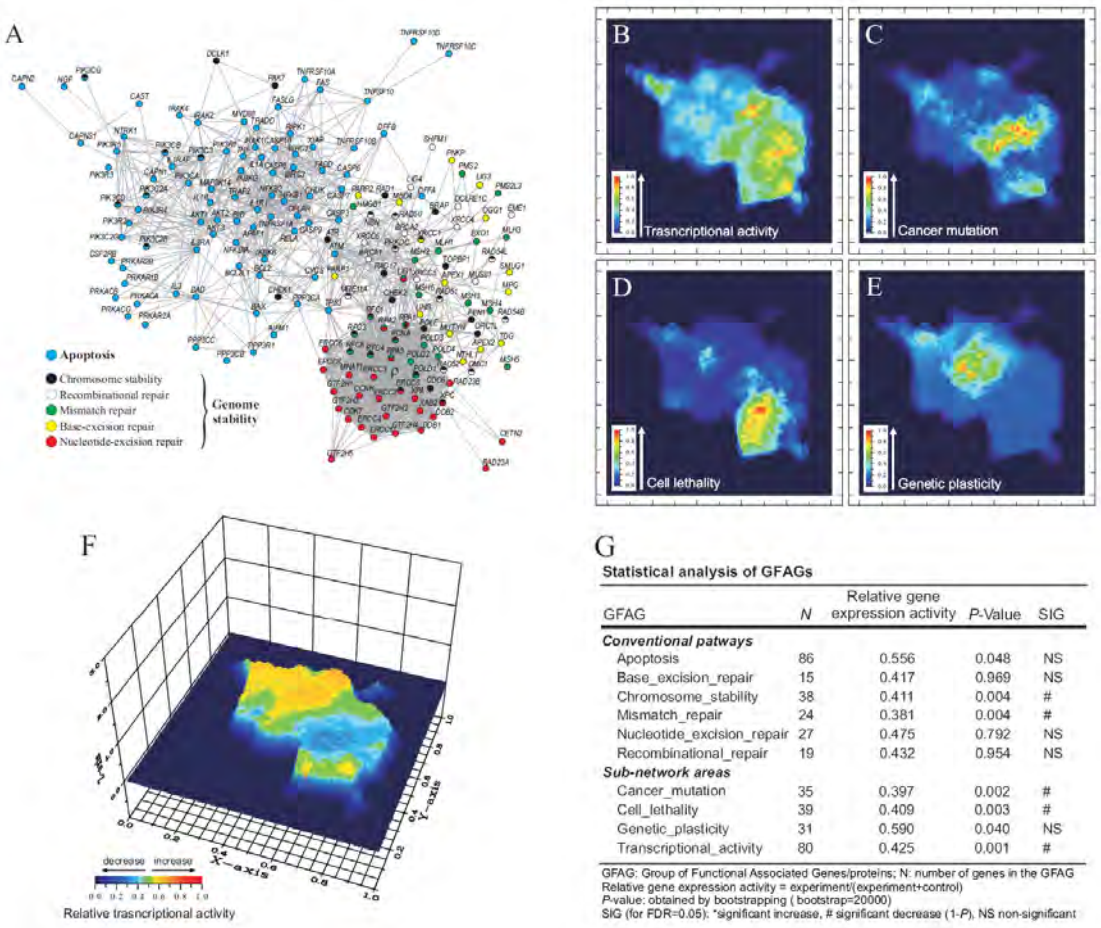
2 IMPLEMENTATION

ViaComplex program code is written for Linux and Windows Intel FORTRAN compilers (version 10.1.025) and is linked with Dislin 9.4, a scientific plotting library (http://www.dislin.de/). The main advantage of this program is that it is able to distribute a given quantity (quantitative or qualitative data) onto gene/protein interaction networks. To do this, ViaComplex overlaps functional information with interaction information (*e.g.* the network-based model of GMM).

The GMM network model comprises 180 genes that participate in human apoptosis and genome-stability functions as previously described (Castro et al., 2007) and is depicted in Figure 1A. As an example of ViaComplex capabilities, in Figure 1B we show a microarray data analysis processed by the landscape module where gene expression activity is plotted over the network topology. In this figure the software distributed the microarray signal according to the coordinates of the network objects (*i.e.* nodes and links). By default, ViaComplex will distribute the signal on both nodes and links, but the user can change this option together with other ones available in the console. Also, the same algorithm can map qualitative data, as exemplified with cancer mutations (Fig. 1C), cell lethality (Fig. 1D) and genetic plasticity (Fig. 1E). Alternatively, user can compare two different functional states of the same gene network topology (Fig. 1F, methylated *vs.* nonmethylated states).

The install package includes a comprehensive help file that provides the user all necessary details to prepare the data input and to execute the data analysis. ViaComplex can read the common gene/protein identifiers (e.g. EMBL, ENTREZ, UniProt, HGNC, RefSeq, and UniGene) and the resulting graphs can be previewed as XWIN format or saved as PDF, EPS or PostScript files.

*To whom correspondence should be addressed.



GFAG	*N*	Relative gene expression activity	*P*-Value	SIG
Conventional patways				
Apoptosis	86	0.556	0.048	NS
Base_excision_repair	15	0.417	0.969	NS
Chromosome_stability	38	0.411	0.004	#
Mismatch_repair	24	0.381	0.004	#
Nucleotide_excision_repair	27	0.475	0.792	NS
Recombinational_repair	19	0.432	0.954	NS
Sub-network areas				
Cancer_mutation	35	0.397	0.002	#
Cell_lethality	39	0.409	0.003	#
Genetic_plasticity	31	0.590	0.040	NS
Transcriptional_activity	80	0.425	0.001	#

GFAG: Group of Functional Associated Genes/proteins; N: number of genes in the GFAG
Relative gene expression activity = experiment/(experiment+control)
P-value: obtained by bootstrapping (bootstrap=20000)
SIG (for FDR=0.05): *significant increase, # significant decrease (1-*P*), NS non-significant

Fig. 1. Landscape analysis of Genome Maintenance Mechanisms (GMM). **(A)** Graph of interactions among genes involved GMM, as previously described (Castro et al., 2007) **(B)** Example of gene expression data analysis using breast MCF7 cell line (GEO accession n. GSM155194). Color gradient represents the transcriptional activity mapped onto graph. The same algorithm can be used to map different data, *e.g.*, **(C)** genes causally implicated in human cancer, **(D)** yeast lethality data, and **(E)** genetic plasticity, as defined in Castro et al. (2008). **(F)** Two-state landscape analysis. It compares the transcription profile of MCF7 cells in hypomethylated state (*a*) *vs.* hipermethylated state (*b*) (GEO accessions GSE5816 *vs.* GSM155194), where the color gradient $Z=a/(a+b)$. A summary of the statistical analysis of this data is presented in **(G)**. Fig 1A reprinted/adapted with permission from NAR/Oxford University Press.

The install package also comprises an extensive library of published studies that exemplifies all procedures by a simple mouse click. In this sense, the GMM network model can be used to observe the functionality of the algorithm, which can analyze different gene networks. **Supplementary Figure 1** illustrates this possibility for a large network (with 1892 genes). Such option is available at the "custom model" module. It is semantically focused in genes, matching gene IDs and node IDs. If there is more than one microarray probe interrogating the expression of a given gene, then the software takes the average of the expression values, which allows the use of different microarray platforms (*i.e.* it does not involve comparisons between network nodes and probe tag IDs). Other numerical samples of different sizes are available at ViaComplex home page.

Additional features of the software include a statistical module where two microarray datasets can be compared following the method described at Castro et al., (2007), as exemplified in Figure 1G for the data used to build the Figure 1F. We anticipate that ViaComplex will be useful to mine graph patterns from high-throughput experimental data.

ACKNOWLEDGEMENTS

Funding: Brazilian Agencies FAPERGS, CAPES and CNPq.

Conflict of Interest: none declared.

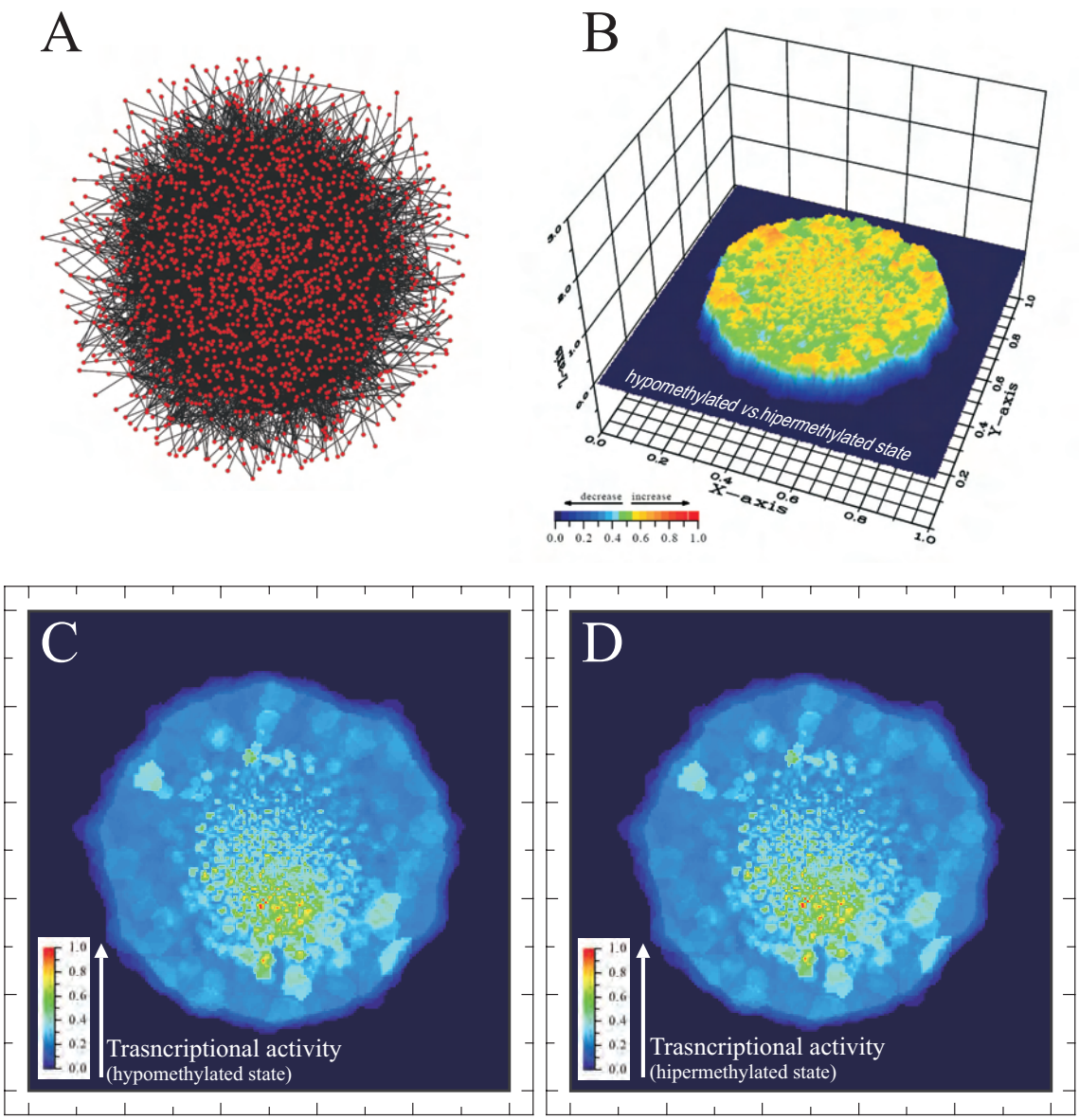
REFERENCES

Castro,M.A.A. et al. (2007) Impaired expression of NER gene network in sporadic solid tumors. *Nucl. Acids Res*., **35**, 1859-1867.

Castro,M.A.A. et al. (2008) Evolutionary origins of human apoptosis and genome-stability gene networks. *Nucl. Acids Res*., **36**, 6269-6283.

Hoeijmakers,J.H.J. (2001) Genome maintenance mechanisms for preventing cancer. *Nature*, **411**, 366-374.

Zhivotovsky,B. and Kroemer,G. (2004) Apoptosis and genomic instability. *Nat. Rev. Mol. Cell Biol*., **5**, 752-762.



Supplementary Figure 1. Example of landscape analysis using large network models. **(A)** Graph of interactions among 1892 human genes. Genes were selected by chance and then their interactions (links) were retrieved from STRING database (http://string.embl.de/). **(B)** This example illustrates the analysis of the same gene expression data presented in the paper (Fig.1) where MCF7 cell line transcriptome is plotted onto GMM network model. Here, this data is plotted onto a large network using the two-state landscape analysis. It compares the transcription profile of MCF7 cells in hypomethylated state (a) *vs.* hipermethylated state (b) (GEO accessions GSE5816 vs. GSM155194), where the color gradient Z=a/(a+b). **(C)** Landscape analysis of gene expression data from MCF7 cell line in hypomethylated state. **(D)** Landscape analysis of gene expression data from MCF7 cell line in hipermethylated state

Capítulo 6: *Estudo da expressão gênica do biomarcador CFL1 para o prognóstico de câncer de pulmão de células não pequenas.*

Artigo em processo de submissão

CFL1 expression levels as a prognostic and drug resistance marker in non-small cell lung cancer

Mauro Antônio Alves Castro[1], Felipe Dal-Pizzol[2], Márcio Soares[3], Caroline Beatriz Müller[1], Fernanda Lopes[1], Alfeu Zanotto-Filho[1], Rosalva Meurer[4], Marilda Cruz Fernandes[4], Stéphanie Zdanov[5], José Cláudio Fonseca Moreira[1], Emily Shacter[5] and Fábio Klamt[1]**.

[1]Department of Biochemistry, ICBS/UFRGS, Porto Alegre/RS, 90035-003 Brazil; [2]Laboratory of Experimental Physiopathology, UNESC, Criciúma/SC, 88806-000 Brazil; [3]Intensive Care Unit, National Cancer Institute, Rio de Janeiro/RJ, 20230-130 Brazil; [4]Laboratory of Pathology Research, UFCSPA, Porto Alegre/RS, 90050-170 Brazil; [5]Division of Therapeutic Proteins, Center for Biological Evaluation and Research, Food and Drug Administration, Bethesda/MD, 20892-4555 USA.

**To whom correspondence should be addressed: Prof. Fábio Klamt, Department of Biochemistry, ICBS/Federal University of Rio Grande do Sul (UFRGS), 2600 Ramiro Barcelos St. Porto Alegre, RS 90035-003, Brazil. Phone: +55 51 3308-5577; Fax: +55 51 3308-5535; e-mail: 00025267@ufrgs.br

Non-small-cell lung cancer is the major determinant of cancer mortality worldwide. Clustering data from biopsies according to *CFL1* gene expression (main member of invasion/metastasis pathways) and outcome, revealed that patients with high *CFL1* levels have lower overall survival. *CFL1* mRNA and protein correlate with invasiveness and resistance against alkylating drugs *in vitro*. Our analysis indicates that cofilin can be used as a NSCLC biomarker with high sensitivity and specificity.

Lung cancer accounts for 1.3 million deaths annually[1] of which 85% are of non-small-cell lung cancer (NSCLC) patients[2]. These patients in average survive for 10 months and only 15% for five years[3]. Currently, prognosis of NSCLC patients is done by considering patient performance status and tumour staging. However, accumulating data[4,5] have shown that these have unsatisfactory power in predicting patient outcome or in guiding physicians on the best course of action for each patient. A novel prognostic method for early-stage NSCLC patients can potentially increase survival rates by indicating those in need of more aggressive treatment.

Lung cancers in particular show poor prognosis because of their ability to generate early metastasis within the lungs and then in distant organs. This behaviour requires great cell motility, which is performed by several proteins that act on the actin cytoskeleton by regulating cycles of polymerization and depolymerization of actin filaments, which in turn generates cell motion.

One of the main proteins in charge of cell motility is cofilin (*CFL1*, cofilin-1; non-muscle isoform; Gene ID: 1072)[6], which is regulated by phosphorylation, pH, binding of phosphoinositides, and subcellular compartmentalization. Its activation

occurs locally in response to EGFR signalling in chemotaxis[7]. The role of the cofilin pathway in cell mobility has been shown extensively and its activity correlated with breast cancer invasion and metastasis[8]. Thus, we raised the hypothesis that cofilin amount in NSCLC could provide relevant information about tumour aggressiveness and therefore be used as a prognostic marker.

Herein, we assessed the potential prognostic value of *CFL1* as an NSCLC biomarker. We used a large, homogeneous, well-defined collection of samples from lung cancer biopsies from a cohort[9] of 111 patients (**Fig. 1**) (cohort description can be found in **Supplementary Methods** and **Supplementary Table 1**). Data on microarray gene expression and patient information such as age, sex, cancer histological type, and NSCLC staging were considered.

Kaplan-Meier estimates of patient cumulative survival by time (months) according to the expression level of *CFL1* showed that when grouped by *CFL1* gene expression (upper fifth *vs.* lower fifth of transcript abundance levels), the expression levels discriminate patients in early disease stages (IA, IB, IIA, and IIB stages) between good and bad outcome (**Fig. 1a**). Cox multivariate regression revealed that lower *CFL1* expression was significantly associated with a high overall survival (hazard ratio for high risk *vs.* low risk, 2.7; 95% C.I., 1.5 to 4.7, P = 0.001) (**Fig. 1a and 1b**).

Analysis of 85 patients with disease stages I or II (the testing cohort), revealed that patients with high *CFL1* expression (n=42) had an overall survival rate shorter than those with low *CFL1* expression (n=43) (**Fig. 1c**). To test the robustness of this finding, we analyzed a second, independent data set of 67 patients (the validation cohort[10]) (**Fig. 1d)** and found that high *CFL1* levels are associated with shorter overall survival in both cohorts. ROC curve analysis showed that *CFL1*

sensitivity/specificity is high enough to indicate outcome in patients with early disease stages (area under ROC curve = 0.787) (**Fig. 1e**). Immunohistochemical stains revealed an increased cofilin immunocontent within the neoplasic tissue (**Fig. 1f**). The information in Figures **1a** to **1f** suggests that *CFL1* levels can be used to indicate patient outcome.

We also asked whether *CFL1* levels could provide additional insights into the pathophysiology of NSCLC by predicting tumour aggressiveness and/or chemotherapy response. We then used data from the US National Cancer Institute (NCI) 60 human tumour cell line anticancer drug screen (NCI60)[11,12]. Six human NSCLC cell lines of the three major histological types of NSCLC, namely adenocarcinomas cells (H-23, A549, EKVX), squamous-cells carcinomas (H-226), and large-cells carcinomas (H-460, HOP-92) were analyzed. We found that the relative levels of *CFL1* gene expression obtained by microarray varied among cell types (**Fig. 2a**; symbols) and matched our data on the amount of cofilin immunocontent (**Fig. 2a**; bars). Using the BD BioCoat™ Matrigel™ Invasion Chamber to assess the tumour's metastatic potential, we found that different histological types expressing higher *CFL1* levels presented higher invasion indexes, which indicated a more aggressive invasiveness behaviour (**Fig. 2b**) (*P < 0.02, Mann Whitney test; **P < 0.0001, One-way ANOVA).

Analysis of microarray data of the six cell lines and of the GI_{50} values of 118 standard chemotherapy agents (from the NCI-60 drug discovery pipeline) revealed that high levels of *CFL1* correlated with resistance against different anticancer drugs but mainly alkylating agents (**Fig. 2c**) (for a list of all correlated alkylating drugs see **supplementary table 2**). Exposure of the cell lines to different concentrations of five widely used chemotherapy drugs, namely cisplatin, carboplatin, 5-fluorouracil,

hydroxyurea, and taxol, revealed significant correlations between *CFL1* levels and resistance to cisplatin and carboplatin, the two alkylating agents tested (**Fig. 2d).** This finding may have great impact on survival rates, as currently there is no way to identify potential responders. Curiously, cisplatin-based chemotherapy has long been the cornerstone of NSCLC management[13,14] and even though the treatment improves patient survival, the benefit is stage-dependent[15]. Additional studies are needed to better elucidate the correlation between cisplatin-based chemotherapy and *CFL1* levels. We also built a network-based model for the cofilin biological pathway (**Fig. 2e**) and evaluated the resistance profile of each gene that interacts directly with *CFL1* (based on data at http://string.embl.de/) against alkylating agents. Four cofilin partners (*CAP1, ACTB, SSH3, YWHAZ* gene products) showed a resistance profile similar to cofilin's, suggesting that a functional network is responsible for this tumour phenotype (**Fig. 2e,** red nodes).

Through a systematic MEDLINE literature inspection we tested (as in Fig 1a) the clinical relevance of 60 genes suggested by other studies to be used as prognostic biomarkers in NSCLC patients. These biomarkers constitute transcription factors, protein kinases and phosphatases, receptors, and several DNA repair systems (See **supplementary table 3** for complete list). Nevertheless, our statistical analyses show that all potential biomarkers have no prognostic value for NSCLC patient outcomes.

Our analysis suggests that *CFL1* levels can help establish which early-stage patients should receive a more aggressive therapy as an attempt to reverse the poor prognosis. This could lead to a better management of NSCLC and decrease its recurrence. Prospective, large-scale, multicenter studies are necessary to test this idea.

Acknowledgments

We would like to thank Dr Marcia Triunfol at Publicase for editing the article and for manuscript comments. The research was supported in part by the Brazilian MCT/CNPq Universal funds (479860/2006-8) and by PROPESQ/UFRGS. This project was funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract NO1-CO-12400.

References

1. Jemal, A. *et al. CA Cancer J. Clin.* **58**, 71–96 (2008).
2. Potti, A. *et al. N. Engl. J. Med.* **355**, 570–580 (2006).
3. Sawyers, C.L. *Nature* **452**, 548–552 (2008).
4. Hsuan-Yu, C. *et al. N. Engl. J. Med.* **356**, 11–20 (2007).
5. Kaminski, N. & Krupsky, M. *Chest.* **125**, 111–115 (2004).
6. Gosh, M. *et al. Science*. **304**, 743–746 (2004).
7. van Rheenen, J. *et al. J. Cell Biol.* **179**,1247–1259 (2007).
8. Wang, W., Eddy, R., & Condeelis, J. *Nat. Rev. Cancer.* **7**, 429–440 (2007).
9. Bild, A.H. *et al. Nature* **439**, 353–357 (2005).
10. Beer, D.G. *et al. Nat. Med*. **8**, 816–824 (2002).
11. Scherf, U. *et al. Nature Genet.* **24**, 236–244 (2000).
12. Shoemaker, R.H. *Nat. Rev. Cancer* **6**, 813–123 (2006).
13. Arriagada, R. *et al. N. Engl. J. Med.* **350**, 351–360 (2004).
14. Imaizumi, M. *Lung Cancer* **49**, 85–94 (2005).
15. Pignon, J.P. *et al. J. Clin. Oncol.* **26**, 3552–3559 (2008).

Figure Legends

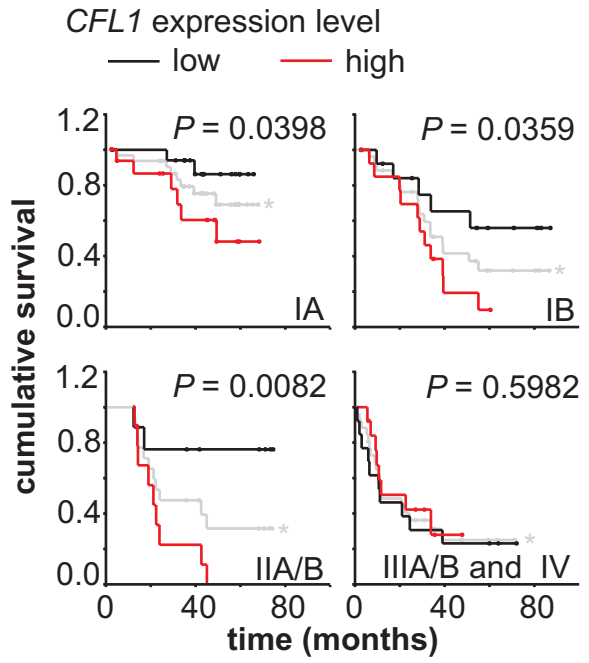
Figure 1 Prognostic value of *CFL1* levels in NSCLC patients. (**a**) Cohort data are grouped according to the International Staging System for Lung Cancer and *CFL1* gene expression level (*i.e.* upper-fifth *vs.* lower-fifth), and plotted as survival probabilities using the Kaplan-Meier method. Black lines represent patients with low *CFL1* expression; red lines with high *CFL1* expression. Differences in survival rates were assessed with the log-rank test. Gray lines represent all patients according to tumour staging. *P* values lower than 0.05 were considered significant. (**b**) Cox multivariable regression analysis to estimate hazard ratios for cohort clinical covariates and *CFL1* expression. Hazard ratios indicate that patients with high *CFL1* expression level presented poor outcome. (**c**) Kaplan Meier plots are shown for patients in stages I and II (*n*=85) in the original cohort (testing cohort) stratified by *CFL1* expression level and (**d**) in an independent cohort (validation cohort) obtained from a different set of published NSCLC microarray data (*n*=67). (**e**) Biomarker performance estimated by Receiver Operating Characteristic (ROC) analysis. (**f)** Representative immunohistochemical (IHC) analysis of *CFL1* gene product in tumour biopsies. Healthy human alveolar tissue obtained from tumour margins is mostly negative to cofilin IHC staining (upper left). High staining for cofilin is found within the neoplasic lung cells (asterisks). Original magnification x200; scale bar = 100 μM.

Figure 2 *in vitro CFL1* levels correlate with tumour invasiveness and resistance against alkylating drugs. Six human NSCLC cell lines composed of adenocarcinomas (H-23, A549, EKVX), large-cell carcinomas (H-460, HOP-92) and squamous-cells carcinomas (H-226) from the NCI-60 panel were selected based on different levels of *CFL1* gene expression

(http://discover.nci.nih.gov/datasetsNature2000.jsp) to establish the role of *CFL1* in tumour aggressiveness, evaluated by assays of cell invasion and drug resistance. (**a**) Western blot analysis shows that the pattern of *CFL1* expression (symbols) matches with the level of cofilin immunocontent (bars). (**b**) *CFL1* levels correlate with tumour invasiveness. Invasion index is obtained by determining the movement of cells through an 8.0 μm pore size, either uncoated (migration) or matrigel-coated (invasion), attracted by a chemotactic gradient of serum. The mean of four fields for each condition in quadruplicates is plotted. *P < 0.02 (Mann Whitney test); **P < 0.0001 (One-way ANOVA). (**c**) *CFL1* expression vs. drug sensitivity/resistance profile. Microarray data of the cell lines are crossed against GI_{50} values of 118 standard chemotherapy agents (from NCI-60 drug discovery pipeline). *P* values have been color coded according to the scale shown; P < 0.05 indicates a significant negative correlation (resistance) while P > 0.95 indicates a significant positive correlation (sensitivity). The major mechanism of drug action is shown (the term “alkylating agents” is used broadly to include platinating agents; Uk: unknown; P90: hsp90 binder; Pi: protein synthesis inhibitor). Each column within the matrix represents the Spearman correlation between gene expression and toxicity of an individual drug. (**d**) De novo validation of the citotoxicity for selected drugs assayed by the sulforhodamine B (SRB) method (upper plots). The obtained drug GI50 values were correlated with cofilin immunocontent (lower plots). (**e**) Biological network of *CFL1* partners *vs*. alkylating drug sensitivity/resistance profile. Nodes represent gene products; connecting lines indicate physical and/or functional associations according to experimental data (http://string.embl.de/). Gene expression data (http://discover.nci.nih.gov/cellminer/home.do) were crossed against GI_{50} values of all alkylating agents identified in the resistance panel at Fig. 2c. Four *CFL1* partners

follow the same resistance profile (red nodes; *n* = number of drugs for which gene expression showed correlation). The network drawn was built using a spring model algorithm. Further details in **Supplementary Methods**.

a



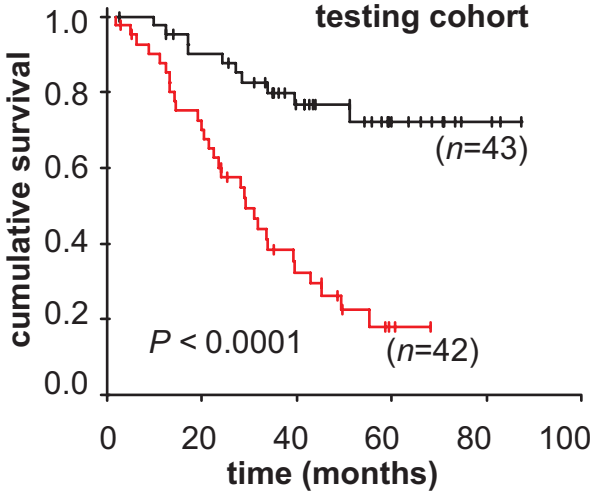
b

Hazard Ratio for patient's survival*

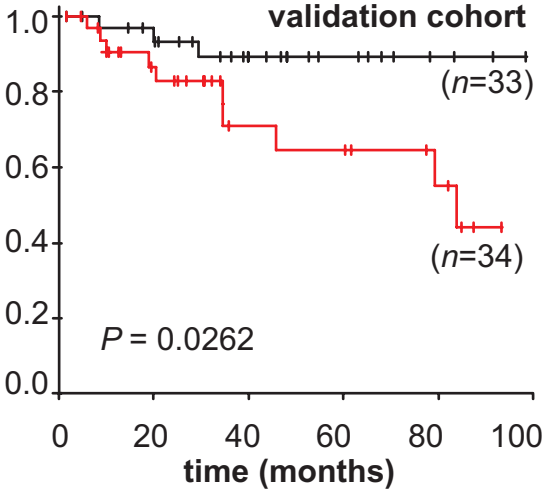
Variables	overall survival HR (95% CI)	*P*-value
Age (Years)	1.02 (0.99-1.05)	0.085
Gender		
Female	1.00	
Male	1.06 (0.61-1.83)	0.845
CFL1 level		
Low	1.00	
High	2.70 (1.54-4.75)	0.001
Type of cancer		
Squamous-cell	1.00	
Adenocarcinoma	1.32 (0.77-2.29)	0.316
Tumour stage		
IA	1.00	
IB	2.35 (1.09-5.07)	0.030
IIA/B	2.76 (1.20-6.37)	0.017
IIIA/B-IV	4.01 (1.90-8.46)	<0.001

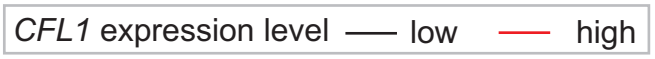
*Cohort description in supplementary materials (*n* = 111)

c

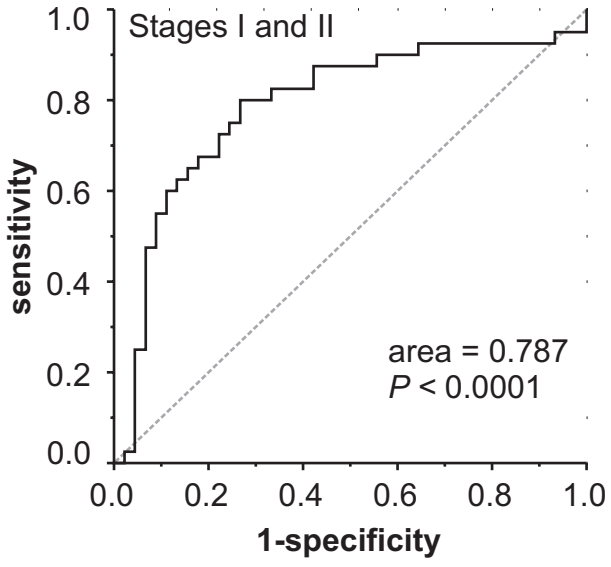


d

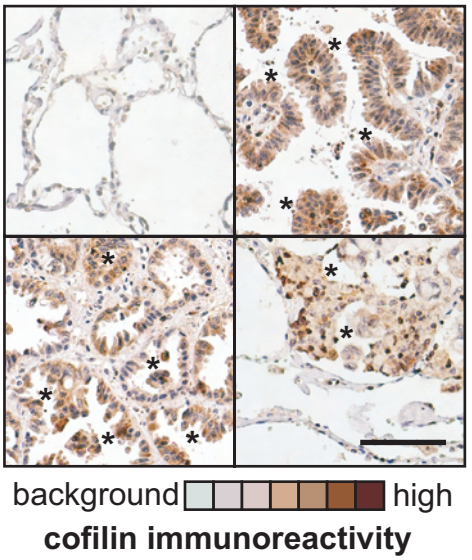


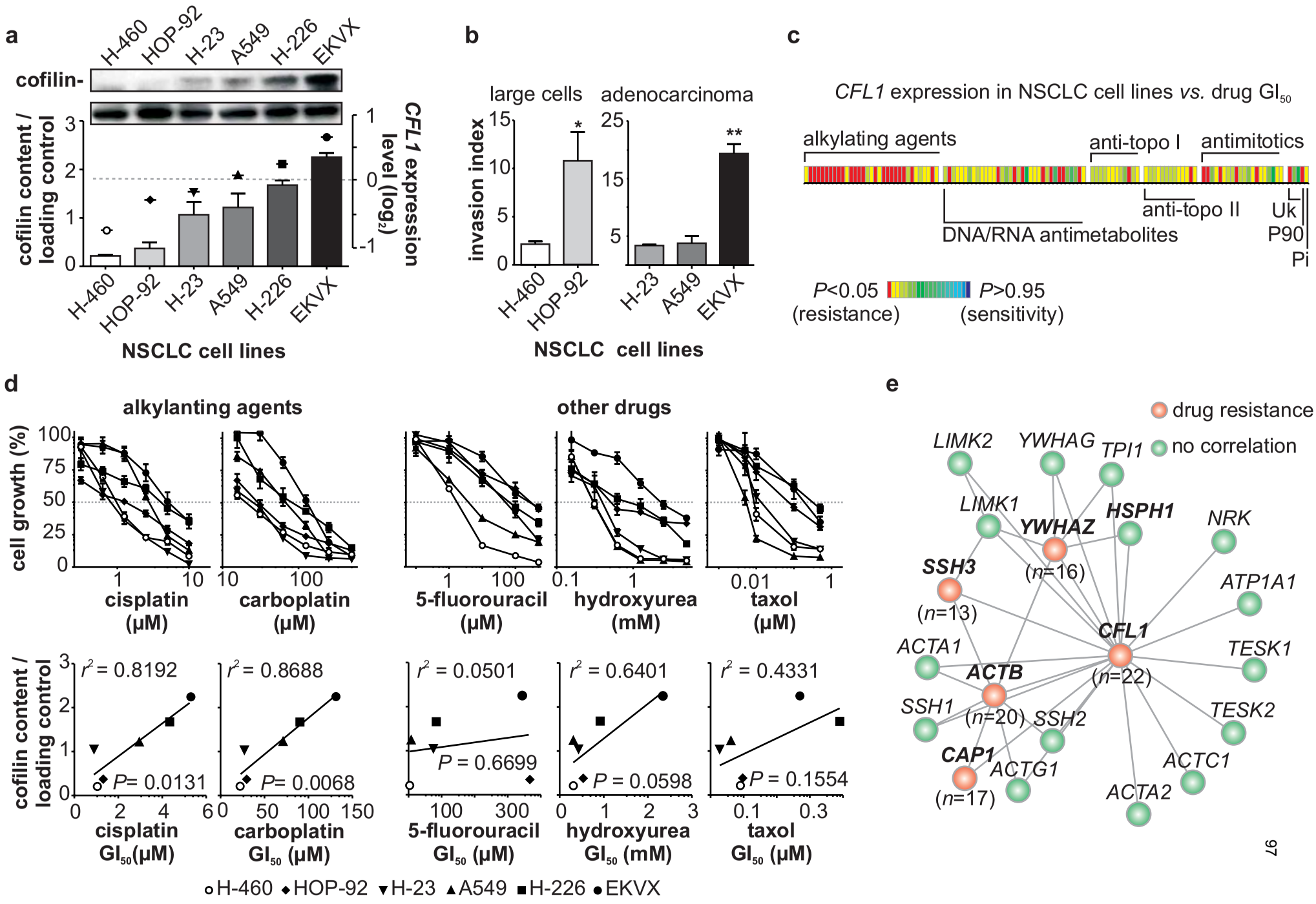


e



f




a
H-460
HOP-92
H-23
A549
H-226
EKVX
cofilin-
cofilin content / loading control
CFL1 expression level (log2)
NSCLC cell lines
b
large cells
adenocarcinoma
invasion index
NSCLC cell lines
c
CFL1 expression in NSCLC cell lines vs. drug GI50
alkylating agents
anti-topo I
antimitotics
anti-topo II
Uk
P90
Pi
DNA/RNA antimetabolites
P<0.05 (resistance)
P>0.95 (sensitivity)
d
alkylanting agents
other drugs
cell growth (%)
cisplatin (μM)
carboplatin (μM)
5-fluorouracil (μM)
hydroxyurea (mM)
taxol (μM)
r2 = 0.8192
r2 = 0.8688
r2 = 0.0501
r2 = 0.6401
r2 = 0.4331
P= 0.0131
P= 0.0068
P = 0.6699
P = 0.0598
P = 0.1554
cisplatin GI50(μM)
carboplatin GI50 (μM)
5-fluorouracil GI50 (μM)
hydroxyurea GI50 (mM)
taxol GI50 (μM)
H-460 HOP-92 H-23 A549 H-226 EKVX
e
drug resistance
no correlation
LIMK2
YWHAG
TPI1
LIMK1
YWHAZ
HSPH1
NRK
SSH3
(n=16)
(n=13)
ATP1A1
CFL1
ACTA1
TESK1
ACTB
(n=22)
SSH1
(n=20)
SSH2
TESK2
CAP1
ACTG1
ACTC1
(n=17)
ACTA2

SUPPLEMENTARY METHODS

Tumor samples and microarray data For NSCLC tumor analysis we accessed a large well-defined collection of lung cancer samples with expression data and relevant clinical and pathologic information on 111 patients (testing cohort), from core biopsies of patients' tumor. The data was obtained from GEO database (http://www.ncbi.nlm.nih.gov/projects/geo/; Series GSE3141) and the Duke Institute for Genome Sciences & Policy website (http://data.cgt.duke.edu/oncogene.php). Gene array data is available on Affymetrix U133A GeneChip[1]. To test the reproducibility of the data we performed the same analysis on a second, independent microarray data set (validation cohort), which is available on different microarray platform (Affymetrix HG_U95Av2 GeneChip)[2]. The validation cohort comprises microarray data from 86 tumor biopsies obtained from sequential patients seen at the University of Michigan Hospital for stage I or stage III lung adenocarcinomas. All gene array data of the validation cohort are available at http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html.

Survival data analysis Standard Kaplan–Meier mortality curves and their significance levels were generated for clusters of patients using SPSS software (SPSS for Windows, release 14.0.0. SPSS Inc., Chicago, IL). The survival curves are compared using the log-rank test and patients are clustered according to either biomarker expression level or NCSLC stage grouping[3].

Cox multivariable regression analysis Multivariate Cox proportional hazards regression models were used to test the independent contribution of each variable on mortality. Graphical assessment was used to assess the Cox model's proportional

hazard assumption. Results of multivariate analysis were summarized by calculating hazard ratios (HR) and corresponding 95% confidence intervals (CI).

Biomarker accuracy The area under the receiver-operating characteristic (ROC) curve was used to evaluate the biomarker's ability in discriminating patients who survived and those who died. An optimal cut-off value was obtained considering the combination of highest sensitivity and specificity.

Immunohistochemical staining Paraffin-embedded sections of lung samples from 20 patients with NSCLC (classified according to World Health Organization criteria) were obtained as archival specimens from the Department of Pathology at the São João Batista Hospital in Criciúma, SC, Brazil. Hematoxylin–eosin (H&E)–stained slides of lung tissue were examined by a national board–certified pathologist. Selected areas of lung cancer and corresponding benign samples were sectioned into 3 μm slices, and immunohistochemical staining was performed according to the standard avidin-biotin immunoperoxidase complex technique. Rabbit anti-human cofilin-1 antibody (Cell Signaling Technology) was diluted in a 1:50 proportion and used as the primary antibody. The brownish-color was considered to be evidence of a positive expression of cofilin-1 in the tumor cells. Unstained red blood cells and labeled macrophages were considered, respectively, as negative and positive internal controls. The Helsinki Declaration of Human Rights was strictly observed when performing these experiments.

Cell culture and western blot immunoassay The human NSCLC cell lines were obtained from NCI-Frederick Cancer DCTD tumor/cell line repository, and

grown in RPMI 1640 medium containing 10% heat-inactivated fetal bovine serum, 2 mM L-Glutamine at 37°C in 5% CO_2 in air. Exponentially growing cells were washed twice with PBS and resuspended in lysis buffer containing 20 mM Tris, pH 7.5, 150 mM NaCl, 1 mM EGTA, 1% Triton, 1 mM Na_3VO_4 and protease inhibitors. After sonication, 30 µg of protein was electrophoresed on 4-12% Bis-Tris NuPage gels (Invitrogen), transferred to PVDF membranes (Immobilon P, Millipore) and blocked with 5% milk. The following antibodies were used for Western blot immunoassay: rabbit polyclonal anti-cofilin (1:1,000), rabbit polyclonal anti-actin (1:2,000) (Cytoskeleton, Denver, CO. USA). Horseradish peroxidase-linked secondary antibody (1:10,000) was from DakoCytomation. Bands were visualized by chemiluminescence using the ECL Detection kit from Amersham Biosciences and exposure of X-ray film. Quantification of band was done with ImageJ 1.36b software (NIH, USA).

Drugs cytotoxicity Drugs GI_{50} was determined as described elsewhere. Briefly, exponential growing NSCLC cell lines were treated with different concentrations of drugs (cisplatin, carboplatin, 5-fluorouracil, hydroxyurea and taxol) (Sigma). After 72 h, the medium was removed and cells were fixed with cold 10% TCA for 1 h at 4°C. Plates were washed five times with distilled water and left to dry at room temperature. Cells were stained with 0.4% of sulforhodamine B (Sigma) (w/v) in 1% acetic acid (v/v) at room temperature for 20 min. SRB was removed and the plates washed five times with 1% acetic acid before air-drying. Bound dye was solubilized with 10 mM unbuffered Tris-base solution and plates were left on a plate shaker for at least 10 min. Absorbance was measured in a 96-well plate reader (VERSA*max*, Molecular Devices) at 492 nm. The growth inhibition (GI_{50}) was

calculated according to the concentration-response curve. The mean of three independent experiments for each condition run in triplicates is plotted.

Cell Migration and Invasion Assays *In vitro* migration and invasion assays were performed using the BioCoat™ Matrigel™ Invasion Chamber System (BD Bioscience). Briefly, Matrigel inserts were rehydrate in RPMI medium and cells ($2.5x10^4$ cells) were seeded at each 24 well chamber. The chemoattractant (medium RPMI with 10% of SFB) were added to the lower wells and the movement of cells through the 8.0 µm pore size Tran-swell cell culture inserts (Falcon), either un-coated (migration) or Matrigel coated (invasion), were determined after 22 h of incubation at 37°C in a humidified incubator with 5% CO_2 atmosphere. At the end of the assay, cells were removed from the top side of the insert using a cotton swab. Cells that penetrated to the underside surfaces of the inserts were fixed and stained with HEMA 3 staining kit (Fisher Scientific) and counted under the microscope. Data is expressed as the percent invasion through the Matrigel relative to the migration though the un-coated membrane, and expressed as invasion index. The mean of three high power fields for each condition run in triplicates is plotted.

Microarray data from NCI-60 cancer cell panel Transcript expression profiles of the six human NSCLC cell lines were obtained from the NCI60 human tumour cell line anticancer drug screen (http://discover.nci.nih.gov/datasetsNature2000.jsp).
To test the reproducibility of the data we assessed a second, independent microarray data set available at http://discover.nci.nih.gov/cellminer/home.do (RMA normalized Affymetrix HG-U133A/B data set). This second microarray platform comprises the human transcriptome and consistently identifies gene probes (*e.g.* it follows approved

gene IDs from the HGNC nomenclature committee – http://www.genenames.org/), allowing the proper identification of *CFL1* partners in the biological network analysis.

The drug database For drug activity analysis we considered those compounds listed in the "mechanism of action" drug activity database, NCI Developmental Therapeutics Program (http://discover.nci.nih.gov/nature2000/). This panel consists of 118 compounds whose mechanisms of action are putatively classified: *i*) alkylating agents; *ii*) topoisomerase I inhibitor; *iii*) topoisomerase II inhibitor; *iv*) DNA/RNA antimetabolites (DNA binder, DNA incorporation, antifols, ribonucleotide reductase inhibitor, DNA synthesis inhibitor, RNA synthesis inhibitor); *v*) antimitotics; and *vi*) others (protein synthesis inhibitor; hsp90 binder or unknown). Drug activity against the NCSLC cell lines is expressed by 50% growth inhibition doses (GI_{50}; also known as IC_{50}) and the entire GI_{50} dataset is available at http://dtp.nci.nih.gov/dtpstandard/cancerscreeningdata/index.jsp.

***CFL1* chemotherapeutic drug resistance/sensitivity data analysis** The relation between the activity of the drug dataset (*i.e.* 118 standard chemotherapy agents) and *CFL1* expression levels was estimated by Spearman correlation analysis in SPSS software (SPSS for Windows, release 14.0.0. SPSS Inc., Chicago, IL). Positive correlations occurred when relatively high levels of gene expression were found in relatively sensitive cell lines. Negative correlations occurred when relatively high levels of gene expression were found in resistant cell lines. Therefore, P values < 0.05 indicate a significant negative correlation (resistance) and $P > 0.95$ indicate a significant positive correlation (sensitivity). Due to multiple comparisons, only drug

categories showing reproducible results were considered for further analysis (*i.e.* consistent results among the drugs of a given class).

Construction of the network-based model of *CFL1* interaction partners Experimental evidences of protein-protein interactions were obtained from STRING database (http://string.embl.de/)[4]. STRING integrates different curated, public databases containing information on direct and indirect functional protein-protein associations. We retrieved all proteins described in that database inferred by experimental evidences and that directly interact with *CFL1* (cofilin-1; non-muscle isoform; Ensembl Peptide ID: ENSP00000309629). The final network was drawn using spring model algorithm and then handled in Medusa software[5]. Gene products are identified according to approved HGNC gene symbols (http://www.genenames.org/).

***CFL1* gene partner analysis** Microarray data of NSCLC cell lines were crossed against GI_{50} values of 118 standard chemotherapy agents in order to estimate drug sensitivity/resistance profile according to the expression levels of *CFL1* gene partners (*i.e.* all genes identified in the network-based model of *CFL1* interaction partners). The statistical analysis follows the original method described in the National Cancer Institute's drug discovery program[6]. Briefly, GI_{50} data were used to populate a correlation matrix along with the expression data for the individual cell lines. For each cell line *i* in the correlation matrix there are *x* genes (*x*= 1...~33,000), corresponding to approximately 44,000 oligonucleotide probes in the HG-U133A/B array, and *y* compounds (*y*=1…118). For the entire cell panel, the correlation between gene expression level and drug sensitivity/resistance profile is given by

$$r = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2 \sum_i (y_i - y_m)^2}} \quad . \quad [1]$$

where x_i is the expression data of the gene x in cell i, y_i is the sensitivity GI_{50} of cell i to drug y, x_m is the mean of the expression data of the gene x, and y_m is the mean sensitivity GI_{50} of drug y. Note that, in HG-U133A array, a given gene may have more than one probe interrogating its expression level. In these cases, the average value of the probe set is considered in the analysis. We estimated the significance of Pearson's correlation coefficients using bootstrap analysis with 50,000 replications. Therefore, the correlation coefficient obtained for each drug–biomarker pair is positioned in the related bootstrap distribution in order to estimate the *P*-value: *P*-values<0.05 indicate a significant negative correlation (resistance), whereas *P*-values>0.95 indicate a significant positive correlation (sensitivity). Bootstrap analysis was performed by in-house program compiled in Microsoft Visual Studio/Intel(R) Fortran Compiler.

Literature search for potential NSCLC biomarkers The list of genes presented in **Supplementary Table 3** was compiled by searching the PubMed database for articles published in English between January 1985 and December 2008. Search criteria included subject heading terms for "biomarker", "prognosis", "gene expression" and "lung cancer". Genes reported in two or more articles during the period of our search, or in one article at least during the last 3 years were selected. Those articles describing pooled biomarkers into the same analysis were excluded from the list (*i.e.* combined performance for multiple genes). Also, a *sine qua non* condition for select a given gene was the presence of two or more microarray probes in the NSCLC cohort study (GEO database series GSE3141;

http://www.ncbi.nlm.nih.gov/geo/). A total of 60 NSCLC potential biomarkers were selected for the analysis. The survival data analysis was performed in SPSS software (as described in the *CFL1* gene analysis) to generate Standard Kaplan–Meier mortality curves with their significance levels, where patients were clustered according to the biomarker expression level (*i.e.* upper fifth *vs.* lower fifth of transcript abundance levels). Survival curves were compared using the log-rank test.

Supplementary References

1. Bild, A.H. *et al. Nature* **439**, 353–357 (2006).
2. Beer, D.G. *et al. Nat. Med.* **8**, 816–824 (2002).
3. Mountain, C.F. *Chest* **111**, 1710–1717 (1997).
4. von Mering,C., *et al. Nucl. Acids Res.* **35**, D358–D362 (2007).
5. Hooper, S.D. & Bork, P. *Bioinformatics* **21**, 4432–4433 (2005).
6. Scherf, U. *et al. Nat. Genet.* **24**, 236–244 (2000).
7. Lu, Y. *et al. PLoS Med* **3**, e467 (2006).
8. Stav, D. *et al. Int.J.Biol Markers* **22**, 108-113 (2007).
9. Oshita, F. *et al. J.Exp.Ther Oncol.* **6**, 49-53 (2006).
10. Takanami, I. & Takeuchi, K. *Jpn.J.Thorac.Cardiovasc.Surg.* **51**, 368-373 (2003).
11. Takanami, I. *et al. Int.J.Cancer* **95**, 384-387 (2001).
12. Karczmarek-Borowska, B. *et al. Lung Cancer* **51**, 61-69 (2006).
13. Karczmarek-Borowska, B. *et al. Folia Histochem.Cytobiol.* **43**, 237-242 (2005).
14. Monzo, M. *et al. J.Clin.Oncol.* **17**, 2100-2104 (1999).
15. Rosell, R. *et al. PLoS ONE.* **2**, e1129 (2007).
16. Guo, N. L. *et al. Clin Cancer Res* **14**, 8213-8220 (2008).
17. Ho, C. C. *et al. Lung Cancer* **59**, 105-110 (2008).
18. Kim, B. *et al. Cancer Res.* **67**, 7431-7438 (2007).
19. Li, R. *et al. Human Pathology* **39**, 1792-1801 (2008).
20. Miyake, M. *et al. Oncogene* **18**, 2397-2404 (1999).
21. Adachi, M. *et al. J.Clin.Oncol.* **16**, 1397-1406 (1998).

22. Oshita, F. *et al. Oncol.Rep.* **16**, 817-821 (2006).
23. Raz, D. J. *et al. Clin Cancer Res* **14**, 5565-5570 (2008).
24. Tsai, M. F. *et al. J Natl Cancer Inst.* **98**, 825-838 (2006).
25. Chen, H. Y. *et al. N.Engl.J.Med.* **356**, 11-20 (2007).
26. Shibata, T. *et al. Cancer Sci.* **98**, 985-991 (2007).
27. Ceppi, P. *et al. Ann.Oncol.* **17**, 1818-1825 (2006).
28. Vallbohmer, D. *et al. Clin.Lung Cancer* **7**, 332-337 (2006).
29. Tang, H. & Goldberg, E. *J Androl* jandrol (2008).
30. Zheng, Z. *et al. N.Engl.J.Med.* **356**, 800-808 (2007).
31. Tian, L. *et al. Cancer* **113**, 1396-1403 (2008).
32. Mano, Y. *et al. Cancer Sci.* **98**, 1902-1913 (2007).
33. Iwakiri, S. *et al. Ann.Surg.Oncol.* **15**, 889-899 (2008).
34. Meyer, B. *et al. Mol Carcinog.* **46**, 503-511 (2007).
35. Gallegos Ruiz, M. I. *et al. PLoS ONE.* **3**, e0001722 (2008).
36. Lee, E. S. *et al. Clin Cancer Res* **14**, 7397-7404 (2008).
37. Kato, T. *et al. Clin.Cancer Res.* **13**, 434-442 (2007).
38. Corson, T. W. *et al. Clin.Cancer Res.* **13**, 3229-3234 (2007).
39. Shibata, T. *et al. Cancer Sci.* **98**, 985-991 (2007).
40. Xing, J. *et al. Br J Cancer* **98**, 1716-1722 (2008).
41. Sun, Z. *et al. Mol Cancer* **3**, 35 (2004).
42. Tomida, S. *et al. Oncogene* **23**, 5360-5370 (2004).
43. Yamashita, S. I. *et al. Clin Oncol.(R.Coll.Radiol.)* **20**, 148-151 (2008).
44. Angulo, B. *et al. J Pathol.* **214**, 347-356 (2008).
45. Hofmann, H. S. *et al. Oncol.Rep.* **16**, 587-595 (2006).
46. Kim, J. H. *et al. Clin Cancer Res* **14**, 2326-2333 (2008).
47. Diederichs, S. *et al. Cancer Res* **64**, 5564-5569 (2004).
48. Amachika, T. *et al. Lung Cancer* **56**, 337-340 (2007).
49. Lin, X. *et al. Clin.Cancer Res.* **12**, 5720-5725 (2006).
50. Miura, N. *et al. Cancer Sci.* **97**, 1366-1373 (2006).
51. Deng, W. G. *et al. J.Biol Chem.* **282**, 26460-26470 (2007).
52. Ceppi, P. *et al. J Thorac.Oncol.* **3**, 583-589 (2008).
53. Shintani, Y. *et al. Lung Cancer* **45**, 189-196 (2004).
54. Shintani, Y. *et al. Int.J.Cancer* **104**, 790-795 (2003).

Supplementary Table 1. Clinical Characteristics of the Original and Validation Cohorts.

Characteristic	*CFL-1* Expression		*P* value
	High	Low	
Original Cohort (n = 111)	55 (49%)	56 (51%)	
Age (years)	64.6 ± 9.6	64.9 ± 9.7	0.842
Gender			
Male	30 (54%)	33 (59%)	0.784
Female	25 (46%)	23 (41%)	
Tumor type			
Adenocarcinoma	28 (51%)	30 (54%)	0.928
Squamous-cell	27 (49%)	26 (46%)	
Tumor TNM Stage			
Ia	20 (36%)	20 (36%)	0.999
Ib	13 (24%)	14 (25%)	
II	9 (16%)	9 (16%)	
III-IV	13 (24%)	13 (23%)	
Validation Cohort (n = 86)	43 (50%)	43 (50%)	
Age (years)	62.3 ± 8.8	65.1 ± 10.7	0.187
Gender			
Male	21 (49%)	14 (33%)	0.198
Female	22 (51%)	29 (67%)	
Tumor type/differentiation			
Adenocarcinoma/well	12 (28%)	12 (28%)	0.964
Adenocarcinoma/ moderate	21 (49%)	20 (47%)	
Adenocarcinoma/ poor	10 (23%)	11 (26%)	
Tumor TNM Stage			
I	34 (79%)	33 (77%)	0.999
III	9 (21%)	10 (23%)	

Supplementary Table 2. List of alkylating agents for which *CFL1* levels is a biomarker[1] for drug resistance.

Class[2]	Drugs	R_s	*P*-value
A2	Porfiromycin	0.771	0.036
A6	Carmustine (BCNU)	1.000	0.000
A6	Chlorozotocin	0.943	0.002
A6	Clomesone	0.943	0.002
A6	Lomustine (CCNU)	0.771	0.036
A6	Mitozolamide	0.943	0.002
A6	PCNU	0.943	0.002
A6	Semustine (MeCCNU)	0.886	0.009
A7	Asaley	0.771	0.036
A7	Carboplatin	0.829	0.021
A7	Chlorambucil	0.829	0.021
A7	Cisplatin	0.829	0.021
A7	Cyclodisone	0.943	0.002
A7	Hepsulfam	0.771	0.036
A7	Iproplatin	1.000	0.000
A7	Mechlorethamine	0.943	0.002
A7	Melphalan	0.771	0.036
A7	Piperazine mustard	0.943	0.002
A7	Piperazinedione	0.771	0.036
A7	Spiromustine	0.886	0.009
A7	Uracil mustard	0.829	0.021
A7	Yoshi-864	0.771	0.036

[1]Meta-analysis data of chemotherapeutic drugs from a panel of 33 alkylating agents (from Fig.2c) tested for positive correlation (resistance) between drug GI_{50} (μM) and the pattern of *CFL1* gene expression in six human NSCLC cell lines (A549, EKVX, HOP-92, NCI-H226, NCI-H23, NCI-H460) obtained from the NCI-60 cell panel.

[2]Mechanism of action codes: A2= alkylating at N-2 position of guanine; A6=alkylating at O-6 position of guanine; A7=alkylating at N-7 position of guanine.

Supplementary Table 3. Potential NSCLC biomarkers previously described.

Symbol[1]	Gene Name[1]	Gene ID[1]	*P-value*[2]	Ref.
ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	4363	0.6239	[7]
AGER	advanced glycosylation end product-specific receptor	177	0.0894	[8]
ALKBH1	alkB. alkylation repair homolog 1	8846	0.5685	[9]
ALKBH5	alkB. alkylation repair homolog 5	54890	0.1704	[9]
ALKBH7	alkB. alkylation repair homolog 7	84266	0.6970	[9]
ALKBH8	alkB. alkylation repair homolog 8	91801	0.5752	[9]
AMFR	autocrine motility factor receptor	267	0.6829	[10,11]
BCL2L1	BCL2-like 1	598	0.0661	[12]
BIRC5	baculoviral IAP repeat-containing 5	332	0.3584	[13,14]
BRCA1	breast cancer 1, early onset	672	0.5103	[15]
CALB1	calbindin 1	793	0.3541	[16]
CAV1	caveolin 1, caveolae protein	857	0.0808	[17]
CBLC	Cas-Br-M (murine) ecotropic retroviral transforming sequence c	23624	0.0608	[18]
CCNB2	cyclin B2	9133	0.8282	[8]
CCND1	cyclin D1	595	0.1177	[19]
CD9	CD9 molecule	928	0.7940	[20,21]
CDK8	cyclin-dependent kinase 8	1024	0.4337	[22]
CRABP1	cellular retinoic acid binding protein 1	1381	0.5485	[7]
CTSB	cathepsin B	1508	0.6717	[23]
DNAJB4	DnaJ (Hsp40) homolog, subfamily B, member 4	11080	0.2820	[24]
DUSP6	dual specificity phosphatase 6	1848	0.2842	[25]
EGFR	epidermal growth factor receptor	1956	0.6074	[26,27]
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2.	2064	0.2379	[26,28]

Supplementary Table 3. *Continued*.

Symbol[1]	Gene Name[1]	Gene ID[1]	*P-value*[2]	Ref.
ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	2065	0.5191	[29]
ERCC1	excision repair cross-complementing rodent repair deficiency.	2067	0.4269	[27,30,31]
FGFR1OP	FGFR1 oncogene partner	11116	0.1871	[32]
FOLR1	folate receptor 1 (adult)	2348	0.3262	[33]
GSTA1	glutathione S-transferase alpha 1	2938	0.6353	[16]
HMGA2	high mobility group AT-hook 2	8091	0.7014	[34]
HSP90AA1	heat shock protein 90kDa alpha (cytosolic), class A member 1	3320	0.4367	[35]
IFI44	interferon-induced protein 44	10561	0.9626	[36]
IGF2BP1	insulin-like growth factor 2 mRNA binding protein 1	10642	0.8041	[37]
IL1A	interleukin 1 alpha	3552	0.4672	[9]
ILF3	interleukin enhancer binding factor 3	3609	0.7057	[16]
KIF14	kinesin family member 14	9928	0.9346	[38]
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	3845	0.4472	[39]
LARS2	leucyl-tRNA synthetase 2, mitochondrial	23395	0.8316	[7]
LCK	lymphocyte-specific protein tyrosine kinase	3932	0.5878	[29]
LST1	leukocyte specific transcript 1	7940	0.2105	[7]
MBD2	methyl-CpG binding domain protein 2	8932	0.4473	[40]
MMD	monocyte to macrophage differentiation-associated	23531	0.1662	[25]
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	4609	0.4586	[41,42]

Supplementary Table 3. *Continued*.

Symbol[1]	Gene Name[1]	Gene ID[1]	*P-value*[2]	Ref.
P53AIP1	p53-regulated apoptosis-inducing protein 1	63970	0.8171	[43]
PIK3CA	phosphoinositide-3-kinase.	5290	0.1111	[44]
PLAU	plasminogen activator. urokinase	5328	0.1314	[45]
PRDX2	peroxiredoxin 2	7001	0.9166	[46]
PRSS3	protease, serine, 3	5646	0.5814	[47]
RRM1	ribonucleotide reductase M1 polypeptide	6240	0.8498	[27,30]
SLC1A7	solute carrier family 1, member 7	6512	0.4883	[16]
SLC2A1	solute carrier family 2 (facilitated glucose transporter), member 1	6513	0.7739	[7]
SP100	SP100 nuclear antigen	6672	0.9490	[18]
STAT1	signal transducer and activator of transcription 1. 91kDa	6772	0.9245	[25]
STC1	stanniocalcin 1	6781	0.1508	[7]
STYK1	serine/threonine/tyrosine kinase 1	55359	0.6834	[48]
TAL2	T-cell acute lymphocytic leukemia 2	6887	0.4593	[16]
TERF2	telomeric repeat binding factor 2	7014	0.1028	[49]
TERT	telomerase reverse transcriptase	7015	0.7542	[50]
TFAP2B	transcription factor AP-2 beta	7021	0.1814	[51]
TOP2A	topoisomerase (DNA) II alpha	7153	0.1998	[52]
TYMS	thymidylate synthetase	7298	0.1742	[53,54]

[1]Gene symbols and names according to HUGO Gene Nomenclature Committee, HGNC database (http://www.genenames.org).

[2]Differences in survival were assessed with the log-rank test using SPSS software, where patients were clustered according to the biomarker expression level (*i.e.* upper fifth *vs.* lower fifth of transcript abundance levels). In all statistical analyses, *P*-value less than 0.05 was considered significant.

DEPÓSITO DE PATENTE OU DE CERTIFICADO DE ADIÇÃO

Capítulo 7: *Processo in vitro para diagnóstico cariotípico e kit para diagnóstico cariotípico in vitro (BRPI0602793-8).*

República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(11) (21) **PI 0602793-8 A**

(22) Data de Depósito: 11/07/2006
(43) Data de Publicação: **26/02/2008**
(RPI 1938)

(51) Int. Cl[7].:
C12Q 1/68 (2008.01)

(54) Título: **PROCESSO IN VITRO PARA DIAGNÓSTICO CARIOTÍPICO E KIT PARA DIAGNÓSTICO CARIOTÍPICO IN VITRO**

(71) Depositante(s): Universidade Federal do Rio Grande do Sul (BR/RS)

(72) Inventor(es): José Cláudio Fonseca Moreira, Antônio Alves Castro, Maria Cunha de Almeida, Gunnar Hugo Onsten

(57) Resumo: Processo in vitro para Diagnóstico Cariotípico e Kit para Diagnóstico Cariotípico in vitro. A presente invenção é relacionada a um kit e a um processo in vitro para estimar a diversidade cariotípica de um eucarioto, ou de um grupo de eucariotos, através da análise de células e/ou tecidos dos mesmos. O processo e o kit da presente invenção são particularmente úteis para avaliação diagnóstica e prognóstica de pacientes com neoplasias sólidas, incluindo carcinomas e adenocarcinomas, tumores mesenquimais e tumores do sistema nervoso central, bem como neoplasias hematológicas, entre outras. O processo da invenção também proporciona a avaliação indireta da instabilidade genômica.

RELATÓRIO DESCRITIVO

Processo *in vitro* para Diagnóstico Cariotípico e
Kit para Diagnóstico Cariotípico *in vitro*

Campo da invenção

A presente invenção é relacionada a um kit e a um processo *in vitro* para estimar a diversidade cariotípica de um eucarioto, ou de um grupo de eucariotos, através da análise de células e/ou tecidos dos mesmos. Mais especificamente, a presente invenção descreve um processo *in vitro* para avaliar o número e a morfologia dos cromossomos presentes em cada célula da amostra, proporcionando: a construção do cariótipo de cada célula; o agrupamento dos cariótipos com igual número de aberrações cromossomais; a obtenção da proporção de cada grupamento de cariótipos em relação ao número total de células analisadas; a classificação de parâmetros de cariótipos de acordo com suas diversidades cariotípicas; a obtenção de índices de diversidade intraneoplásica e interneoplásicas, no caso do uso no diagnóstico cariotípico de humanos. O processo e o kit da presente invenção são particularmente úteis para avaliação diagnóstica e prognóstica de pacientes com neoplasias sólidas, incluindo carcinomas e adenocarcinomas, tumores mesenquimais e tumores do sistema nervoso central, bem como neoplasias hematológicas, entre outras. O processo da invenção também proporciona a avaliação indireta da instabilidade genômica.

Antecedentes da invenção

A morfologia e o número de cromossomos de uma célula são representados pelo seu cariótipo, o qual pode ser construído por diferentes métodos de citogenética. Estes métodos são rotineiramente empregados em humanos para o diagnóstico e prognóstico de neoplasias hematológicas como, por exemplo, leucemias e linfomas, sendo de extrema importância para o tratamento dos pacientes. Estes métodos também podem ser empregados em desordens cromossômicas, tais como Síndrome de Down, porém são muito pouco utilizados em neoplasias sólidas: tumores epiteliais malignos causam cerca de 80% das mortes humanas por câncer, porém constituem apenas 10% das neoplasias cariotipadas. Esta desproporção decorre de um grande número de problemas analíticos, sendo o

principal deles a elevada heterogeneidade dos tumores sólidos, com aberrações cromossomais muito complexas e de difícil caracterização. Em geral, o grau de desordem dos cromossomos é tão severo que impossibilita a identificação de aberrações recorrentes, as quais poderiam ter algum valor diagnóstico. Apenas a constatação de um cariótipo com muitas aberrações tem significado prático, pois indica, em geral, um mau prognóstico. Porém, esta informação tem baixa sensibilidade e especificidade, resultando na subutilização da cariotipagem.

A literatura Patentária encontrada contempla alguns documentos relacionados ao assunto, porém sem antecipar ou sugerir os objetos da presente invenção.

O pedido internacional de patente WO 98/13525 de titularidade de Applied Spectral Imaging Ltd e intitulado “Method For Chromosome Classification By Decorrelation Statistical Analysis and Hardware Therefore”, descreve uma técnica de cariotipagem espectral com propósitos diagnósticos para detectar aberrações cromossomais em diferentes tipos celulares, tais como células fetais, câncer etc., sendo mais sensível que as técnicas convencionais de bandeamento. O método descrito nesse documento realiza classificação *in situ* de cromossomos corados com diferentes fluoróforos, interpretando os resultados por meios estatísticos e matemáticos para posterior classificação dos cromossomos e obtenção do cariótipo. O foco principal do documento é a descrição de uma metodologia para obtenção do cariótipo, de forma mais sensível. Entretanto, o método descrito não proporciona a interpretação do nível de complexidade do cariótipo obtido, ou seja, não descreve o grau de diversidade cariotípica.

O pedido internacional de patente WO 04/046370, de titularidade de ANPA Research Institute e intitulado “Method and apparatus for inserting a bundle of newspaper inserts into a hopper”, descreve a obtenção do cariótipo de modo sistemático e quantitativo a partir da análise do DNA genômico de células eucarióticas empregando técnicas biologia molecular. O documento descreve a estimativa da freqüência de certas partes ou segmentos dos cromossomos, o que permite detectar alguns tipos de alterações cromossomais (principalmente deleções e amplificações). É descrito o cariótipo de modo sistemático, sem interpretar o grau de diversidade.

A patente européia EP 1533618, de titularidade de Ludwig-Maximilians-Universität München e intitulada “Method for distinguishing prognostically definable AM“, descreve um método para distinguir subtipos de leucemias mielóides agudas

(LMA) que apresentam cariótipos normais. É descrito e reivindicado o uso de um painel de marcadores de expressão gênica que, uma vez detectados de acordo com o método, permitem diferenciar subgrupos de LMAs com valor prognóstico. Portanto, não é descrito o uso de cariotipagem, não há interpretação das LMAs, os quais são declarados normais para os subtipos definidos.

O pedido internacional de patente WO 02/054074, de titularidade de Erasmus Universiteit Rotterdam e intitulado “Recognition of Tumor-Specific Gene Products in Cancer“, descreve um método de citometria de fluxo para detecção de aberrações cromossomais associadas a produtos gênicos específicos, tais como proteínas de fusão encontradas em translocações cromossomais que podem causar leucemias ou outras neoplasias como, por exemplo, cromossomo Philadelphia. O documento propõe identificar aberrações cromossomais com valor diagnóstico e não se propõe a descrever o conjunto de todas as aberrações (ou da maior parte) necessário para obtenção de um cariótipo. Portanto, o método desse documento não proporciona a estimativa da diversidade das aberrações cromossomais, diferentemente do método proposto na presente invenção.

A Patente norte-americana US 5,567,593, intitulada “Cytodiagnostic method using alstonine as a selective marker, and diagnostic kit containing marker“, descreve o uso de um alcalóide (alstonina) que exibe a propriedade de se associar com o DNA de células tumorais, além de apresentar propriedades fluorescentes (atua como um corante diferencial). Destas propriedades, resulta a possibilidade de obtenção do cariótipo da célula tumoral de forma específica, diferenciando das demais células normais que por ventura possam estar presentes numa amostra de tecido. Portanto, este documento descreve um método de cariotipagem aplicado a células tumorais e não se propõe estimar a diversidade das aberrações cromossomais que possam estar presentes no cariótipo.

Nenhum dos documentos encontrados na literatura patentária relacionada à presente invenção antecipam ou sugerem os objetos da presente invenção, pois não fornecem uma estimativa do grau de desordem dos cariótipos, ou seja, o estado da técnica atual não proporciona meios para quantificar a desorganização dos cromossomos. A presente invenção, por outro lado, proporciona a quantificação do grau de aberração cromossomal, fornecendo estimativas de diversidade cariótipica. O processo descrito na presente invenção apresenta diversas vantagens em relação aos processos *in vitro* já existentes, pois além de proporcionar a interpretação dos

resultados de exames citogenéticos de uma forma inovadora, quantifica o grau de desordem, criando uma escala de diversidade que outras tecnologias até o momento não conseguiram realizar.

Sumário da Invenção

É um objeto da presente invenção descrever um processo *in vitro* para estimar a diversidade cariotípica de um eucarioto, ou de um grupo de eucariotos, através da análise de células e/ou tecidos dos mesmos.

Um objeto adicional da presente invenção é descrever um processo *in vitro* para avaliar o número e a morfologia dos cromossomos presentes em cada célula da amostra.

Um outro objeto da invenção é um processo para construir o cariótipo de cada célula, agrupando os cariótipos com igual número de aberrações cromossomais.

Ainda outro objeto da presente invenção é um processo para obter a proporção de cada grupamento de cariótipos em relação ao número total de células analisadas

Um objeto adicional da presente invenção é proporcionar um processo para classificar parâmetros de cariótipos de acordo com suas diversidades cariotípicas.

Outro objeto da presente invenção é descrever um processo para obter o índice de diversidade intraneoplásica.

É também um objeto da presente invenção proporcionar um kit para diagnóstico cariotípico *in vitro* de um eucarioto, ou de um grupo de eucariotos, através da análise de células e/ou tecidos dos mesmos.

Estes e outros objetos da presente invenção ficarão mais evidentes a partir da descrição detalhada da invenção e das reivindicações anexas.

Breve descrição das figuras

As figuras descritas a seguir integram a presente invenção e estão incluídas para melhor demonstrar certos aspectos de concretizações preferenciais do processo *in vitro* para diagnóstico cariotípico da invenção.

A figura 1 mostra os índices de diversidade cariotípica que caracterizam um conjunto de amostras biológicas (neste caso, biópsias humanas).

A figura 2 mostra os índices de diversidade intracariótipica (*Hi*) obtidos para cada um dos cinco tipos de cariótipos.

A figura 3 mostra os índices de diversidade intraneoplásica (*H*alfa1) obtidos para cada um dos cinco tipos de biópsias humanas.

Descrição detalhada da invenção

O relato a seguir descreve um kit e um processo *in vitro* para estimar a diversidade cariotípica de um eucarioto, ou de um grupo de eucariotos, através da análise de células e/ou tecidos dos mesmos. Embora a invenção possa ser aplicada a qualquer eucarioto contendo organização genômica e/ou cromossomal detectável, será feita referência à concretização preferencial de uso na análise de cariótipos humanos. Os versados na arte saberão prontamente, a partir da presente descrição, que os objetos da presente invenção podem ser aplicados a outros eucariotos. Conseqüentemente, a descrição a seguir tem o intento de exemplificar, mas não de limitar, as aplicações da presente invenção.

O processo *in vitro* para diagnóstico cariotípico da presente invenção compreende os passos de:

a) obter o cariótipo de uma ou mais amostras de células e/ou tecidos de pelo menos um eucariota;

b) analisar a morfologia dos cromossomos presentes em cada célula da amostra;

c) agrupar os cariótipos com igual número de aberrações cromossomais;

e) obter a proporção de cada agrupamento de cariótipos em relação ao total de células analisadas; e

f) obter um índice de diversidade através da contagem de aberrações cromossomais.

A etapa de obter o cariótipo de uma ou mais amostras de células e/ou tecidos de pelo menos um eucariota é conhecida dos versados na arte. Tal etapa em geral faz uso de marcador(es) colorimétrico(s) conhecidos dos versados na arte. Exemplos incluem aqueles descritos nos documentos citados aqui como antecedentes da invenção, sendo, portanto, incorporados à mesma por referência. Tais marcadores fornecem a cor espectral dos cariótipos e possibilitam a detecção de aberrações cromossômicas. Na presente descrição todas as amostras biológicas podem ser preparadas utilizando-se uma solução aquosa saturada de um reagente específico (*i.e. Giemsa* ou *Wright* - solução de eosina - azul de metileno) e um

120

suporte para incubação da amostra do tecido neoplásico para, posteriormente, visualizar a metáfase em microscopia óptica e então compor o cariótipo.

O kit para diagnóstico cariotípico *in vitro* da presente invenção compreende:

a) meios para a obtenção do cariótipo de uma ou mais amostras de células e/ou tecidos de pelo menos um eucariota;

b) meios para proporcionar a análise da morfologia dos cromossomos presentes em cada célula da amostra;

c) meios para proporcionar o agrupamento dos cariótipos com igual número de aberrações cromossomais;

d) meios para proporcionar a obtenção da proporção de cada agrupamento de cariótipos em relação ao total de células analisadas; e

e) meios para proporcionar a obtenção de um índice de diversidade cromossomal.

Os meios para a obtenção do cariótipo do kit da presente invenção podem ser obtidos por um grande número de variações metodológicas. A título exemplificativo são citados aqui dois métodos preferenciais, que incluem reativos de técnicas de citogenética clássica (*i.e.* bandeamento G), e/ou reativo/técnicas de citogenética molecular (*i.e.* hibridização *in situ*). A técnica de bandeamento G consiste basicamente dos seguintes passos: (i) Bloqueio da divisão celular no estágio de metáfase adicionando-se ao material uma solução diluída de *colchicina* ou um agente similar (*colcemida*); (ii) Dispersão dos cromossomos adicionando-se ao material uma *solução hipotônica* (KCL) para inchar o núcleo da célula e facilitar o espalhamento; (iii) Fixação das células adicionando-se uma solução de *metanol* e *ácido acético* (3:1); (iv) Aquecimento das lâminas por 6 horas a 45º C; (v) Tratamento das células com uma solução de *tripsina* (vi) Bandeamento dos cromossomos com o corante *Giemsa* ou *Wright* (solução de *eosina - azul de metileno*); e (vii) Após a coloração, os cromossomos podem ser visualizados ao microscópio e/ou fotografados para posterior interpretação e construção do cariótipo.

As demais etapas do processo da invenção, a partir da etapa b) indicada acima, bem como os demais meios do kit da presente invenção, a partir de seu item b) descrito acima, serão descritas em referência a uma ou mais de suas configurações preferenciais, na(s) qual(is) o processo e o kit da presente invenção permitem quantificar o grau de desordem dos componentes de células neoplásicas.

121

O processo descrito nessa concretização preferencial fornece cinco índices de diversidade, os quais se aplicam a diversas situações, desde a caracterização mais fundamental, que se refere ao fornecimento do grau de desordem de uma única neoplasia até estudos epidemiológicos, onde se fornece o grau de desordem de muitas neoplasias de um mesmo tipo. Uma vez obtidos os índices de diversidade, pode-se classificar tanto neoplasias pouco alteradas do ponto de vista citogenético até as mais aberrantes, numa escala que varia de zero a um.

A materialização do processo da invenção pelos inventores foi conduzida de diferentes formas, sendo aqui descritas algumas. Numa concretização, o processo da presente invenção foi aplicado a um painel de noventa e um tipos de tumores sólidos, sendo demonstrada existência de correlação entre os índices que compõem a escala de diversidade da presente invenção e a sobrevida em cinco anos de pacientes com tumores sólidos. Primeiramente foi quantificada a proporção de clones, incluindo sub-clones e variantes celulares, as quais são presentes em uma amostra de tecido, em biópsia, por exemplo. Para isso, toma-se o número necessário de células, obtendo-se o cariótipo de cada célula por técnicas de citogenética clássica ou molecular (*i.e.* etapa b descrita acima). Os diferentes cariótipos são então agrupados conforme o número de aberrações cromossomais que possuem (*e.g.* aberrações numéricas ou aberrações estruturais). A forma de composição deste agrupamento de cariótipos pode gerar cinco parâmetros de diversidade cariotípica, que podem ser classificados em diversidade intraneoplásica; diversidade interneoplásica; diversidade intercariotípica; diversidade intracariotípica; e por fim a diversidade global de cariótipos.

Cada um destes parâmetros tem uma finalidade distinta na avaliação da diversidade dos cromossomos, os quais podem ser utilizados para dois tipos de análise que são divididas em: análise de uma única amostra individualmente, tal como uma biópsia ou um conjunto de diferentes amostras que podem ser biópsias de um grupo de pacientes com um dado tipo de neoplasia, de acordo com os exemplos abaixo:

Exemplo 1: Protocolo para estimar a diversidade cariotípica em um único paciente, através da realização de uma única biópsia:

A partir de uma biópsia de tecido neoplásico, por exemplo, uma amostra de tumor ou neoplasia maligna, descreve-se *in vitro* o número e a morfologia dos cromossomos presentes em cada célula da amostra. O número total de células a ser

analisada é dependente do grau de confiabilidade desejado. Em seguida é necessário construir o cariótipo de cada célula e, após isto, cariótipos com igual número de aberrações são, então, agrupados. Deve-se considerar separadamente aberrações numéricas e aberrações estruturais para análise. Feito isso, obtém-se a proporção $p(i)$ de cada agrupamento de cariótipos em relação ao número total de células analisadas. A análise do índice de diversidade pode ser conduzida de diferentes formas, sendo aqui apresentada uma forma simples e prática. Embora o tratamento dos dados mostrado a seguir simplifique a condução da presente invenção, tal tratamento não constitui, em si, nenhum dos objetos da invenção. Uma das melhoras maneiras de se obter o índice de diversidade intraneoplásica $H\alpha_1$ é pela soma de $-p(i)\ln p(i)$, através da fórmula descrita a seguir:

$$H_{\alpha 1} = -\frac{1}{\ln(M)}\sum_i p(i)\ln p(i)$$

onde M representa o número de possibilidades na contagem de aberrações, sendo que o valor M pode ser estimado tanto para um dado tipo neoplasia como para o conjunto de todas as neoplasias humanas (neste caso, $M \equiv 181$ para aberrações numéricas e $M \equiv 81$ para aberrações estruturais). A divisão pelo fator $\ln(M)$ normaliza o índice $H\alpha 1$ e, desta forma, o índice de diversidade intraneoplásica pode variar de 0 (zero) a 1,0 (um).

Exemplo 2: Protocolo para estimar a diversidade cariotípica em grupos de pacientes com o mesmo tipo de neoplasia:

A análise é realizada a partir de uma amostra de cariótipos de n biópsias de um dado tipo de neoplasia, tais como neoplasias sólidas ou hematológicas, onde cada biopsia α da amostra (α =1,...,n) tem $K\alpha$ diferentes cariótipos, entre M possíveis cariótipos. Os cariótipos são caracterizados pelo tipo de alterações que possuem (aberrações numéricas e aberrações estruturais). Considere, por ora, i como indicando o número de cromossomos, o qual pode variar de i_0 a i_f, por exemplo. Nesse caso, o número de possíveis cariótipos será $M = i_f - i_0 + 1$. Assim, para cada biópsia α, define-se $s(i,\alpha)$ como sendo o número de cariótipos com i cromossomos. $K\alpha$ é dado por:

$$K_\alpha = \sum_{i=i_0}^{i=i_f} s(i,\alpha)$$

Em seguida, obtêm-se outras quantidades a partir dos valores de $s(i,\alpha)$ para toda amostra. Começar pelo número L_i de cariótipos do tipo *i* para toda amostra de *n* biópsias

$$L_i = \sum_{\alpha=1}^{n} s(i,\alpha)$$

Agora, obtêm-se a partir de $K\alpha$ ou L_i o número total de *N* cariótipos*:*

$$N = \sum_{\alpha=1}^{n} K_\alpha = \sum_{i=i_0}^{i=i_f} L_i$$

Em seguida, a partir das equações citadas anteriormente, definiu-se as diferentes funções normalizadas de probabilidade que seguem. Inicia-se, pela mais geral de todas, $z(i,\alpha)$*,* ou seja a probabilidade de escolher aleatoriamente, entre todos os *N* cariótipos, um que é do tipo *i* e pertence à biópsia α e a expressão é definida como segue:

$$z(i,\alpha) = \frac{s(i,\alpha)}{N}$$

Definiu-se também para cada biópsia α a probabilidade $p(i,\alpha)$ de escolher aleatoriamente um cariótipo do tipo *i*, a qual foi calculada pela seguinte expressão:

$$p(i,\alpha) = \frac{s(i,\alpha)}{K_\alpha}$$

e, para cada tipo i de cariótipo, definiu-se a probabilidade $\eta(i,\alpha)$ de escolher aleatoriamente um cariótipo que pertence à biópsia α,

$$\eta(i,\alpha) = \frac{s(i,\alpha)}{L_i}$$

Finalmente, definiu-se $P\alpha$ como a probabilidade de escolher aleatoriamente, entre todos os *N* cariótipos, um que pertence à biópsia α*:*

$$P_\alpha = \frac{K_\alpha}{N}$$

e, analogamente, definiu-se a probabilidade Q_i de escolher aleatoriamente, entre todos os *N* cariótipos, um que seja do tipo *i:*

$$Q_i = \frac{L_i}{N}$$

Todas as funções de probabilidade acima são então normalizadas. As condições de normalização são escritas como:

$$\sum_i \sum_\alpha z(i,\alpha) = 1;\ \sum_i p(i,\alpha) = 1;\ \sum_\alpha \eta(i,\alpha) = 1\ ;\ \sum_\alpha P_\alpha = 1 \text{ e } \sum_i Q_i = 1$$

Associa-se, para cada uma das funções de probabilidade descritas acima, uma função de informação, como segue:

$$H_{global} = -\frac{1}{\ln(Mn)} \sum_i \sum_\alpha z(i,\alpha) \ln z(i,\alpha)$$

$$H_\alpha = -\frac{1}{\ln(Mn)} \sum_i p(i,\alpha) \ln p(i,\alpha)$$

$$H_i = -\frac{1}{\ln(Mn)} \sum_\alpha \eta(i,\alpha) \ln \eta(i,\alpha)$$

$$H^* = -\frac{1}{\ln(Mn)} \sum_\alpha P_\alpha \ln P_\alpha$$

$$H' = -\frac{1}{\ln(Mn)} \sum_i Q_i \ln Q_i$$

onde dividiu-se todos os termos pelo fator ln(*Mn*) com objetivo de normalizar todas as quantidades. Dessa forma, todo *H* irá variar entre 0 (zero) e 1,0 (um). O valor obtido para cada um das funções acima revela um aspecto da diversidade cariotípica.

Para a interpretação dos índices de diversidade cariotípica, utilizou-se a função $H_{global,}$ a qual reflete o espalhamento da distribuição $z(i,\alpha)$, ou seja, a diversidade global de cariótipos, enquanto que H_α representa o espalhamento existente no interior de uma dada biópsia, ou seja, mede a diversidade cariotípica que existe em uma amostra de tecido, denominada diversidade intraneoplásica. (quando n=1, o valor de H_α será igual à $H\alpha_1$). Analogamente, H_i é o espalhamento associado a um dado tipo *i* de cariótipo, denominado diversidade intracariotípica. A diversidade interneoplásica (H^*) fornece uma medida do espalhamento devido a diferenças no número de cariótipos entre biópsias. Finalmente, H' está relacionado

ao espalhamento entre *i* tipos de cariótipos, que é denominada diversidade intercariotípica. A relação entre estas cinco funções de informação é dada por:

$$H_{global} = H^* + \sum_{\alpha} P_{\alpha} H_{\alpha} = H^* + H_t$$

A expressão acima relata o fato de H_{global} poder ser obtido como a soma de $H_t \equiv \Sigma_\alpha P_\alpha H_\alpha$ (média da diversidade intraneoplásica) e *H*,* indicando uma possível diferença entre no número de cariótipos entre biópsias.

Por outro lado, a equação descrita abaixo fornece H_{global} pela soma de $H_k \equiv \Sigma_i Q_i H_i$ (média da diversidade intracariotípica) e *H'*, indicando uma possível diferença entre os *M* tipos de cariótipos.

$$H_{global} = H' + \sum_{i} Q_i H_i = H' + H_k$$

Para a exemplificação dos ensaios laboratoriais, considere cinco biópsias de cinco diferentes neoplasias (*n*= 5,0), com cinco possíveis cariótipos (*M*=5,0). Para uma dada biópsia α, $s(i,\alpha)$ representa o número de células com *i* cromossomos; $K\alpha$ representa o número de células cariotipadas em cada biópsia. Para todas elas, $K\alpha$= 10,0 e L_i representa o número de cariótipos do tipo *i* descritos entre todas as alfa-biópsias. Os resultados estão agrupados na tabela 1 demonstrada a seguir:

Tabela 1: Ilustração de tabulação amostral

biópsia α	*i cromossomos*					$K\alpha = \sum s(i,\alpha)$
	47	48	49	50	51	
a	**10**	0	0	0	0	10
b	**5**	**5**	0	0	0	10
c	**4**	**3**	**3**	0	0	10
d	**3**	**3**	**2**	**2**	0	10
e	**2**	**2**	**2**	**2**	**2**	10
$L_i = \sum s(i,\alpha)$	24	13	7	4	2	N = 50

Os dados da tabela acima indicam que as diferentes distribuições de cariótipos podem ser consideradas tanto para aberrações numéricas, quanto para aberrações estruturais.

Os diferentes índices de diversidade obtidos a partir desta amostragem estão apresentados nas Figuras 1, 2 e 3. A figura 1 mostra os índices de diversidade cariotípica que caracterizam um conjunto de biópsias. A análise dos resultados da Figura 3 indica que a biópsia α_a apresenta o menor índice de diversidade intraneoplásica ($H\alpha 1$=zero), uma vez que possui um único tipo de cariótipo, ao passo que a biópsia α_5 apresenta o maior índice de diversidade ($H\alpha_1$=1.0), com cinco tipos de cariótipos representados em proporções iguais.

Os demais índices de diversidade caracterizam a amostra como um todo e seus significados estão também associados à caracterização da diversidade no conjunto amostral. O índice diversidade intracariotípica (H_i), pode ser observado na figura 2 e mostra que cariótipos com 47 cromossomos (i=47) estão mais distribuídos entre as biópsias, indicando maior recorrência desta aberração numérica. Em relação à diversidade global (H_{global}), quanto maior a distribuição do número total de cariótipos na tabulação dos dados maior será o índice de diversidade H_{global}. Da mesma forma, quanto mais distribuídos estiverem os cariótipos entre as M possibilidades, maior será a diversidade intercariotípica (H') e quanto mais distribuído o número total de cariótipos entre as biópsias, maior será a diversidade interneoplásica (H^*). Já em relação à média da diversidade intraneoplásica (H_t) e a média da diversidade intracariotípica (H_k), estas são utilizadas para identificar qual a principal contribuição para compor a diversidade global H_{global}, cuja origem pode vir tanto de uma grande diversidade intraneoplásica média (H_t alto) como de uma grande diversidade intracariotípica média (H_k alto).

Portanto, este processo quantifica a diversidade cariotípica sob deferentes aspectos. O mais fundamental deles diz respeito à produção de um índice de diversidade intraneoplásica, $H_{\alpha 1}$, o qual pode ser utilizado para o diagnóstico do grau de complexidade dos cariótipos encontrados na biópsia de um dado paciente. Os demais índices de diversidade dizem respeito à análise de conjuntos de biópsias, o que permite uma avaliação epidemiológica do grau de complexidade encontrado nos cariótipos de grupos de pacientes com o mesmo tipo de neoplasia.

Os versados na técnica apreciarão que outras formas de concretização da presente invenção são viabilizadas a partir dos conhecimentos fornecidos nesta descrição, e que, portanto, pequenas modificações na forma de condução do processo aqui descrito devem ser incluídas dentro do espírito da invenção e do alcance das reivindicações anexas.

REIVINDICAÇÕES

Processo *in vitro* para Diagnóstico Cariotípico e
Kit para Diagnóstico Cariotípico *in vitro*

1. Processo *in vitro* para diagnóstico cariotípico, compreendendo a obtenção do cariótipo de uma ou mais amostras de células e/ou tecidos de pelo menos um eucariota, caracterizado por compreender os passos de:

a) analisar a morfologia dos cromossomos presentes em cada célula da amostra;

b) agrupar os cariótipos com igual número de aberrações cromossomais;

c) obter a proporção de cada agrupamento de cariótipos em relação ao total de células analisadas; e

d) obter um índice de diversidade através da contagem de aberrações cromossomais.

2. Kit para diagnóstico cariotípico *in vitro* compreendendo meios para a obtenção do cariótipo de uma ou mais amostras de células e/ou tecidos de pelo menos um eucariota, caracterizado por adicionalmente compreender:

- meios para proporcionar a análise da morfologia dos cromossomos presentes em cada célula da amostra;
- meios para proporcionar o agrupamento dos cariótipos com igual número de aberrações cromossomais;
- meios para proporcionar a obtenção da proporção de cada agrupamento de cariótipos em relação ao total de células analisadas; e
- meios para proporcionar a obtenção de um índice de diversidade cromossomal.

3. Kit, conforme reivindicação 2, caracterizado pelo fato de que os referidos meios para a obtenção do cariótipo incluem um suporte para incubação da amostra biológica; um marcador colorimétrico; e uma solução alcoólica.

RESUMO

Processo *in vitro* para Diagnóstico Cariotípico e
Kit para Diagnóstico Cariotípico *in vitro*

A presente invenção é relacionada a um kit e a um processo *in vitro* para estimar a diversidade cariotípica de um eucarioto, ou de um grupo de eucariotos, através da análise de células e/ou tecidos dos mesmos. O processo e o kit da presente invenção são particularmente úteis para avaliação diagnóstica e prognóstica de pacientes com neoplasias sólidas, incluindo carcinomas e adenocarcinomas, tumores mesenquimais e tumores do sistema nervoso central, bem como neoplasias hematológicas, entre outras. O processo da invenção também proporciona a avaliação indireta da instabilidade genômica.

130

FIGURAS

Figura 1

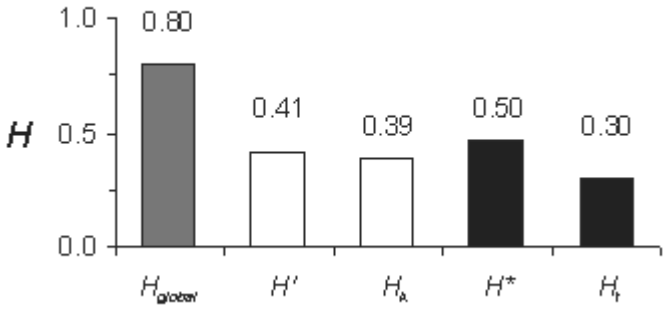
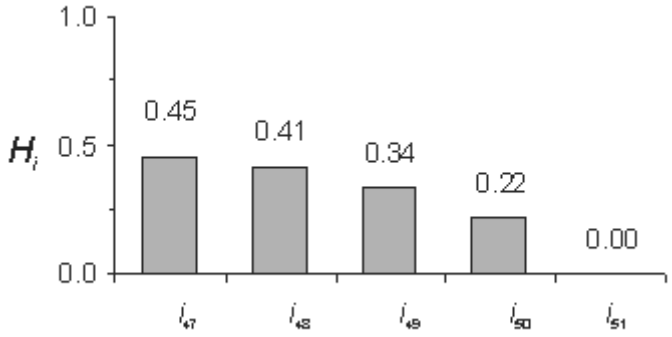
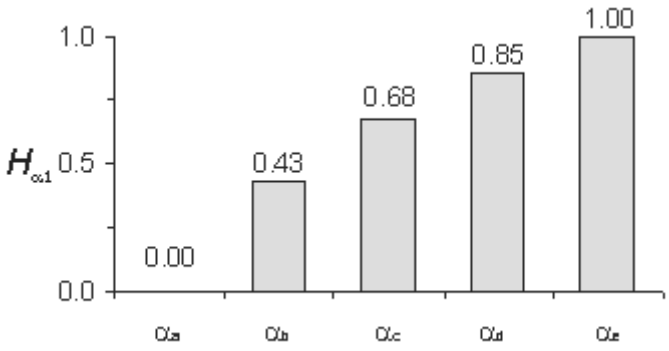


Figura 2



131

Figura 3



PARTE III

DISCUSSÃO GERAL

Diversidade citogenética de neoplasias sólidas

Uma das maiores limitações no uso de exames citogenéticos para o diagnóstico de tumores sólidos está na complexidade das aberrações cromossômicas observadas nos cariótipos (Hoglund et al., 2002b). O grau de desordem é tão severo que impossibilita a identificação de aberrações recorrentes, as quais poderiam ter algum significado prático, a exemplo do que ocorre com a citogenética das neoplasias hematológicas (Mitelman, 2000). Apenas a presença de um cariótipo complexo (com muitas aberrações) tem algum valor clínico, pois indica um mau prognóstico (Albertson et al., 2003; Mitelman et al., 1997). Entretanto, esta informação é irrelevante frente a outros parâmetros utilizados para descrever o grau de agressividade dos tumores (*i.e.* estadiamento e exames anatomopatológicos) (Greene et al., 2002). Do ponto vista teórico, a complexidade dos cariótipos de tumores sólidos é pouco compreendida. Teorias como (*i*) instabilidade genômica ao nível dos cromossomos (Albertson et al., 2003) (*ii*) seleção clonal (Cahill et al., 1999; Sieber et al., 2003), (*iii*) aneuploidização e desequilíbrio cromossômico (Duesberg et al., 2005) e (*iv*) surgimento de um fenótipo mutador (Bielas et al., 2006) apresentam importantes divergências em relação ao grau de aleatoriedade do processo que gera as aberrações. Por exemplo, a hipótese da *instabilidade genômica* sugere que a assinatura da progressão tumoral seria a presença de múltiplas aberrações distribuídas aleatoriamente pelo genoma, ao passo que a hipótese da *seleção clonal* aponta para um padrão não-randômico em que prevaleceria o fenótipo mais adaptado. Essas divergências teóricas refletem, em parte, a ausência de um consenso sobre a dinâmica da progressão tumoral, bem como a ausência de parâmetros seguros para quantificar a complexidade das aberrações cromossômicas.

Qualquer que seja o processo responsável pelo aumento da desordem cromossômica em neoplasias sólidas, é conveniente estimar o desfecho desse processo usando parâmetros bem definidos e robustos frente à reconhecida variabilidade amostral (*i.e.* alta complexidade dos cariótipos). Até onde pudemos constatar, a vasta maioria dos trabalhos em citogenética de neoplasias sólidas apenas descreve o grau de aberração cromossômica, mas não o quantifica.

O trabalho apresentado no *Capitulo I* intitulado "*Profiling cytogenetic diversity with entropy-based karyotypic analysis*" propõem o uso da teoria da informação neste sentido. Neste trabalho nós quantificamos a conteúdo de informação de amostras de cariótipos agrupados usando a informação de Shannon, ou melhor, seu valor positivo, a entropia de Shannon (Shannon, 1948)

$$H' = -\frac{1}{\ln(M)} \sum_{i}^{M} p(i) \ln p(i) \quad , \qquad [1]$$

onde *p* representa a probabilidade (freqüência) de ocorrência de aberrações cromossômicas e *M* representa o número de possibilidades na contagem de aberrações. A divisão pelo fator ln(*M*) normaliza *H'* de forma que índice de diversidade varia de 0,0 (zero) a 1,0 (um).

Em uso coloquial, informação significa conhecimento; na teoria da informação (Shannon, 1948) isso denota uma quantia mensurável (*e.g.* Shannon *H'*). Por exemplo, o aumento da heterogeneidade cromossômica implica em aumento da desordem, ou entropia, e perda de informação sobre o sistema em questão (Gatenby & Frieden, 2002). Por definição, entropia mede variação ou diversidade em uma distribuição de itens ou eventos; padrões invariáveis – tais como os observados em cariótipos normais – têm distribuição cariotípica com entropia zero. Tumores que apresentam diferentes tipos de cariótipos, por outro lado, têm entropia maior que zero, indicando maiores níveis de diversidade (Fuhrman et al., 2000; Kendal, 1990). Em uma interpretação alternativa, o índice *H'* quantifica o espalhamento amostral (algumas ilustrações gráficas e numéricas estão disponíveis nos *Capítulos I e II*).

O ponto de partida para o trabalho apresentado no *Capitulo I* foi explorar padrões de distribuição de uma amostral de 1232 cariótipos, de 14 tipos de neoplasias sólidas, em que constatamos acentuada assimetria e irregularidade, o que não se trata de novidade (Hoglund et al., 2002a), mas permitiu descartarmos abordagens convencionais baseadas em distribuições gaussianas. A sugestão de usar a diversidade de Shannon ocorreu em analogia aos estudos populacionais realizados em ecologia. Nesta disciplina o problema da heterogeneidade amostral é bastante conhecido e inúmeros índices de diversidade já foram propostos, sendo o índice de Shannon (*H'*) um dos mais utilizados (Adami et al., 2000; Desrochers & Anand, 2004; Ricotta, 2003; Ricotta & Szeidl, 2006).

O resultado geral do trabalho apresentado no *Capitulo I* apontou que o índice de Shannon pode ser adaptado para quantificar a diversidade de amostras de cariótipos. Além disso, essa quantificação refletiu o grau de malignidade dos tumores avaliados. Aprimoramos a metodologia ao descrever funções complementares de entropia (entropia global e suas contribuições parciais) em um trabalho subseqüente, o qual está apresentado no *Capítulo II* intitulado "*Chromosome aberrations in solid tumors have a stochastic nature*". Neste novo trabalho utilizamos 79 tipos de tumores sólidos, totalizando 12789 cariótipos, e observamos que a principal contribuição para entropia global em um dado tipo de tumor vem da análise do espalhamento das aberrações entre os tumores que compõem a amostra (diversidade intertumoral), sendo que o grau de espalhamento das aberrações cromossômicas é específico para cada tipo de tumor. Observamos também que o grau de espalhamento das aberrações cromossômicas está diretamente correlacionado com a malignidade (*i.e.* tipos de tumores com menor espalhamento amostral apresentam estatísticas populacionais com a melhor sobrevida), corroborando os achados do trabalho anterior. Finalmente, mostramos que o valor normalizado da entropia global cresce com o tamanho da amostra de maneira assintótica, deixando uma assinatura do grau de aleatoriedade do processo que gerou as aberrações

cromossômicas. Este dado, juntamente com análise dos limites do modelo proposto, nos permitiu inferir o padrão da dinâmica de formação das aberrações cromossômicas, o qual não é nem determinístico nem completamente aleatório, mas situa-se entre estas duas fronteiras. Desse modo, desenvolvemos uma metodologia que quantifica a complexidade dos cariótipos de tumores sólidos, fornecendo um valor de diversidade. Entretanto, é importante reforçar que tais medidas não visam explicar as causas da diversidade, mas sim quantificá-la. Se esta quantificação é capaz de capturar algum novo aspecto da dinâmica de progressão tumoral, então ela própria pode ter algum valor prognóstico.

Apesar de essa metodologia ter antecedentes teóricos (Gatenby et al., 2002; Kendal, 1990), a aplicação em citogenética é inovadora e só foi possível devido à disponibilidade de grandes amostras de cariótipos compiladas no banco de dados Mitelman (Mitelman Database, 2009). Entretanto, merece referência uma trabalho independente (realizado no mesmo período dos artigos apresentados nos *Capítulos I* e *II*) em que os autores empregaram a entropia de Shannon de forma análoga para estimar a diversidade clonal de adenocarcinomas de esôfago (Maley et al., 2006). Este estudo demonstrou a existência de correlação entre o índice de diversidade Shannon e sobrevida em cinco anos, resultado esse que apóia a abordagem desenvolvida em nossos trabalhos [estudos revisados em (Merlo et al., 2006)], bem como a possibilidade de uso da diversidade cariotípica para fins prognósticos em neoplasias sólidas.

Tendo em vista tal possibilidade, levamos a metodologia ao conhecimento da Secretaria de Desenvolvimento Tecnológico – SEDETEC/UFRGS, onde foi realizado um estudo de viabilidade técnica que resultou na patente de invenção BRPI0602793-8 (documento disponível na sessão *Depósito de Patente ou de Certificado de Adição*). Como pode ser observada neste documento, a literatura patentária encontrada contempla algumas invenções relacionadas ao assunto, porém sem antecipar ou sugerir os objetos da invenção: o estado da técnica atual não fornece meios para estimatimar do grau de desordem dos

cariótipos. Já a presente invenção quantifica o grau de aberração cromossômica, fornecendo estimativas de diversidade cariotípica, além de avaliar, de modo indireto, a instabilidade genômica.

Assim, o patente BRPI0602793-8 propõe a utilização de exames citogenéticos para avaliação da diversidade de neoplasias sólidas, o que constitui uma nova aplicação para uma técnica amplamente disponível, uma vez que exames citogenéticos não são empregados no estadiamento de neoplasias sólidas. Considerando apenas o Sistema Único de Saúde (SUS) e o atual estado da técnica de cariotipagem, trata-se de um universo de mais de 22.000 cariótipos/ano destinados, na sua maioria, para neoplasias hematológicas (DATASUS, 2009). A materialização desta invenção está disponibilizada no protótipo KARYOCOMPLEX (software em anexo) o qual incorpora a metodologia descrita na patente BRPI0602793-8 e fornece os meios necessários para a implementação da análise de diversidade de cariótipos.

Diversidade de expressão gênica de neoplasias sólidas

A extensão natural dos resultados discutidos acima foi investigar diferentes fontes de diversidade tumoral com base na teoria da informação. Se o índice *H'* de fato é um bom estimador da diversidade tumoral, então é possível que outros aspectos da heterogeneidade das neoplasias sólidas sejam captados por esta mesma metodologia, como a diversidade de expressão gênica.

Alterações nos sistemas de estabilização do genoma estão diretamente envolvidas no aumento da heterogeneidade das neoplasias sólidas. Estes sistemas estão concentrados em três diferentes e parcialmente sobrepostos processos celulares: (*i*) mecanismos de reparo do DNA (Hoeijmakers, 2001), (*ii*) mecanismos de segregação/estabilização dos cromossomos (Jefford et al., 2006) e (*iii*) mecanismos de morte celular programada (apoptose) (Zhivotovsky & Kroemer, 2004).

Para estimar a diversidade de expressão gênica nestes sistemas nós propusemos um modelo de redes de interação protéica com o intuito caracterizar os componentes dos mecanismos de manutenção do genoma e quantificar sua atividade em neoplasias sólidas. Este trabalho está apresentado no *Capitulo III* intitulado "*Impaired expression of NER gene network in sporadic solid tumors*". O ponto de partida foi definir um conjunto de vias e genes, os quais foram representados coletivamente num grafo de associação contendo: 25 genes do sistema de reparo por recombinação (RER), 25 do sistema de reparo de erros de pareamento (MMR), 17 do sistema de reparo por excisão de bases (BER), 28 do sistema reparo por excisão de nucleotídeos (NER), 77 genes envolvidos na estabilização dos cromossomos e 98 envolvidos em apoptose. Uma vez definido este conjunto com base em extensiva revisão bibliográfica, então adaptamos a entropia de Shannon para quantificar a diversidade de expressão gênica. Nesse sentido, o foco da análise não se constituiu na atividade dos componentes individualmente, mas na expressão da via como um todo onde *H*' quantifica o espalhamento da expressão gênica de cada uma das vias (*e.g.* diversidade de transcritos). O conteúdo de informação dos agrupamentos de genes foi quantificado de forma análoga ao apresentado na equação [1], mas aqui *p* representa a probabilidade (freqüência) de expressão de um dado gene e *M* representa o número total de genes da via. Os dados de expressão foram obtidos da coleção de bibliotecas de transcriptomas disponíveis no projeto genoma do câncer (Lash et al., 2000), e a íntegra dos resultados foi depositada nos suplementos do artigo do *Capitulo III* (Castro et al., 2007).

Este trabalho sugere que as redes de genes de apoptose e do sistema de reparo do DNA por excisão de nucleotídeos estão funcionalmente alteradas em neoplasias sólidas, com aumento da diversidade de expressão gênica e diminuição da abundância de transcritos. A novidade nestes resultados contrapõe-se ao fato do sistema NER não ser descrito como estruturalmente alterado em câncer, salvo em Xeroderma Pigmentosum (XP), uma rara

síndrome associada a mutações germinativas em genes NER e que predispõe pacientes ao desenvolvimento de neoplasias de pele (Friedberg, 2001; Garfinkel et al., 2002; Lengauer et al., 1998).

Como discutido na seção *introdução*, a instabilidade genômica ao nível dos nucleotídeos serve de modelo para a teoria do *fenótipo mutador* (Loeb, 2001), contudo a ausência de mutações somáticas causalmente associadas ao desenvolvimento de câncer no sistema de reparo NER (Futreal, 2009) enfraquece essa teoria, a qual tem sustentação em um possível aumento nas taxas de mutações somáticas. No trabalho apresentado no *Capitulo IV* intitulado "*On the absence of mutations in nucleotide excision repair genes in sporadic solid tumors*" nós propusemos a existência de um limiar de alteração na rede de manutenção do genoma acima do qual a célula se torna inviável. A abordagem que empregamos para sugerir essa hipótese foi bastante simples, apenas comparamos o número de interações estabelecidas entre os componentes do reparo NER com o número de interações observadas nos genes causalmente associados ao desenvolvimento de câncer. Embora este estudo seja especulativo, ele está fundamentado em trabalhos prévios que sugerem que a conectividade (número de interações) de uma dada proteína é proporcional à essencialidade desta proteína para o funcionamento do sistema em que ela atua (Albert et al., 2000; Goh et al., 2007; Jeong et al., 2001). Em relação aos genes somaticamente mutados catalogados no Censo de Genes Câncer (Futreal, 2009), NER apresentou maior nível de conectividade, e muito acima da média de conexões observadas na rede de manutenção do genoma.

Soma-se a isso um estudo paralelo em que nosso grupo demonstrou a ausência de registrados de mutações germinativas que causam nulidade da expressão de genes NER em humanos (nulidade em homozigose), dado esse que reforça o papel essencial do sistema NER para o desenvolvimento embrionário (Castro et al., 2008). Em outras palavras, alelos com mutações germinativas que causam a disfunção total da expressão de genes NER não

aparecem em homozigose, indicando que há desvio da freqüência mendeliana esperada para esses alelos mutados, possivelmente letais quando em homozigose. Essa possibilidade é corroborada por modelos animais e celulares em que dupla nulidade de genes NER causa letalidade embrionária/perinatal em camundongos (Eppig et al., 2007) e letalidade celular em leveduras (Hirschman et al., 2006). Além disso, mutações somáticas que diminuem a capacidade adaptativa ao nível celular não são mutações com bom potencial carcinogênico, um critério *sine qua non* utilizado para caracterizar genes causalmente associados ao desenvolvimento de câncer (Futreal, 2009).

Se por um lado alterações estruturais em NER podem não conferir vantagens adaptativas para as células de câncer, por outro isso favorece a ocorrência de instabilidade genômica ao nível dos nucleotídeos (Friedberg, 2001; Garfinkel et al., 2002; Lengauer et al., 1998). Então, apesar de alterações estruturais em NER poderem ultrapassar o limiar de viabilidade celular, como demonstrado em levedura, as subletais podem acelerar a progressão tumoral por aumentarem a taxa de mutação, algo previsto pela teoria do *fenótipo mutador* (Bielas et al., 2006; Loeb, 2001). Nosso trabalho aponta evidências de que alterações funcionais em NER, detectadas pela alteração do padrão expressão gênica, poderiam contribuir para o desenvolvimento deste fenótipo.

Apesar das evidências serem indiretas, pois não sabemos a natureza destas alterações nem observamos seus efeitos, nosso trabalho sugere um cenário em que modificações epigenéticas se enquadram para explicar a disfunção do sistema de reparo do ADN ao nível dos nucleotídeos. Nossos resultados também indicam que a rede de genes de apoptose está funcionalmente alterada – com aumento da diversidade de expressão global e diminuição da abundância de transcritos. Além disso, nosso modelo de rede de genes também ilustra a grande sobreposição funcional entre todos os sistemas de reparo do ADN – algo bastante reconhecido (Friedberg, 2001; Wood et al., 2005).

Contudo, é importante perceber que a análise desta rede, quando limitada às subdivisões consensuais de cada sistema, pode não captar certos aspectos da diversidade de expressão gênica que, de outro modo, poderiam ser mais informativos (*e.g.* outras subdivisões da rede de manutenção do genoma poderiam incluir intersecções entre NER e MMR, apoptose e BER, etc.).

Assim, dada a sobreposição funcional dos sistemas de manutenção do genoma, nós desenvolvemos um novo *software* de bioinformática na tentativa de não limitar a análise de expressão às subdivisões consensuais. Esta ferramenta incorpora alguns métodos de bioinformática utilizados previamente e apresenta como novidade a capacidade de projetar transcriptomas (*i.e.* dados de expressão em escala genômica) sobre grafos de interação para gerar planícies de expressão gênica (*i.e.* visualização 3D de transcriptomas). O artigo que descreve o programa está apresentado no *Capítulo V* intitulado "*ViaComplex: software for landscape analysis of genome maintenance mechanisms*" e o pacote de instalação está disponível no endereço http://lief.if.ufrgs.br/pub/biosoftwares/viacomplex/.

Embora este trabalho esteja focado na descrição do programa, ele apresenta um resultado interessante ao demonstrar sua habilidade em diferenciar padrões de expressão gênica, uma vez que o exemplo usado compara células tratadas e não tratadas com 5-aza-2'-desoxicitidina (5-aza), um agente quimioterápico que causa hipometilação do ADN (*Capítulo V*, Figura 1F). Ainda que novos trabalhos devam ser conduzidos para avaliar a robusteza da abordagem, o resultado geral sugere que a alternância de estados hipo/hipermetilado causa modificações na planície de expressão gênica de tal forma que NER e apoptose respondem em concordância e de maneira oposta aos demais sistemas. Neste cenário, há um padrão de alteração global no funcionamento dos sistemas de manutenção do genoma em que a supressão de NER e apoptose por hipermetilação está corroborando os resultados apresentados no artigo do *Capítulo III*, uma hipótese razoável considerando que a

hipermetilação de ilhas CpG é descrita como causa de silenciamento de genes supressores tumorais (Feinberg, 2004). Se a coordenação de expressão gênica entre NER e apoptose for de fato consistente (e oposta aos demais sistemas), então talvez tenhamos no futuro um meio de avaliar essa idéia em bases estatísticas.

Desenvolvimento de novos biomarcadores tumorais

Uma das motivações nos objetivos traçados nesta tese está necessidade em desenvolver testes moleculares capazes de lidar com a grande heterogeneidade intra e intertumoral, uma reconhecida barreira para o desenvolvimento de biomarcadores de progressão tumoral (Kulasingam et al., 2008; Ransohoff, 2005). Portanto, desenvolvimento de biomarcadores e caracterização da diversidade tumoral são objetivos interligados.

Embora o estadiamento tumoral, via de regra, não seja realizado com auxilio de biomarcadores (Greene et al., 2002), novas tecnologias em escala genômica tem sido consideradas promissoras para contornar este quadro (Kulasingam et al., 2008). Estratégias que empregam microarranjos de ADN complementar, por exemplo, avaliam simultaneamente múltiplos candidatos em busca de alguns biomarcadores (Bild et al., 2006), o que exige controles estatísticos específicos para diminuir a ocorrência de falsos positivos, como o controle das taxas de falsas descobertas – FDR (Benjamini & Hochberg, 1995). Apesar disso, a lista de potenciais biomarcadores em geral é extensa, os biomarcadores não são relacionados (pois atuam nas mais variadas funções) e a reprodutibilidade em coortes de validação é muito baixa (Ransohoff, 2005).

Na tentativa de aplicar as ferramentas de bioinformática que desenvolvemos previamente nós invertemos a abordagem usualmente empregada em estudos de microarranjo. Este trabalho está apresentado no *Capítulo VI* intitulado "*CFL1 gene is a biomarker for patient outcome and drug resistance in non-small cell lung cancer*". A estratégia consistiu na

caracterização de um gene específico (*CFL1*, cofilina-1; isoforma não-muscular) e no aprimoramento de seu poder discriminatório através do estudo conjunto dos componentes de sua rede de interação protéica. Escolhemos estudar o gene *CFL1* em câncer de pulmão de células não pequenas (NSCLC) por possuirmos meios para este fim e porque a proteína cofilina está envolvida em processos de adesão e invasão celular, um fenótipo essencial para a progressão tumoral (Ghosh et al., 2004; Wang et al., 2007). Por exemplo, de um lado desenvolvemos uma abordagem experimental para estudos de quimiorresistência usando seis linhagens de células do tipo NSCLC e, de outro, utilizamos uma abordagem retrospectiva para estudo clínico usando dados de uma coorte de 111 pacientes de câncer de pulmão que tiveram o perfil de expressão gênica mapeado por microarranjos de ADN complementar (Bild et al., 2006).

Os principais resultados da caracterização do gene *CFL1* são: (i) a análise das curvas de mortalidade mostrou que o biomarcador apresenta bom potencial prognóstico, com melhor sensibilidade/especificidade para pacientes em estágio inicial da doença; (ii) os resultados puderam ser reproduzidos em uma segunda coorte não relacionada; (iii) linhagens celulares NSCLC tratadas com agentes alquilantes são mais resistentes quando expressam alto imunoconteúdo de cofilina, sugerindo um potencial preditivo deste biomarcador no tratamento do câncer de pulmão e (iv) em relação à rede de interação protéica, quatro outros componentes seguiram o mesmo perfil de quimiorresistência da cofilina.

Uma vez que este trabalho apresenta evidências clínicas e experimentais sobre potencial prognóstico/preditivo de um biomarcador ainda não descrito, nós propusemos um novo método para a estratificação do câncer de pulmão de células não pequenas com base na expressão do gene *CFL1* (pedido de depósito de patente PI0802917-2 em anexo). Mesmo assumindo que a capacidade deste biomarcador em lidar com o problema da heterogeneidade tumoral não seja melhor que a capacidade dos muitos biomarcadores já propostos, nossa

intenção é propor o uso conjunto de sua de rede de interação protéica de modo a aperfeiçoar o potencial do biomarcador. O fato de quatro outros genes da rede de cofilina apresentarem o mesmo perfil de quimiorresistência sugere que tal objetivo pode ser viável. A seqüência natural deste estudo será o uso do software VIACOMPLEX para investigar o perfil de expressão gênica do modelo de rede proposto, a exemplo do que foi feito com a de rede de genes de manutenção do genoma.

CONCLUSÕES

A heterogeneidade das neoplasias sólidas ocorre em vários níveis, com acúmulo de múltiplas alterações. A partir do estudo dos padrões de acúmulo destas alterações desenvolvemos novos métodos para caracterizar a diversidade citogenética e molecular, bem como descrevemos alterações nos mecanismos envolvidos na estabilização do genoma capazes de explicar alguns aspectos da progressão tumoral. Estes estudos nos permitiram formular as seguintes conclusões:

Capítulos I e II

- A diversidade citogenética pode ser quantificada pela entropia de Shannon;
- O índice de diversidade *H'* reflete o grau de malignidade dos tumores avaliados;
- Funções complementares de entropia (entropia global e suas contribuições parciais) indicam que o grau de espalhamento das aberrações cromossômicas é específico para cada tipo de tumor;
- O grau de espalhamento das aberrações cromossômicas está correlacionado com a malignidade (*i.e.* tipos de tumores com menor espalhamento amostral apresentam estatísticas populacionais com a melhor sobrevida);
- O valor normalizado da entropia global sugere que o padrão da dinâmica de formação das aberrações cromossômicas não é nem determinístico nem completamente aleatório, mas situa-se entre estas duas fronteiras.

Capítulos III e IV

- A diversidade de expressão gênica também pode ser quantifica pela entropia de Shannon;
- Redes de genes de apoptose e do sistema de reparo por excisão de nucleotídeos estão funcionalmente alteradas no painel de neoplasias sólidas analisado, com aumento da diversidade de expressão gênica e diminuição da abundância de transcritos;

- A rede de genes NER apresenta uma topologia com alto grau de conectividade, compatível com sistemas essenciais de baixa tolerância a mutações. Trabalhos paralelos suportam esta hipótese tanto em humanos como em modelos animais e celulares.

Capítulo V

- O modelo de rede dos mecanismos de manutenção do genoma serviu de base para o desenvolvimento de novos métodos computacionais capazes de mapear o padrão de expressão gênica em escala genômica.

Capítulo VI

- O gene *CFL1* apresentou bom potencial prognóstico para o câncer de pulmão de células não pequenas, com melhor sensibilidade/especificidade para pacientes em estágio inicial da doença;
- Os métodos computacionais desenvolvidos sugerem que é possível aumentar a robusteza deste potencial biomarcador considerando sua rede de interação protéica, pois outros componentes da rede seguiram o mesmo perfil de quimiorresistência.

***Capítulo VII* / anexos**

- A padronização dos métodos computacionais desenvolvidos ao longo desta tese resultou no depósito de duas patentes de invenção (aplicadas ao diagnóstico/prognóstico de neoplasias sólidas) e em dois produtos tecnológicos na forma de programas de computador (aplicados ao estudo da diversidade citogenética e molecular).

PERSPECTIVAS

Capítulos I e II

- Estudar a diversidade citogenética intratumoral com o uso técnicas de hibridização *in situ* por fluorescência.

Capítulos III e IV

- Estudar a diversidade molecular de outras redes de expressão gênica envolvidas na progressão tumoral (*e.g.* genes envolvidos em adesão e invasão tumoral).

Capítulo V

- Mapear o padrão de expressão gênica dos mecanismos de manutenção do genoma com uso do programa VIACOMPLEX.

Capítulo VI

- Estudar a rede de interação protéica do gene *CFL1* com objetivo de aumentar o potencial prognóstico do biomarcador tumoral.
- Estudar o perfil de quimiorresistência desta rede, bem como seu potencial preditivo no tratamento do câncer de pulmão de células não pequenas.

***Capítulo VII* / anexos**

- Aprimorar o protótipo do programa KARYCOMPLEX para testes de campo com usuários.
- Buscar parcerias para dar seguimento aos métodos patenteados.

REFERÊNCIAS BIBLIOGRÁFICAS

ADAMI, C.; OFRIA, C.;COLLIER, T. C. Special Feature: Evolution of biological complexity. **Proc. Natl. Acad. Sci. U.S.A.,** v. 97, n. 9, p. 4463-4468, 2000.

AHLGREN, J. et al. Angiogenesis in invasive breast carcinoma--a prospective study of tumour heterogeneity. **Eur. J. Cancer,** v. 38, n. 1, p. 64-69, 2002.

ALBERT, R.; JEONG, H.;BARABASI, A. L. Error and attack tolerance of complex networks. **Nature,** v. 406, n. 6794, p. 378-382, 2000.

ALBERTSON, D. G. et al. Chromosome aberrations in solid tumors. **Nat. Genet.,** v. 34, n. 4, p. 369-376, 2003.

ANDERSON, A. R. A. ; QUARANTA, V. Integrative mathematical oncology. **Nat. Rev. Cancer,** v. 8, n. 3, p. 227-234, 2008.

AQUILINA, G. ; BIGNAMI, M. Mismatch repair in correction of replication errors and processing of DNA damage. **J. Cell. Physiol.,** v. 187, n. 2, p. 145-154, 2001.

ARMITAGE, P. ; DOLL, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. **Br. J. Cancer,** v. 8, n. 1, p. 1-12, 1954.

Atlas de Mortalidade por Câncer. Instituto Nacional de Câncer (INCA). Disponível em: <http://mortalidade.inca.gov.br/> Acesso em: 4 fevereiro. 2009.

BACKVALL, H. et al. Genetic tumor archeology: microdissection and genetic heterogeneity in squamous and basal cell carcinoma. **Mutat. Res.,** v. 571, n. 1-2, p. 65-79, 2005.

BENJAMINI, Y. ; HOCHBERG, Y. Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. **J. R. Stat. Soc.,** v. 57, n. 1, p. 289-300, 1995.

BENNETT, J. M. et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. **Br. J. Haematol,** v. 33, n. 4, p. 451-458, 1976. Disponível em: PMID: 188440.

BERCZI, C. et al. Prognostic value of DNA ploidy status in patients with rectal cancer. **Anticancer Res.,** v. 22, n. 6B, p. 3737-3741, 2002.

BIELAS, J. H. ; LOEB, L. A. Quantification of random genomic mutations. **Nat. Met.,** v. 2, n. 4, p. 285-290, 2005.

BIELAS, J. H. et al. Human cancers express a mutator phenotype. **Proc. Natl. Acad. Sci. U.S.A.,** v. 103, n. 48, p. 18238-18242, 2006.

BILD, A. H. et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. **Nature,** v. 439, n. 7074, p. 353-357, 2006.

BREIVIK, J. Don't stop for repairs in a war zone: Darwinian evolution unites genes and environment in cancer development. **Proc. Natl. Acad. Sci. U.S.A.,** v. 98, n. 10, p. 5379-5381, 2001.

BREIVIK, J. ; GAUDERNACK, G. Resolving the evolutionary paradox of genetic instability: a cost-benefit analysis of DNA repair in changing environments. **FEBS Lett.,** v. 563, n. 1-3, p. 7-12, 2004.

BRITTEN, R. A. et al. Intratumoral heterogeneity as a confounding factor in clonogenic assays for tumour radioresponsiveness. **Radiother. Oncol,** v. 39, n. 2, p. 145-153, 1996.

CAHILL, D. P. et al. Genetic instability and darwinian selection in tumours. **Trends Cell Biol.,** v. 9, n. 12, p. M57-M60, 1999.

CASTRO, M. A. A. et al. Evolutionary origins of human apoptosis and genome-stability gene networks. **Nucleic Acids Res.,** v. 36, n. 19, p. 6269-6283, 2008.

CASTRO, M. A. A. et al. Impaired expression of NER gene network in sporadic solid tumors. **Nucleic Acids Res.,** v. 35, n. 6, p. 1859-1867, 2007. Disponível em: doi:10.1093/nar/gkm061.

CEOL, C. J.; PELLMAN, D.;ZON, L. I. APC and colon cancer: two hits for one. **Nat Med,** v. 13, n. 11, p. 1286-1287, 2007.

COTRAN, R.; COLLINS, K. V.; ROBBINS, T. **Robbins: Patologia Estrutural e Funcional**. Rio de Janeiro: Guanabara Koogan, 2000.

DATASUS. Produção Ambulatorial do SUS - 2008 - Brasil. Disponível em: <http://www.datasus.gov.br> Acesso em: 12 de março. 2009.

DENG, Y.; CHAN, S. S.;CHANG, S. Telomere dysfunction and tumour suppression: the senescence connection. **Nat. Rev. Cancer,** v. 8, n. 6, p. 450-458, 2008.

DESROCHERS, R. E. ; ANAND, M. Traditional Diversity Indices to Taxonomic Diversity Indices. **Int. J. Ecol. Environ. Sci.,** v. 30, p. 85-92, 2004.

DUESBERG, P. et al. The chromosomal basis of cancer. **Cell. Oncol.,** v. 27, n. 5-6, p. 293-318, 2005.

DUESBERG, P.; STINDL, R.;HEHLMANN, R. Explaining the high mutation rates of cancer cells to drug and multidrug resistance by chromosome reassortments that are catalyzed by aneuploidy. **Proc. Natl. Acad. Sci. U.S.A.,** v. 97, n. 26, p. 14295-14300, 2000.

EPPIG, J. T. et al. The mouse genome database (MGD): new features facilitating a model system. **Nucleic Acids Res.,** v. 35, n. suppl_1, p. D630-D637, 2007.

FABARIUS, A.; HEHLMANN, R.;DUESBERG, P. H. Instability of chromosome structure in cancer cells increases exponentially with degrees of aneuploidy. **Cancer Genet. Cytogenet.,** v. 143, n. 1, p. 59-72, 2003.

FEARON, E. R. ; VOGELSTEIN, B. A genetic model for colorectal tumorigenesis. **Cell,** v. 61, n. 5, p. 759-767, 1990.

FEINBERG, A. P. The epigenetics of cancer etiology. **Semin. Cancer Biol.,** v. 14, n. 6, p. 427-432, 2004.

FEINBERG, A. P.; OHLSSON, R.;HENIKOFF, S. The epigenetic progenitor origin of human cancer. **Nat. Rev. Genetics,** v. 7, n. 1, p. 21-33, 2006.

FIDLER, I. J. Tumor Heterogeneity and the Biology of Cancer Invasion and Metastasis. **Cancer Res.,** v. 38, n. 9, p. 2651-2660, 1978.

FLYGER, H. L. et al. DNA ploidy in colorectal cancer, heterogeneity within and between tumors and relation to survival. **Cytometry,** v. 38, n. 6, p. 293-300, 1999.

FODDE, R.; SMITS, R.;CLEVERS, H. APC, signal transduction and genetic instability in colorectal cancer. **Nat. Rev. Cancer,** v. 1, n. 1, p. 55-67, 2001.

FORBES, S. A. et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). **Curr. Protoc. Hum. Genet,** v. Chapter 10, p. Unit2008.

FRANK, S. A. **Dynamics of cancer: incidence, inheritance and evolution**. Princeton: Princeton University Press, 2007.

FRIEDBERG, E. C. How nucleotide excision repair protects against cancer. **Nat. Rev. Cancer,** v. 1, n. 1, p. 22-33, 2001.

FUHRMAN, S. et al. The application of Shannon entropy in the identification of putative drug targets. **Biosystems,** v. 55, n. 1-3, p. 5-14, 2000.

Futreal, P. A. The Cancer Gene Census. Disponível em: <http://www.sanger.ac.uk/genetics/CGP/Census/> Acesso em: 21 de fevereiro. 2009.

FUTREAL, P. A. et al. A CENSUS OF HUMAN CANCER GENES. **Nat. Rev. Cancer,** v. 4, n. 3, p. 177-183, 2004.

GARFINKEL, D. J. ; BAILIS, A. M. Nucleotide Excision Repair, Genome Stability, and Human Disease: New Insight from Model Systems. **J. Biomed. Biotechnol.,** v. 2, n. 2, p. 55-60, 2002.

GATENBY, R. A. ; VINCENT, T. L. An evolutionary model of carcinogenesis. **Cancer Res.,** v. 63, n. 19, p. 6212-6220, 2003.

GATENBY, R. A. ; FRIEDEN, B. R. Application of Information Theory and Extreme Physical Information to Carcinogenesis. **Cancer Res.,** v. 62, n. 13, p. 3675-3684, 2002.

GHOSH, M. et al. Cofilin Promotes Actin Polymerization and Defines the Direction of Cell Motility. **Science,** v. 304, n. 5671, p. 743-746, 2004.

GLOCKNER, S. et al. Marked intratumoral heterogeneity of c-myc and CyclinD1 but not of c-erbB2 amplification in breast cancer. **Lab. Invest.,** v. 82, n. 10, p. 1419-1426, 2002.

GOH, K. I. et al. The human disease network. **Proc. Natl. Acad. Sci. U.S.A.,** v. 104, n. 21, p. 8685-8690, 2007.

GONZALEZ-GARCIA, I.; SOLE, R. V.;COSTA, J. Metapopulation dynamics and spatial heterogeneity in cancer. **Proc. Natl. Acad. Sci. U.S.A.,** v. 99, n. 20, p. 13085-13089, 2002.

GREENE, F. L. et al. **AJCC Cancer Staging Manual**. A. Fritz. 6.ed. New York: Springer, 2002.

GREENMAN, C. et al. Patterns of somatic mutation in human cancer genomes. **Nature,** v. 446, n. 7132, p. 153-158, 2007.

GRIFFITHS, A. J. S. et al. **Modern Genetic Analysis**. 2a. ed. New York: W. H. Freeman, 2002.

HAHN, W. C. ; WEINBERG, R. A. Rules for Making Human Tumor Cells. **N. Engl. J. Med.,** v. 347, n. 20, p. 1593-1603, 2002.

HALAZONETIS, T. D.; GORGOULIS, V. G.;BARTEK, J. An Oncogene-Induced DNA Damage Model for Cancer Development. **Science,** v. 319, n. 5868, p. 1352-1355, 2008.

HARADA, K. et al. Intratumoral cytogenetic heterogeneity detected by comparative genomic hybridization and laser scanning cytometry in human gliomas. **Cancer Res.,** v. 58, n. 20, p. 4694-4700, 1998.

HARRIS, N. L. et al. A revised European-American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. **Blood,** v. 84, n. 5, p. 1361-1392, 1994.

HARRIS, N. L. et al. World Health Organization Classification of Neoplastic Diseases of the Hematopoietic and Lymphoid Tissues: Report of the Clinical Advisory Committee MeetingAirlie House, Virginia, November 1997. **J. Clin. Oncol.,** v. 17, n. 12, p. 3835-3849, 1999.

HENG, H. H. et al. Clonal and non-clonal chromosome aberrations and genome variation and aberration. **Genome,** v. 49, n. 3, p. 195-204, 2006.

HIRSCHMAN, J. E. et al. Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. **Nucleic Acids Res.,** v. 34, n. suppl_1, p. D442-D445, 2006.

HOEIJMAKERS, J. H. J. Genome maintenance mechanisms for preventing cancer. **Nature,** v. 411, n. 6835, p. 366-374, 2001.

HOGLUND, M. et al. Statistical behavior of complex cancer karyotypes. **Genes Chromosomes Cancer,** v. 42, n. 4, p. 327-341, 2005.

HOGLUND, M. et al. Dissecting karyotypic patterns in colorectal tumors: two distinct but overlapping pathways in the adenoma-carcinoma transition. **Cancer Res.,** v. 62, n. 20, p. 5939-5946, 2002a.

HOGLUND, M. et al. Coping with complexity. multivariate analysis of tumor karyotypes. **Cancer Genet. Cytogenet.,** v. 135, n. 2, p. 103-109, 2002b.

HOGLUND, M. et al. Identification of cytogenetic subgroups and karyotypic pathways in transitional cell carcinoma. **Cancer Res.,** v. 61, n. 22, p. 8241-8246, 2001.

HUMPHRIES, A. ; WRIGHT, N. A. Colonic crypt organization and tumorigenesis. **Nat. Rev. Cancer,** v. 8, n. 6, p. 415-424, 2008.

ICD-6. **International Statistical Classification of Diseases, Injuries, and Causes of Death.** World Health Organization. 6.ed. rev. Geneva:1948.

ICD-7. **International Statistical Classification of Diseases, Injuries and Causes of Death.** 7.ed. rev. Geneva: World Health Organization, 1955.

ICD-O-1. **International Classification of Disease for Oncology**. 1.ed. Geneva: World Health Organization, 1976.

ICD-O-3. **International Classification of Diseases for Oncology**. 3.ed. Geneva: World Health Organization, 2000.

ISCN. **ISCN (1995): An international system for human cytogenetic nomenclature.** F. Mitelman. Basel: S Karger, 1995.

JALLEPALLI, P. V. ; LENGAUER, C. Chromosome segregation and cancer: Cutting through the mystery. **Nat. Rev. Cancer,** v. 1, n. 2, p. 109-117, 2001.

JASS, J. R.; KELLOFF, G. J.;SCHILSKY, R. L. Limitations of the Adenoma-Carcinoma Sequence in Colorectum. **Clin. Cancer Res.,** v. 10, n. 17, p. 5969-5970, 2004.

JEFFORD, C. E. ; IRMINGER-FINGER, I. Mechanisms of chromosome instability in cancers. **Crit. Rev. Oncol. Hematol.,** v. 59, n. 1, p. 1-14, 2006.

JEONG, H. et al. Lethality and centrality in protein networks. **Nature,** v. 411, n. 6833, p. 41-42, 2001.

JIN, C. et al. Karyotypic heterogeneity and clonal evolution in squamous cell carcinomas of the head and neck. **Cancer Genet. Cytogenet.,** v. 132, n. 2, p. 85-96, 2002.

JIRICNY, J. The multifaceted mismatch-repair system. **Nat. Rev. Mol. Cell Biol.,** v. 7, n. 5, p. 335-346, 2006.

JONES, S. et al. Comparative lesion sequencing provides insights into tumor evolution. **Proc. Natl. Acad. Sci. U.S.A.,** v. 105, n. 11, p. 4283-4288, 2008.

KENDAL, W. S. The use of information theory to analyze genomic changes in neoplasia. **Math. Biosci.,** v. 100, n. 2, p. 143-159, 1990.

KINZLER, K. W. ; VOGELSTEIN, B. Lessons from hereditary colorectal cancer. **Cell,** v. 87, n. 2, p. 159-170, 1996.

KLEIN, C. A. Random mutations, selected mutations: A PIN opens the door to new genetic landscapes. **Proc. Natl. Acad. Sci. U.S.A.,** v. 103, n. 48, p. 18033-18034, 2006.

KUFE, D. et al. Cardinal Manifestations of Cancer. In: D.KUFE; R. BAST; W. HAIT; W. HONG; R. POLLACK; R. WEICHSELBAUM; J. HOLLAND; E. FREI (Eds.). **Cancer Medicine.** 2005. p. ISBN 1-55009-307-X.

KULASINGAM, V. ; DIAMANDIS, E. P. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. **Nat. Clin. Prac. Oncol.,** v. 5, n. 10, p. 588-599, 2008.

LASH, A. E. et al. SAGEmap: A Public Gene Expression Resource. **Genome Res.,** v. 10, n. 7, p. 1051-1060, 2000.

LEERS, M. P. G. ; NAP, M. Steroid Receptor Heterogeneity in Relation to DNA Index in Breast Cancer: A Multiparameter Flow Cytometric Approach on Paraffin-Embedded Tumor Samples. **Breast J.,** v. 7, n. 4, p. 249-259, 2001.

LENGAUER, C.; KINZLER, K. W.;VOGELSTEIN, B. Genetic instability in colorectal cancers. **Nature,** v. 386, n. 6625, p. 623-627, 1997.

_____. Genetic instabilities in human cancers. **Nature,** v. 396, n. 6712, p. 643-649, 1998.

LI, R. et al. Aneuploidy vs. gene mutation hypothesis of cancer: Recent study claims mutation but is found to support aneuploidy. **Proc. Natl. Acad. Sci. U.S.A.,** v. 97, n. 7, p. 3236-3241, 2000.

LOEB, L. A.; LOEB, K. R.;ANDERSON, J. P. Multiple mutations and cancer. **Proc. Natl. Acad. Sci. U.S.A.,** v. 100, n. 3, p. 776-781, 2003.

LOEB, L. A. A Mutator Phenotype in Cancer. **Cancer Res.,** v. 61, n. 8, p. 3230-3239, 2001.

LOSI, L. et al. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. **Carcinogenesis,** v. 26, n. 5, p. 916-922, 2005.

LUDWIG, J. A. ; WEINSTEIN, J. N. Biomarkers in Cancer Staging, Prognosis and Treatment Selection. **Nat. Rev. Cancer,** v. 5, n. 11, p. 845-856, 2005.

LUEBECK, E. G. ; MOOLGAVKAR, S. H. Multistage carcinogenesis and the incidence of colorectal cancer. **Proc. Natl. Acad. Sci. U.S.A.,** v. 99, n. 23, p. 15095-15100, 2002.

MALEY, C. C. et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. **Nat. Genet.,** v. 38, n. 4, p. 468-473, 2006.

MARX, J. Debate surges over the origins of genomic defects in cancer. **Science,** v. 297, n. 5581, p. 544-546, 2002.

MERLO, L. M. F. et al. Cancer as an evolutionary and ecological process. **Nat. Rev. Cancer,** v. 6, n. 12, p. 924-935, 2006.

MITELMAN DATABASE **Mitelman Database of Chromosome Aberrations in Cancer**: http://cgap.nci.nih.gov/Chromosomes/Mitelman., 2009

MITELMAN, F. Recurrent chromosome aberrations in cancer. **Mutat. Res.,** v. 462, n. 2-3, p. 247-253, 2000.

MITELMAN, F. et al. Clinical significance of cytogenetic findings in solid tumors. **Cancer Genet. Cytogenet.,** v. 95, n. 1, p. 1-8, 1997.

MURNANE, J. P. Telomeres and chromosome instability. **DNA Repair,** v. 5, n. 9-10, p. 1082-1092, 2006.

NOWAK, M. A. et al. Evolutionary dynamics of tumor suppressor gene inactivation. **Proc. Natl. Acad. Sci. U.S.A.,** v. 101, n. 29, p. 10635-10638, 2004.

ODA, S. et al. Two modes of microsatellite instability in human cancer: differential connection of defective DNA mismatch repair to dinucleotide repeat instability. **Nucleic Acids Res.,** v. 33, n. 5, p. 1628-1636, 2005.

OGINO, S. ; GOEL, A. Molecular Classification and Correlates in Colorectal Cancer. **J. Mol. Diagn.,** v. 10, n. 1, p. 13-27, 2008.

ORNDAL, C. et al. Cytogenetic intratumor heterogeneity in soft tissue tumors. **Cancer Genet. Cytogenet.,** v. 78, n. 2, p. 127-137, 1994.

OSTERHELD, M. C. et al. Evaluation of heterogeneity of DNA ploidy in early gastric cancers. **Anal.Cell. Pathol.,** v. 19, n. 2, p. 67-72, 1999.

PELTOMAKI, P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. **Hum. Mol. Genet.,** v. 10, n. 7, p. 735-740, 2001.

QUIRKE, P. et al. The future of the TNM staging system in colorectal cancer: time for a debate? **Lancet Oncol.,** v. 8, n. 7, p. 651-657, 2007.

RAJAGOPALAN, H. et al. The significance of unstable chromosomes in colorectal cancer. **Nat. Rev. Cancer,** v. 3, n. 9, p. 695-701, 2003.

RANSOHOFF, D. F. Bias as a threat to the validity of cancer molecular-marker research. **Nat. Rev. Cancer,** v. 5, n. 2, p. 142-149, 2005.

RAPPAPORT, H. et al. Report of the Committee on Histopathological Criteria Contributing to Stagin of Hodgkin's Disease. **Cancer Res.,** v. 31, n. 11, p. 1864-1865, 1971. Disponível em: PMID:5121694.

RICKE, R. M.; VAN REE, J. H.;VAN DEURSEN, J. M. Whole chromosome instability and cancer: a complex relationship. **Trends Genet.,** v. 24, n. 9, p. 457-466, 2008.

RICOTTA, C. On parametric evenness measures. **J. Theor. Biol.,** v. 222, n. 2, p. 189-197, 2003.

RICOTTA, C. ; SZEIDL, L. Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. **Theor. Popul. Biol.,** v. 70, n. 3, p. 237-243, 2006.

Ries, L. A. G., Melbert, D., Krapcho, M., Stinchcomb, D. G., Howlader, N., Horner, M. J. et al. SEER Cancer Statistics Review, 1975-2005, National Cancer Institute. Disponível em: <http://seer.cancer.gov/csr/1975_2005/> Acesso em: 4 fevereiro. 2009.

ROUDIER, M. P. et al. Phenotypic heterogeneity of end-stage prostate carcinoma metastatic to bone. **Hum. Pathol.,** v. 34, n. 7, p. 646-653, 2003.

RUSAN, N. M. ; PEIFER, M. Original CIN: reviewing roles for APC in chromosome instability. **J. Cell Biol.,** v. 181, n. 5, p. 719-726, 2008.

SAWAN, C. et al. Epigenetic drivers and genetic passengers on the road to cancer. **Mutat. Res.,** v. 642, n. 1-2, p. 1-13, 2008.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell Systems Tech. J.,** v. 27, n. 3, p. 379-423, 1948.

SHEN, L. et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. **Proc. Natl. Acad. Sci. U.S.A.,** v. 104, n. 47, p. 18654-18659, 2007.

SHIBATA, D. Clonal diversity in tumor progression. **Nat Genet.,** v. 38, n. 4, p. 402-403, 2006.

SHIOMI, H. et al. Cytogenetic heterogeneity and progression of esophageal squamous cell carcinoma. **Cancer Genet. Cytogenet.,** v. 147, n. 1, p. 50-61, 2003.

SIEBER, O. M.; HEINIMANN, K.;TOMLINSON, I. P. Genomic instability--the engine of tumorigenesis? **Nat. Rev. Cancer,** v. 3, n. 9, p. 701-708, 2003.

SJOBLOM, T. et al. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. **Science,** v. 314, n. 5797, p. 268-274, 2006.

SKLAREW, R. J.; BODMER, S. C.;PERTSCHUK, L. P. Comparison of Microscopic Imaging Strategies for Evaluating Immunocytochemical (Pap) Steroid-Receptor Heterogeneity. **Cytometry,** v. 12, n. 3, p. 207-220, 1991.

SMITH, G. et al. Mutations in APC, Kirsten-ras, and p53ΓÇöalternative genetic pathways to colorectal cancer. **Proc. Natl. Acad. Sci. U.S.A.,** v. 99, n. 14, p. 9433-9438, 2002.

SOREIDE, K. et al. Microsatellite instability in colorectal cancer. **Br. J. Surg.,** v. 93, n. 4, p. 395-406, 2006.

SPREMULLI, E. N. ; DEXTER, D. L. Human tumor cell heterogeneity and metastasis. **J. Clin. Oncol.,** v. 1, n. 8, p. 496-509, 1983.

VENKATESAN, R. N.; BIELAS, J. H.;LOEB, L. A. Generation of mutator mutants during carcinogenesis. **DNA Repair,** v. 5, n. 3, p. 294-302, 2006.

VOGELSTEIN, B. ; KINZLER, K. W. Cancer genes and the pathways they control. **Nat. Med.,** v. 10, n. 8, p. 789-799, 2004.

WANG, T. L. et al. Prevalence of somatic alterations in the colorectal cancer cell genome. **Proc. Natl. Acad. Sci. U.S.A.,** v. 99, n. 5, p. 3076-3080, 2002.

WANG, W.; EDDY, R.;CONDEELIS, J. The cofilin pathway in breast cancer invasion and metastasis. **Nat. Rev. Cancer,** v. 7, n. 6, p. 429-440, 2007.

WANG, Z. H. et al. Three classes of genes mutated in colorectal cancers with chromosomal instability. **Cancer Res.,** v. 64, n. 9, p. 2998-3001, 2004.

WOOD, L. D. et al. The Genomic Landscapes of Human Breast and Colorectal Cancers. **Science,** v. 318, n. 5853, p. 1108-1113, 2007.

WOOD, R. D.; MITCHELL, M.;LINDAHL, T. Human DNA repair genes, 2005. **Mutat. Res.,** v. 577, n. 1-2, p. 275-283, 2005.

YUNG, W. K.; SHAPIRO, J. R.;SHAPIRO, W. R. Heterogeneous Chemosensitivities of Subpopulations of Human Glioma Cells in Culture. **Cancer Res.,** v. 42, n. 3, p. 992-998, 1982.

ZHIVOTOVSKY, B. ; KROEMER, G. Apoptosis and genomic instability. **Nat. Rev. Mol. Cell Biol.,** v. 5, n. 9, p. 752-762, 2004.

ANEXO I: PEDIDO DE DEPÓSITO DE PATENTE

Método e kit diagnóstico e/ou prognóstico de câncer de pulmão (PI0802917-2).

Espaço reservado para protocolo

DEPÓSITO DE PEDIDO DE PATENTE OU DE CERTIFICADO DE ADIÇÃO

Ao Instituto Nacional da Propriedade Industrial:
O requerente solicita a concessão de um privilégio na natureza e nas condições abaixo indicadas:

1. Depositante (71):

1.1 Nome: UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

1.2 CNPJ/CPF: 92969856000198

1.3 Endereço completo: AV. PAULO GAMA, 110 CENTRO PORTO ALEGRE RS BRASIL

1.4 CEP: 90040-060

1.5 Telefone: ( ) :51 3308 4232 1.6 FAX: ( ) 51 3308 4237

1.7 Email: SEDETEC@UFRGS.BR

☐ continua em folha anexa

2. Natureza: ☒ Invenção ☐ Modelo de Utilidade ☐ Certificado de Adição

Escreva, obrigatoriamente, e por extenso, a Natureza desejada: Patente de Invenção

3. Título da Invenção, do Modelo de Utilidade ou do Certificado de Adição (54):

Método e Kit de Diagnóstico e/ou Prognóstico de Câncer de Pulmão

☐ continua em folha anexa

4. Pedido de Divisão: do pedido Nº : Data de Depósito: 20/06/2008

5. Prioridade: ☐ interna ☐ unionista

O depositante reinvindica a(s) seguintes(s) prioridade(s)

País ou organização de origem	Número do depósito	Data do depósito
		/ /
		/ /
		/ /

6. Inventor (72):

☐ Assinale aqui se o(s) mesmo(s) requer(em) a não divulgação de seu(s) nome(s)

6.1 Nome: Fábio Klamt

6.2 Qualificação: DOCENTE 6.3 CPF: [redacted]

6.4 Endereço completo: [redacted]

6.5 CEP: 90420130

6.6 Telefone: ( ) 51-81863013 6.7 FAX: ( )

6.8 Email: 00025267@ufrgs.br

☒ continua em folha anexa

7. Declaração na forma do item 3.2 do Ato Normativo nº 127/97:

160

[X] 7.1 Declaro que os dados fornecidos no presente formulário são idênticos ao da certidão de depósito ou documento equivalente do pedido cuja prioridade está sendo reivindicada.

[ ] em anexo

8. Declaração de divulgação anterior não prejudicial (Período de graça):
(art. 12 da LPI e item 2 do AN nº 127/97)

[ ] em anexo

9. Declaração na forma do art. 2º da Resolução/INPI nº 134 de 13/12/06

[ ] 9.1 Declaro que o objeto do presente pedido de patente de invenção foi obtido em decorrência de acesso a amostra de componente do patrimônio genético nacional, realizado a partir de 30 de junho de 2000, e que foram cumpridas as determinações da Medida Provisória 2.186-16, de 23/08/01, informando ainda:

9.2 Número e a data da Autorização do acesso correspondente:

Nº Data: / /

9.3 Origem do material genético e do conhecimento tradicional associado, quando for o caso:

[ ] continua em anexo

[ ] 9.4 Declaro que o objeto do presente pedido de patente de invenção não foi obtido em decorrência de acesso a amostra de componente do patrimônio genético nacional, realizado a partir de 30 de junho de 2000.

10. Procurador (74):

10.1 Nome:

10.2 CPF/CNPJ: 10.3 API/OAB:

10.4 Endereço completo:

10.5 CEP:

10.6 Telefone: ( ) 10.7 FAX: ( )

11. Documentos anexados (assinale e indique também o número de folhas):
(Deverá ser indicado o nº total de somente uma das vias de cada documento)

X	11.1 Guia de recolhimento	1 fls.	X	11.5 Relatório descritivo	15 fls.
	11.2 Procuração	0 fls.	X	11.6 Reivindicações	3 fls.
	11.3 Documentos de prioridade	0 fls.		11.7 Desenhos	0 fls.
	11.4 Doc. de contrato de trabalho	0 fls.	X	11.8 Resumo	1 fls.
X	11.9 Outros (especificar): Portaria de Competência e cópi a DOU.				3 fls.
	11.10 Total de folhas anexadas:				23 fls.

12. Declaro, sob penas da Lei, que todas as informações acima prestadas são completas e verdadeiras

Porto Alegre, 20/06/2008

Local e Data

Assinatura e Carimbo

Prof Ricardo N. Ayup Zouain
Secretário de Desenvolvimento Tecnológico
UFRGS

ANFXO DE INVENTORES

Título: Método e Kit de Diagnóstico e/ou Prognóstico de Câncer de Pulmão **Página : 3**

Nome: MAURO ANTÔNIO ALVES CASTRO
Qualificação: DOUTORANDO
Endereço: Instituto de Ciências Básicas da Saúde - Departamento de Bioquímica Centro Porto Alegre RS
Cep:
Nac: BRASILEIRA

Nome: José Cláudio Fonseca Moreira
Qualificação: DOCENTE
Endereço:
Cep: 91751830
Nac: BRASILEIRA

Nome: Felipe Dal-Pizzol
Qualificação: DOCENTE
Endereço:
Cep: 88800000
Nac: BRASILEIRA

Nome: Emily Shacter
Qualificação: DOCENTE
Endereço:
Cep: 20912
Nac: AMERICANA

ANEXO II: FERRAMENTAS DE BIOINFORMÁTICA

Pacotes de instalação dos programas ViaComplex e KaryoComplex estão disponíveis nos seguintes endereços eletrônicos:

- http://lief.if.ufrgs.br/pub/biosoftwares/viacomplex/

- http://lief.if.ufrgs.br/pub/biosoftwares/karycomplex/

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo