Cancer and genomics

P. Andrew Futreal*, Arek Kasprzyk†, Ewan Birney†, James C. Mullikin‡, Richard Wooster*§ & Michael R. Stratton*§

** Cancer Genome Project and ‡ Informatics Division, Sanger Centre, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK*
† EBI-EMBL, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
§ Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK

Identification of the genes that cause oncogenesis is a central aim of cancer research. We searched the proteins predicted from the draft human genome sequence for paralogues of known tumour suppressor genes, but no novel genes were identified. We then assessed whether it was possible to search directly for oncogenic sequence changes in cancer cells by comparing cancer genome sequences against the draft genome. Apparently chimaeric transcripts (from oncogenic fusion genes generated by chromosomal translocations, the ends of which mapped to different genomic locations) were detected to the same degree in both normal and neoplastic tissues, indicating a significant level of false positives. Our experiment underscores the limited amount and variable quality of DNA sequence from cancer cells that is currently available.

All cancers are caused by abnormalities in DNA sequence. Throughout life, the DNA in human cells is exposed to mutagens and suffers mistakes in replication, resulting in progressive, subtle changes in the DNA sequence in each cell. Occasionally, one of these somatic mutations alters the function of a critical gene, providing a growth advantage to the cell in which it has occurred and resulting in the emergence of an expanded clone derived from this cell. Additional mutations in the relevant target genes, and consequent waves of clonal expansion, produce cells that invade surrounding tissues and metastasize. Cancer is the most common genetic disease: one in three people in the western world develop cancer, and one in five die from it[1].

Around 30 recessive oncogenes (tumour suppressor genes) and more than 100 dominant oncogenes have been identified. In the past, the most successful way to identify such genes was to narrow their location to a small part of the genome using mapping strategies, and then to screen candidate genes in the region for mutations in cancer cases. However, this strategy has its limitations. Mapping information can be confusing or misleading. Moreover, some cancer genes leave no obvious 'identifiers' in the genome and therefore cannot be readily positioned using such maps (for example, dominant oncogenes that are activated by single nucleotide substitutions leading to a single amino-acid change). These mutated genes are essentially invisible to conventional detection and their identification has usually depended upon selection of likely candidates on the basis of the biological features associated with cancer.

How will the genome sequence help us to identify the remaining cancer genes? One possibility is to generate a longer list of plausible candidates by searching for paralogues of known cancer genes. To this end we searched a protein set which represents the current 'best guess' of proteins encoded by the human genome[2] for paralogues of known recessive oncogenes/tumour suppressor genes (see Supplementary Information; Table 1). Although we detected most of the known family members, we found no convincing evidence of new ones at the level of stringency of this search (50% or greater amino-acid identity over a minimum of 50 amino acids). The lack of detectable paralogues could be due to deficiencies of the protein set, despite the fact that new paralogues have been found for many other families[2]. To confirm our results, we ran an exhaustive search of one of the genes, *TP53*, against the total draft genome sequence, considering every possible gene prediction and reading frame[3]. This additional search revealed no unknown paralogues.

The lack of novel paralogues may reflect the biological and medical importance of these gene families; most of their members may have already been found. But additional paralogues could be hiding in unsequenced regions of the genome or may have gone undetected owing to the deficiencies of gene prediction. Clearly, detection of paralogues also depends upon the search criteria used. Predictably, when we lowered the stringency of the search the number of hits increased substantially—but many of these have questionable biological validity.

Was this really a useful way of generating likely candidate cancer genes? Mutated cancer genes do recur in certain gene families (for example, signalling kinases and GDP binding proteins), but these families are often large and, in most cases, only a minority of members are implicated in cancer. In fact, the diversity of structure and function of cancer genes is striking. For example, hardly any of the known recessive oncogenes have strong homology to any other, and their proteins are associated with diverse biological and biochemical functions. Moreover, many close relatives of important cancer-related genes (for example, the genes for p73 and p63, which show sequence similarity to *TP53* (ref. 4)) are not known to be mutated in cancer (although it is possible that they are significantly altered by changes in expression). So we may learn more about the mutations driving cancer if we are not too heavily influenced by past experience. Instead, we should persevere in exploring every gene or protein, whatever its structure or putative function, as a possible candidate.

By simply mining the genome sequence for similarity to known cancer genes, we may miss an opportunity. In addition to many more gene sequences, the working draft contains information concerning the organization of genes, their structures and ordering along the chromosomes. Cancers are characterized by disruption and disorganization of genomes. Perhaps we could use the genome sequence as a template against which we can detect structural alterations of the genome in cancer cells.

To do this we need to compare the working draft (and ultimately the finished sequence) with corresponding sequences from cancer cell genomes. However, there is very little cancer genome sequence available, and what is available is patchy. The largest body of sequence from cancer genomes originates from programs that sample clones in complementary DNA libraries constructed from neoplastic tissues (for example, the Cancer Genome Anatomy Project (CGAP, http://www.ncbi.nlm.nih.gov/CGAP/)). In principle, we could try and compare these with the working draft for the somatic base substitutions and small insertions and deletions that often result in inactivation of tumour suppressor genes or activation of dominantly acting oncogenes. In practice, however, these databases contain relatively little sequence and do not sample most transcripts in any single tumour. Moreover, the available sequence is mostly from untranslated regions of genes (whereas the cancer-causing mutations cluster in the coding regions). Even if the rare, meaningful somatic mutations could be detected, they would be buried in the debris of sequence errors (both in the cDNA libraries and in the genomic sequence) or camouflaged in a forest of innocuous polymorphisms.

We attempted to use these cDNA library sequences in conjunction with the working draft to look for a different type of alteration in cancer. Gene rearrangements that result in activation of oncogenes can arise as a result of chromosomal translocation. This type of abnormality is common in leukaemias, lymphomas and sarcomas[5,6], and often results in the formation of a chimaeric transcript, the product of a fusion gene derived from portions of transcribed genes on either side of the chromosomal breakpoints (although in some instances the translocation simply results in dysregulation of an intact transcript). Intriguingly, this pattern of oncogenesis has not been frequently documented amongst most of the common, adult epithelial cancers. Whether the rearrangements exist but are hidden in the disorganized complexity of epithelial cancer cell genomes or are simply not present in epithelial cancers is a question that may be addressed in the near future.

To look for such gene rearrangements we obtained all the sequences from the CGAP program (derived from cDNA libraries constructed from both normal tissue samples and neoplastic samples) and selected those clones from which two sequences had been obtained (normal, 215,889 clones; cancer, 25,446 clones). Most of these sequences were from either end of the clones. We then looked in the genomic sequence for matches to these paired cDNA sequences that would allow their chromosomal localization (see Supplementary Information).

Most pairs of sequences from the same cDNA clone mapped to the same position in the genome, as would be expected if they had originated from a single normal transcript. But there were a few pairs derived from a single cDNA clone that mapped to two different parts of the genome. Could they represent transcripts of chimaeric genes generated by chromosomal translocation? Possibly, but this strategy has limitations. Even this conceptually simple experiment was dogged by the intrinsic complexities of the genome, such as low-frequency repeats and multiple, high-fidelity copies of some genes. Moreover, our analyses yielded proportionally the same number of apparently chimaeric transcripts derived from normal tissues as from cancers (3% of starting clones in both cases), indicating a significant rate of false positives. These could result from chimaeric clones arising as artefacts of cDNA library construction, mistracking of sequencing gels, errors in annotating and curating databases or misassembly of the draft sequence.

Of course, for this experiment we used a resource that had not

Table 1 Recessive oncogenes used in the search for paralogues

Symbol	Accession	Name	Paralogues identified*	Cancer syndrome	Cancer types (germline and/or somatic mutations)
APC	P25054	Adenomatous polyposis of the colon gene	APC-like gene	Familial polyposis of the colon	Colorectal, pancreatic cancers, desmoids, hepatoblastoma
BRCA1	P38398	Familial breast/ovarian cancer gene 1	None	Hereditary breast/ovarian cancer	Hereditary breast/ovarian cancers
BRCA2	P51587	Familial breast/ovarian cancer gene 2	None	Hereditary breast/ovarian cancer	Hereditary breast/ovarian cancers
CDH1	P12830	Cadherin 1 gene	N-cadherin, P-cadherin, R-cadherin	Familial gastric carcinoma	Lobular breast cancer
CDKN2A	P42771	Cyclin-dependent kinase inhibitor 2A (p16) gene	None	Cutaneous malignant melanoma 2	Melanoma, other tumour types
CYLD	CAB93533	Familial cylindromatosis gene	None	Familial cylindromatosis	Cylindromas
EP300	Q09472	300-kD E1A-binding protein gene	Rubinstein-Taybi gene	n/a	Colorectal, breast, pancreatic cancers
EXT1	Q16394	Multiple extoses type 1 gene	EXT-like 1	Multiple exostoses type 1	Exostoses, osteosarcoma
EXT2	Q93063	Multiple extoses type 2 gene	EXT-like 3	Multiple exostoses type 2	Exostoses, osteosarcoma
CDKN1C	P49918	Cyclin-dependent kinase inhibitor 1C gene	None	Beckwith-Wiedemann syndrome	Wilms' tumour, rhabdomyosarcoma
STK11	Q15831	Serine/threonine kinase 11 gene	None	Peutz-Jeghers syndrome	Jejunal harmartomas, ovarian tumours, testicular and pancreatic cancers
MAP2K4	P45985	Mitogen-activated protein kinase kinase 4	Multiple map kinase family members	n/a	Pancreatic, breast, colon cancers
MEN1	O00255	Multiple endocrine neoplasia type 1 gene	None	Multiple endocrine neoplasia type 1	Parathyroid/pituitary adenoma, islet cell carcinoma, carcinoid tumours
MLH1	P40692	*E. coli* MutL homologue gene	None	Familal non-polyposis colorectal cancer	Colorectal, endometrial, ovarian cancer
MSH2	P43246	*E. coli* MutS homologue 2 gene	None	Familal non-polyposis colorectal cancer	Colorectal, endometrial, ovarian cancer
NF1	P21359	Neurofibromatosis type 1 gene	None	Neurofibromatosis type 1	Neurofibroma, glioma
NF2	P35240	Neurofibromatosis type 2 gene	Moesin	Neurofibromatosis type 2	Meningioma, acoustic neuroma
PRKAR1A	P10644	Protein kinase A type 1-α regulatory subunit gene	None	Carney complex	Myxoma, endocrine tumours
PTCH	Q13635	Homologue of *Drosophila patched* gene	Patched 2	Nevoid basal cell carcinoma syndrome	Basal cell carcinoma, medulloblastoma
PTEN	O00633	Phosphatase and tensin homologue gene	None	Cowdens syndrome	Harmartomas, glioma, prostate and endometrial cancers
RB1	P06400	Retinoblastoma gene	None	Familial retinoblastoma	Retinoblastoma, sarcomas, breast cancer, small cell lung cancer
SDHD	O14521	Succinate dehyrdogenase cytochrome B small subunit gene	None	Familial paraganglioma	Paraganglioma
MADH4	NP_005350	Homologue of *Drosophila mothers against decapentaplegic 4* gene	None	Juvenile polyposis	Gastrointestinal polyps, colorectal and pancreatic cancers
SMARCB1	Q12824	Swi/Snf5 matrix-associated actin-dependent regulator of chromatin gene	None	Rhabdoid predisposition syndrome	Malignant rhabdoid tumours
TP53	P04637	Tumour suppressor p53 gene	p73, p63 genes	Li-Fraumeni syndrome	Sarcoma, adrenocortical carcinoma, glioma, other tumour types
TSC1	Q92574	Tuberous sclerosis 1 gene	None	Tuberous sclerosis 1	Hamartomas, renal cell carcinoma
TSC2	P49815	Tuberous sclerosis 2 gene	None	Tuberous sclerosis 2	Hamartomas, renal cell carcinoma
VHL	P40337	Von Hipple-Lindau syndrome gene	None	Von Hipple-Lindau syndrome	Renal cell carcinoma, hemangioma, phaeochromocytoma
WT1	P19544	Wilms tumour 1 gene	Zinc finger-containing genes	Familial Wilms tumour	Wilms tumour

*50% identity over at least 50 amino acids.

been designed to support such investigations, so it is not surprising that it did not bear much fruit. In addition to the problems of false positives, relatively few clones have been sequenced from most of the libraries, many of the libraries are not normalized (leading to undersampling of less abundant transcripts) and many of the sampled cDNAs are not full length. This was illustrated by the fact that when we searched the CGAP annotations of libraries constructed from five cancer types with known chimaeric genes, we found only one of the ten genes involved in the five chimaeric transcripts. This exercise underscores the point that elucidating the complexity of cancer at the genomic level will require much more sequence data from cancer genomes, which will need to be configured appropriately for the task at hand.

So, the working draft will not immediately reveal the natures of the abnormalities in cancer cell genomes. To facilitate these analyses we will need the finished sequence, which will form a structural framework for a new generation of massive-scale comparisons of cancer cell and normal genomes. Ultimately, it may also prompt a shift away from strategies that depend upon primary genomic localization and allow systematic genome-wide searches for mutations. New technology will be required; there is no single technology at present that will detect all the types of abnormality (large deletions, rearrangements, base substitutions, small insertions and deletions, amplifications, and epigenetic changes such as methylation) that are present in cancer cells. Sequencing of genomic libraries constructed from cancer genomes would come closest to this goal, but given the diversity of cancers and the effort and cost required to obtain reasonable coverage of a human genome this is a daunting challenge. □

1. Higginson, J., Muir, C. & Munoz, N. *Human Cancer: Epidemiology and Environmental Causes* (Cambridge Monographs on Cancer Research, Cambridge, UK, 1992).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
3. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10,** 547–548 (2000).
4. Kaelin, W. G. Jr The p53 gene family. *Oncogene* **18,** 7701–7705 (1999).
5. Rowley, J. D. The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.* **32,** 495–519 (1998).
6. Bell, R. S., Wunder, J. & Andrulis, I. Molecular alterations in bone and soft-tissue sarcoma. *Can. J. Surg.* **42,** 259–266 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Correspondence and requests for materials should be addressed to M.R.S. (e-mail: mrs@sanger.ac.uk).