Histopathology

Histopathology 2014, **64**, 151–162. DOI: 10.1111/his.12272

REVIEW

Evolving techniques for gene fusion detection in soft tissue tumours

Fredrik Mertens & Johnbosco Tayebwa
Department of Clinical Genetics, University and Regional Laboratories, Lund University, Lund, Sweden

Mertens F & Tayebwa J
(2014) *Histopathology* **64**, 151–162

Evolving techniques for gene fusion detection in soft tissue tumours

Chromosomal rearrangements resulting in the fusion of coding parts from two genes or in the exchange of regulatory sequences are present in approximately 20% of all human neoplasms. More than 1000 such gene fusions have now been described, with close to 100 of them in soft tissue tumours. Although little is still known about the functional outcome of many of these gene fusions, it is well established that most of them have a major impact on tumorigenesis. Furthermore, the strong association between type of gene fusion and morphological subtype makes them highly useful diagnostic markers. Until recently, the vast majority of gene fusions were identified through molecular cytogenetic characterization of rearrangements detected at chromosome banding analysis, followed by use of the reverse transcriptase–polymerase chain reaction (RT–PCR) and Sanger sequencing. With the advent of next-generation sequencing (NGS) technologies, notably of whole transcriptomes or all poly-A^+ mRNA molecules, the possibility of detecting new gene fusions has increased dramatically. Already, a large number of novel gene fusions have been identified through NGS approaches and it can be predicted that these technologies soon will become standard diagnostic clinical tools.

Keywords: gene fusion, next-generation sequencing, sarcoma, soft tissue

Somatic mutations

A century ago, Theodor Boveri postulated that neoplasia was caused by chromosomal rearrangements, a hypothesis known later as the somatic mutation theory of cancer.[1] The validity of his hypothesis could not, however, be evaluated properly until the advent of investigative approaches allowing for more refined analysis of the genomes of neoplastic cells, a process requiring decades of methodological improvements. With the vast array of technologies available to us today, it has been proved beyond doubt that Boveri's idea was correct – all neoplasms show more or less extensive genetic abnormalities. Furthermore, different tumour types are characterized by different, although sometimes overlapping, spectra of genetic changes. The types of mutation identified vary considerably among tumour types, ranging from single nucleotide mutations to large-scale genomic alterations involving ploidy shifts and losses or gains of entire chromosomes. By studying patterns of mutations in tumours of different lineages or at different stages of development, and by evaluating the cellular effects of mutated genes in experimental model systems, an ever more refined network of neoplasia-associated genes and signalling pathways has been revealed. Combined with the steadily accumulating information on how our genome regulates normal physiological processes such as cell division, differentiation and life span, our understanding of how neoplasms develop has increased dramatically in the last few decades. Apart from unravelling cellular processes involved in tumour development, the genetic data have also provided clinicians with valuable

Address for correspondence: F Mertens, Department of Clinical Genetics, University and Regional Laboratories, Lund University, SE-221 85 Lund, Sweden. e-mail: fredrik.mertens@med.lu.se

diagnostic and prognostic markers; and in some cases, the underlying genetic aberrations have become important therapeutic targets.[2,3]

Gene fusions

One particular type of neoplasia-associated mutation that has attracted much attention from a biological as well as a clinical viewpoint is gene fusion. Gene fusions are formed by structural chromosomal rearrangements (translocations, inversions, interstitial deletions) that result in chimeric genes or in the exchange of regulatory sequences. The prototypical examples are the *BCR–ABL1* fusion resulting from a t(9;22)(q34;q11) in chronic myeloid leukaemia and the *IGH@–MYC* fusion in Burkitt lymphoma with t(8;14)(q24;q32), where the former represents a gene fusion creating a chimeric protein and the latter is an example of transcriptional up-regulation of an oncogene (*MYC*) through exchange of regulatory sequences.[4] As gene fusions are often seen as the sole cytogenetic change, with few accompanying mutations detected even using high-resolution, genome-wide technologies, they are assumed to have a very strong pathogenetic impact. This conclusion has been supported further by results from *in-vitro* studies and from experimental animal models, showing that the gene fusion, at least if it occurs in a permissive cellular context, is sometimes sufficient for malignant transformation.[5] The strong impact of gene fusions on tumour cells, coupled with the fact that chimeric genes are specific for tumour cells, make them very attractive as potential targets for treatment. Indeed, novel treatments based on the presence of *BCR–ABL1* in chronic myeloid leukaemia and various *ALK* chimeras in lung cancer and other neoplasms constitute excellent examples of the feasibility of this approach.[6,7] Furthermore, as the gene fusions are often associated very strongly with tumour morphology, i.e. a specific fusion tends to be present in only one or a few tumour types, they have become highly useful for diagnostic purposes.[4]

Detection of gene fusions: the classical route

The first gene fusions in human neoplasia were described in the early 1980s, i.e. some 10 years after the invention of chromosome banding techniques. The latter breakthrough not only greatly facilitated the detection of structural chromosomal changes, such as inversions or reciprocal translocations in metaphase spreads from cultured tumour cells, but also made it possible to assign the breakpoints of these rearrangements to specific chromosomes and chromosome bands. Hence, chromosome banding analysis provided an opportunity for systematic screening of recurrent chromosomal breakpoints in neoplasms that could be cultured *in vitro* which, to a large extent, was initially synonymous with haematological neoplasms. It was soon discovered that certain structural aberrations, notably translocations, were associated non-randomly with particular leukaemia subtypes, implying strongly that genes of importance for tumorigenesis were located at or near the chromosomal breakpoints. Nevertheless, even with this type of information to hand, cloning of the genes affected by the chromosome rearrangements was initially an arduous task. However, further technical improvements during the 1980s, in parallel with the increasingly detailed map of the human genome, radically reduced the time needed to pinpoint the genes affected by chromosomal breakpoints. Two new technologies emerged as being particularly fruitful for gene fusion detection: first, fluorescence *in-situ* hybridization (FISH) allowed cytogeneticists to narrow the breakpoint regions to a few hundred Kb, compared to a resolution level of 5–10 Mb when using chromosome banding alone.[8] Secondly, the invention of the RT–PCR technique made it possible to test directly for potential fusion transcripts (Figure 1). Illustrating the additive effects of increased amounts of cytogenetic information and technical improvements, only 11 gene fusions were identified during the 7-year period 1982–88, followed by 82 during 1989–95 and 171 during 1996–2002.[9,10]

New approaches to identify gene fusions

While the classical route of detecting gene fusions (chromosome banding followed by FISH followed by RT–PCR) has a number of merits, it also suffers from some very important shortcomings. First, metaphase spreads of high quality are not obtained easily from some tumour types; certain cells simply do not thrive *in vitro*, or at least require highly sophisticated culturing conditions in order to spawn a sufficient number of mitoses. Furthermore, many malignant neoplasms display such complex karyotypes that it is next to impossible to identify the chromosomes, let alone the chromosomal bands, involved in structural rearrangements. Secondly, some tumour types could have frequent gene fusions without having any corresponding, microscopically visible chromosome aberration.

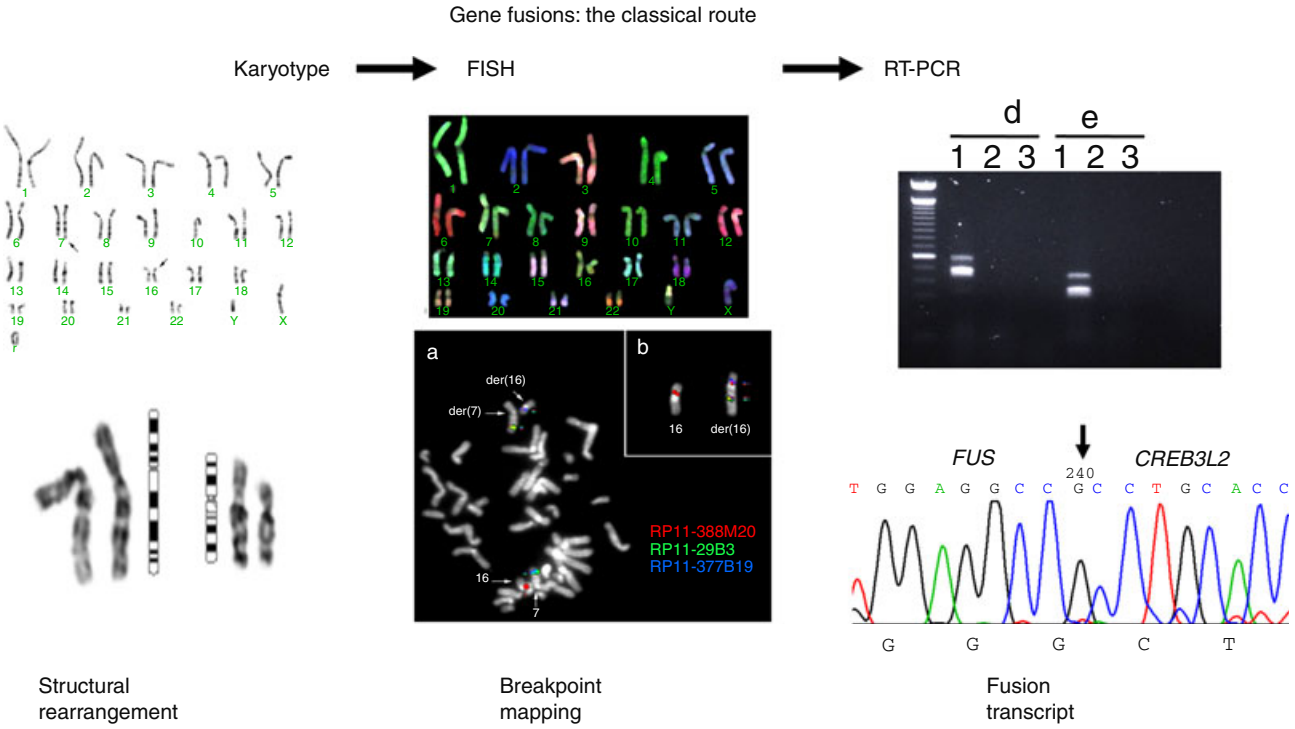

Figure 1. The classical route for gene fusion detection. After short-term culturing of cells from a fresh tumour biopsy, chromosome banding is performed on metaphase spreads. The breakpoints of balanced translocations, such as the t(7;16) depicted here, can then be delineated by FISH, followed by RT–PCR for putative fusion transcripts. Products found at RT–PCR can then be sequenced using traditional Sanger sequencing.

For instance, gene fusions due to small interstitial deletions or inversions, or translocations affecting the ends of chromosomes, may all result in gene fusions without being detectable even in high-quality chromosome preparations.[4]

The emergence in the 1990s of array-based technologies for high-resolution analysis of tumour genomes and transcriptomes provided new options to circumvent the need for cell culturing in the search for gene fusions.[11] In principle, gene fusions due to unbalanced chromosomal rearrangements or else associated with small gains or losses could be detected by analysis of genomic DNA; many seemingly balanced translocations resulting in gene fusions are actually accompanied by more or less extensive deletions and/or duplications in the breakpoint regions.[12] However, this approach has not really been used as a stand-alone technique, hampered as it is by the sheer number of breakpoints found typically in malignancies, as well as by extensive constitutional copy number variation. Nevertheless, as an adjunct to other sets of data, e.g. cytogenetics, FISH or gene expression profiling, this approach might be useful in pinpointing genes affected by chromosomal rearrangements. In contrast, global gene expression profiling has emerged as a very fruitful method when searching for gene fusions. By focusing on genes showing outlier expression values, tentative targets for genes affected by chromosomal rearrangements could be identified. The best example of this approach is perhaps the finding of recurrent gene fusions in prostate cancer, the first example of a common epithelial malignancy showing gene fusions.[13] An even more sophisticated way of utilizing expression data for gene fusion discovery is to look at expression levels at the exon level, rather than at the gene level; gene expression data are presented typically as mean expression levels for several exons of a gene, but some arrays allow for the analysis of each exon individually. Consequently, genes showing discrepancies between expression levels of their 5′- and 3′-parts could be suspected of being affected by rearrangements separating the two parts from each other. This approach has been applied successfully,[14] and attempts have been made to develop such assays for clinical, diagnostic purposes.[15] Although proven to be successful, as well as being theoretically pleasing, this approach is unlikely to attract much attention in the future: not all gene fusions result in differential expression of the

separated parts of the genes involved, only known gene fusions are tested for and the bioinformatic analysis is far from trivial, etc. Furthermore, it is difficult to see what advantages this approach may offer compared to deep sequencing of tumour genomes and transcriptomes.

Next-generation sequencing: gene fusions galore

While each of the above-mentioned technical improvements have certainly contributed substantially to the detection of gene fusions in human neoplasia, their merits pale in comparison with what can now be achieved by using so-called next-generation sequencing (NGS) technologies. NGS (also known as second-generation sequencing, deep sequencing, massively parallel sequencing, etc.) is an umbrella term for various solutions to obtain simultaneously both width (i.e. multiple nucleotide sequences are analysed at the same time) and depth (i.e. each target nucleotide sequence is analysed several times, allowing for the detection of rare, mosaic variants) in the analysis of genetic material. It is beyond the scope of the present review to discuss the strategies behind the many different NGS platforms; several excellent reviews on technical aspects of NGS technologies are already available.[16–19] It is sufficient here to mention that, until 2005, for almost 30 years sequencing had been based on the methods developed by Sanger and co-workers (so-called Sanger sequencing).[20,21] Although this sequencing method provides data of high accuracy for sequences up to 500–1000 nucleotides in length, it is poorly suited for the analysis of large numbers of nucleotide sequences. Thus, two papers that appeared in 2005, describing cost-efficient, automated approaches to obtaining large amounts of sequence data (known as reads) in parallel, opened unprecedented opportunities for querying genomes.[22,23] Spurred by competition among companies developing and launching different platforms and solutions for NGS, the pace of this technological evolution has been, and still remains, staggering: the quality and quantity of data that can be obtained in a single run keep increasing, while at the same time the amounts of starting material and the time needed for the analysis keep decreasing. Not the least important, the costs for such analyses have dropped dramatically, making the technologies widely affordable. In this context it could be pointed out, however, that while the costs for equipment, flow cells, and reagents have decreased, the costs for data processing and interpretation or for the complementary experiments needed to verify NGS results are still considerable.[24]

In parallel with the development of sophisticated solutions for sequencing of nucleotides, improvements of methods for obtaining and labelling DNA and RNA molecules have contributed significantly to optimizing the results of sequencing, such as the various end-sequence profiling methods that allow analysis of paired rather than single reads.[25,26] Finally, prompted by the need to be able to handle extremely large (giga- to terabyte level) sets of data, there have been major achievements in the field of bioinformatics, with scripts, algorithms and pipelines being developed for all types of data. For instance, there are now large numbers of different types of software to choose from for the alignment of sequencing data to a reference genome (e.g. CASAVA, STAR, TopHat and TrueSight; reviewed by Fonseca *et al.*[27]) and for the identification of fusion transcripts (e.g. ChimeraScan, defuse, FusionMap and FusionSeq; reviewed by Carrara *et al.*[28]). In summary, it is currently possible to obtain information within a few days on the entire genome or transcriptome of any type of cell or organism, using as little starting material as the DNA or RNA of a few cells. An excellent recent review of the potential benefits of these technologies for clinical and research aspects of tumour biology has been provided by Vogelstein and co-workers.[3]

Soon after the successful implementation of BAC-end sequencing and serial analysis of gene expression (SAGE) sequencing for the detection of gene fusions,[26,29,30] the first study taking full advantage of the new sequencing possibilities was published in 2008 by Campbell and co-workers.[31] Using a genome-wide, paired-end sequencing approach, two lung cancer genomes were analysed at the DNA level, revealing numerous mutations, including some interchromosomal rearrangements that could be confirmed to result in fusion transcripts. This report was followed quickly by an avalanche of NGS studies employing different variants of transcriptome sequencing (also known as RNA-Seq), usually on RNA molecules with poly-A tails, i.e. protein-coding genes, revealing hundreds of novel gene fusions in common malignancies such as carcinomas of the breast, lung and prostate.[32–36] The main reason for RNA being the preferred starting material when searching for gene fusions is that most of those detected so far arise due to breaks within introns, which are sometimes very large. As splicing of these non-coding intervening sequences occurs during RNA processing, resulting in end-to-end joining of the exons, the mRNA molecule is much smaller than the

corresponding DNA sequence. Thus, in order to find the fusion at the DNA level, either the whole genome or an enriched region of DNA has to be analysed; exon-based sequencing of all genes (whole-exome sequencing) or of a panel of selected genes will often not identify fusion events associated with intronic breakpoints. This notwithstanding, it should be kept in mind that several gene fusions have been found by analysing DNA,[31,37–39] and as the costs and bioinformatic problems decrease whole-genome DNA approaches will also become increasingly valuable for gene fusion detection. Nevertheless, a major advantage of using RNA as the starting material is, of course, that analyses at the transcript level provides information not only about potential gene fusions, but also about expression levels and transcript variants.

Although RNA-Seq has already been highly successful in identifying gene fusions, there are several technical and bioinformatic pitfalls to consider. Apart from obvious limitations and artefacts related to suboptimal RNA quality and potential errors introduced when converting RNA to complementary DNA (cDNA), RNA-Seq will not detect chromosomal rearrangements resulting in promoter swapping, i.e. the fusion breakpoints have to be located within the mature mRNA molecules in order to be picked out. However, such gene fusions constitute only a small fraction of the gene fusions identified through the classical route,[10] suggesting that this might be a relatively minor problem. Finally, RNA-Seq will cover genes on the basis of their expression levels, i.e. there are more reads for highly expressed genes than for poorly expressed ones. There are already some reports suggesting that the functional outcome of a gene fusion sometimes might be transcriptional silencing of one of the two genes involved.[40] Potentially, such fusion events could be difficult to detect by RNA-Seq. In addition to false-negative results, there is a very high likelihood of obtaining false-positive results. As demonstrated in most RNA-Seq studies, a common phenomenon in neoplastic as well as in normal cells is so-called read-through transcripts, also known as transcription-induced chimeras. The transcription of a gene usually stops at a specific termination point, identified through its nucleotide sequence. However, this mechanism for controlling the activity of RNA polymerase is sometimes bypassed, and the transcription continues to the next gene on the same strand. This results in a chimeric transcript in which the intervening, non-coding region between the two genes is removed from the final, fused mRNA. The likelihood of such an event occurring is greater when the distance between the two neighbouring genes is small.[41] Although read-through transcripts, representing extreme variants of alternate splice forms of a gene, have been shown to generate functional proteins and hence may have effects on both non-neoplastic and neoplastic cells,[42] in the vast majority of cases they are unlikely to represent pathogenetically important, neoplasia-associated fusion transcripts. In addition to chimeric *cis*-fusion transcripts, *trans*-splicing events (also known as non-co-linear transcripts) have also been found in both normal and neoplastic cells.[43,44] *Trans*-splicing, i.e. fusion of transcripts from non-adjacent genes without a corresponding fusion at the DNA level, is a common phenomenon in certain organisms such as trypanosomes and nematodes, but has also been identified in humans and other mammalians.[43] At least two neoplasia-associated *trans*-spliced RNA molecules have been described: *SLC45A3–ELK4* in prostate cancer and *JAZF1–SUZ12* in endometrial stromal sarcomas.[45,46] However, only the former fusion has been verified in independent studies, whereas the validity of the latter fusion has been questioned.[47] All aspects combined, it seems prudent at present to verify any potential fusion transcript identified by RNA-Seq at the transcript level by RT–PCR and at the DNA level by genomic PCR, FISH or cytogenetics.

Gene fusions in soft tissue tumours

When more thorough chromosome banding analyses of soft tissue tumours were initiated in the early 1980s, it soon became apparent that each morphological subtype has its own characteristic cytogenetic profile, ranging from single numerical or structural aberrations to highly complex karyotypes.[48] Somewhat surprisingly at that time, analysis of benign tumours also revealed typical chromosomal aberration patterns, including tumour-specific balanced translocations. As had been the case for leukaemias and lymphomas, recurrent balanced rearrangements attracted particular attention, and in 1992 the first sarcoma-associated gene fusion – the *EWSR1–FLI1* chimera resulting from t(11;22)(q24;q12) in Ewing sarcoma – was discovered.[49] During the following 20 years, new gene fusions were added at a fairly constant rate, and now amount to 94 fusions in more than 30 distinct entities (Table 1). So far, no particular clinicomorphological feature has been identified separating fusion-positive from fusion-negative entities. Apart from nerve sheath tumours, all major lineages are represented in the fusion-positive group

Table 1. Gene fusions in soft tissue tumours*

Tumour	Gene fusion	Chromosome aberration
Adipocytic tumours		
Lipoma	*EBF1-LOC204010*	t(5;12)(q33;q14)
	HMGA2-CXCR7	t(2;12)(q37;q14)
	HMGA2-EBF1	t(5;12)(q33;q14)
	HMGA2-LHFP	t(12;13)(q14;q13)
	HMGA2-LPP	t(3;12)(q28;q14)
	HMGA2-NFIB	t(9;12)(p22;q14)
	HMGA2-PPAP2B	t(1;12)(p32;q14)
	LPP-C12orf9	t(3;12)(q28;q14)
Lipoblastoma	*COL1A2-PLAG1*	t(7;8)(q21;q12)
	HAS2-PLAG1	del(8)(q12q24)
Chondroid lipoma	*C11orf95-MKL2*	t(11;16)(q13;p13)
Myxoid/round cell liposarcoma	*FUS-DDIT3*	t(12;16)(q13;p11)
	EWSR1-DDIT3	t(12;22)(q13;q12)
Dedifferentiated liposarcoma	*CNOT2-ASTN2*	t(9;12)(q33;q15)
	CTDSP2-FAM19A2	?t(12)(q14q14)
	NR6A1-TRHDE	t(9;12)(q33;q21)
	NUP107-LGR5	?t(12)(q15q21)
	NUP107-PAPPA	t(9;12)(q33;q15)
	RCOR1-WDR70	t(5;14)(p13;q32)
Fibroblastic/Myofibroblastic tumours		
Soft tissue angiofibroma	*AHRR-NCOA2*	t(5;8)(p15;q13)
	GTF2I-NCOA2	t(7;8;14)(q11;q13;q31)
Dermatofibrosarcoma protuberans	*COL1A1-PDGFB*	t(17;22)(q21;q13)
Solitary fibrous tumour	*NAB2-STAT6*	inv(12)(q13q13)
Infantile fibrosarcoma	*ETV6-NTRK3*	t(12;15)(p13;q25)
Low-grade fibromyxoid sarcoma	*FUS-CREB3L2*	t(7;16)(q34;p11)
	FUS-CREB3L1	t(11;16)(p11;p11)
Sclerosing epithelioid fibrosarcoma	*FUS-CREB3L2*	t(7;16)(q34;p11)

Table 1. *(Continued)*

Tumour	Gene fusion	Chromosome aberration
Inflammatory myofibroblastic tumour	*ATIC-ALK*	inv(2)(p23q35)
	CARS-ALK	t(2;11)(p23;p15)
	CLTC-ALK	t(2;17)(p23;q23)
	PPFIBP1-ALK	t(2;12)(p23;p11)
	RANBP2-ALK	t(2;2)(p23;q13)
	RREB1-TFE3	t(X;6)(p11;p24)
	SEC31A-ALK	t(2;4)(p23;q21)
	TPM3-ALK	t(1;2)(q21;p23)
	TPM4-ALK	t(2;19)(p23;p13)
So-called fibrohistiocytic tumours Tenosynovial giant cell tumour	*COL6A3-CSF1*	t(1;2)(p13;q37)
Smooth muscle tumours Leiomyoma of the uterus	*CUX1-AGR3*	inv(7)(p21q22)
	HMGA2-CCNB1IP1	t(12;14)(q14;q11)
	HMGA2-COG5	t(7;12)(q31;q14)
	HMGA2-COX6C	t(8;12)(q22;q14)
	HMGA2-RAD51L1	t(12;14)(q14;q24)
Pericytic (perivascular) tumours Pericytoma with t(7;12)	*ACTB-GLI1*	t(7;12)(p22;q13)
Skeletal muscle tumours Alveolar rhabdomyosarcoma	*FOXO1-FGFR1*	t(8;13;9)(p11;q14;q32)
	PAX3-FOXO1	t(2;13)(q36;q14)
	PAX3-FOXO4	t(X;2)(q13;q36)
	PAX3-NCOA1	t(2;2)(p23;q36)
	PAX3-NCOA2	t(2;8)(q36;q13)
	PAX7-FOXO1	t(1;13)(p36;q14)
Spindle cell rhabdomyosarcoma	*SRF-NCOA2*	t(6;8)(p21;q13)
	TEAD1-NCOA2	t(8;11)(q13;p15)
Vascular tumours Epithelioid hemangioendothelioma	*WWTR1-CAMTA1*	t(1;3)(p36;q25)
	YAP1-TFE3	t(X;11)(p11;q22)

Table 1. *(Continued)*

Tumour	Gene fusion	Chromosome aberration
Tumours of uncertain differentiation		
Angiomatoid fibrous histiocytoma	*EWSR1-CREB1*	t(2;22)(q33;q12)
	FUS-ATF1	t(12;16)(q13;p11)
	EWSR1-ATF1	t(12;22)(q13;q12)
Ossifying fibromyxoid tumour	*EP400-PHF1*	t(6;12)(p21;q24)
Myoepithelioma/mixed tumour	*EWSR1-ATF1*	t(12;22)(q13;q12)
	EWSR1-PBX1	t(1;22)(q23;q12)
	EWSR1-POU5F1	t(6;22)(p21;q12)
	EWSR1-ZNF444	t(19;22)(q13;q12)
Synovial sarcoma	*SS18-SSX1, SS18-SSX2* or *SS18-SSX4*	t(X;18)(p11;q11)
	SS18L1-SSX1	t(X;20)(p11;q13)
Alveolar soft part sarcoma	*ASPSCR1-TFE3*	t(X;17)(p11;q25)
Clear cell sarcoma	*EWSR1-CREB1*	t(2;22)(q33;q12)
	EWSR1-ATF1	t(12;22)(q13;q12)
Extraskeletal myxoid chondrosarcoma	*TAF15-NR4A3*	t(9;17)(q31;q12)
	TFG-NR4A3	t(3;9)(q12;q31)
	TCF12-NR4A3	t(9;15)(q31;q21)
	EWSR1-NR4A3	t(9;22)(q31;q12)
Desmoplastic small round cell tumour	*EWSR1-WT1*	t(11;22)(p13;q12)
	EWSR1-ERG	t(21;22)(q22;q12)
Ewing sarcoma	*EWSR1-ERG*	t(21;22)(q22;q12)
	EWSR1-ETV1	t(7;22)(p21;q12)
	EWSR1-ETV4	t(17;22)(q21;q12)
	EWSR1-FEV	t(2;22)(q35;q12)
	EWSR1-FLI1	t(11;22)(q24;q12)
	EWSR1-NFATC2	r(20;22)(q13;q12)
	EWSR1-PATZ1	inv(22)(q12q12)
	EWSR1-SMARCA5	t(4;22)(q31;q12)
PEComa	*SFPQ-TFE3*	t(X;1)(p11;p34)

Table 1. *(Continued)*

Tumour	Gene fusion	Chromosome aberration
Undifferentiated/unclassified sarcomas		
Undifferentiated/unclassified sarcomas	*BCOR-CCNB3*	inv(X)(p11p11)
	CIC-DUX4	t(4;19)(q35;q13)
	CIC-DUX4L10	t(10;19)(q26;q13)
	EWSR1-POU5F1	t(6;22)(p21;q12)
	EWSR1-SP3	t(2;22)(q31;q12)
Chondro-osseous tumours		
Soft tissue chondroma	*HMGA2-LPP*	t(3;12)(q28;q14)
Mesenchymal chondrosarcoma	*HEY1-NCOA2*	t(8;8)(q13;q21 or del(8)(q13q21)
	IRFBP2-CDX1	t(1;5)(q42;q32)
Miscellaneous tumours		
Endometrial stromal sarcoma	*EPC1-PHF1*	t(6;10;10)(p21;q22;p11)
	JAZF1-PHF1	t(6;7)(p21;p15)
	JAZF1-SUZ12	t(7;17)(p15;q11)
	MEAF6-PHF1	t(1;6)(p34;p21)
	YWHAE-FAM22A	t(10;17)(q23;p13)
	YWHAE-FAM22B	t(10;17)(q22;p13)
	ZC3H7B-BCOR	t(X;22)(p11;q13)
Epithelioid sarcoma of the ovary	*CMKLR1-HNF1A*	?t(12;12)(q23;q24)
	ERBB3-CRADD	?t(12;12)(q13;q22)
	SMARCB1-WASF2	t(1;22)(p36;q11)
Primary pulmonary myxoid sarcoma	*EWSR1-CREB1*	t(2;22)(q33;q12)

*Gene fusions were retrieved from Mitelman *et al.*,[10] Queried on April 30, 2013. Gene fusions in black were identified using the classical route (chromosome banding and FISH), gene fusions in blue were identified through global gene expression profiling and gene fusions in red were found using next-generation sequencing data.

by at least one tumour type; and both benign lesions, such as conventional lipoma and tenosynovial giant cell tumour, and highly malignant ones, such as Ewing sarcoma and synovial sarcoma, harbour gene fusions (Table 1).

The vast majority of the currently known gene fusions in soft tissue tumours were identified through the classical route (shown in black in Table 1), i.e. cytogenetics followed by FISH and RT–PCR (Figure 1), but a few of them, such as *PAX3–NCOA1* in alveolar rhabdomyosarcoma and *HEY1–NCOA2* in mesenchymal chondrosarcoma,[14,50] were identified largely through aberrant gene expression profiles (blue in Table 1). Initial NGS studies have been focused on common epithelial malignancies, such as carcinomas of the breast, lung and prostate, but a quickly growing number of studies on less common malignancies such as sarcomas have also been published,[39,51–60] already resulting in 18 newly identified gene fusions (red in Table 1, Figure 2). So far, only one sarcoma-associated gene fusion has been discovered by DNA-based NGS analysis, namely the *NAB2–STAT6* fusion in solitary fibrous tumour.[39,57,60] Interestingly, this fusion of two neighbouring genes in chromosome band 12q13 was discovered in three independent studies using three different NGS approaches: RNA-Seq,[60] whole-exome sequencing[57] and deep sequencing of an enriched genomic

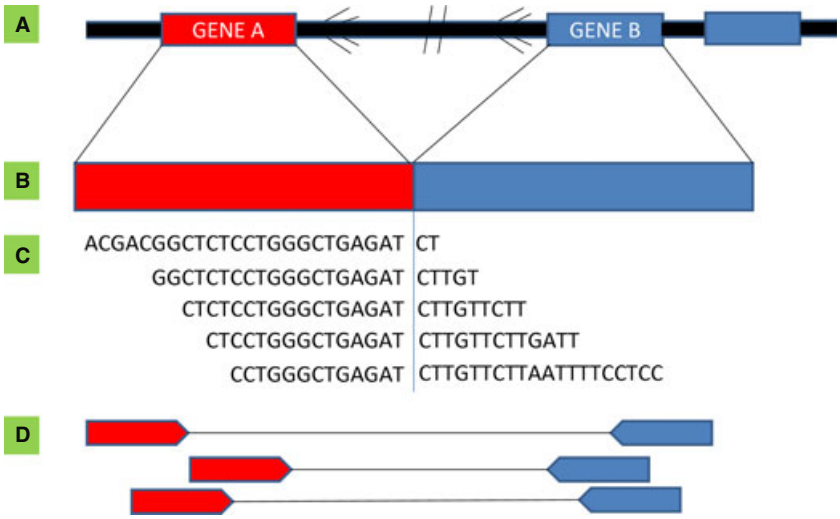

Figure 2. Schematic of gene fusion detection using data generated by next-generation sequencing. **A,** Chromosomal context of the gene fusion partners, gene A and gene B, which may be located on the same chromosome or on different chromosomes. **B,** Gene fusion resulting from a translocation, insertion or interstitial deletion. **C,** Fusion junction-spanning reads. Each such read contains a variable number of nucleotides from genes A and B. The junction-spanning reads allow for exact delineation of the breakpoints in the two genes. **D,** Mate-pair fusion-spanning reads. The two reads in a pair map to genes A and B, respectively. These mate-pairs indicate the presence of a gene fusion, but do not show the exact breakpoints in the two genes.

region.[39] The RNA-Seq approach was the most successful in terms of few false-negative results, while the enrichment strategy had the added value of providing information on breakpoint localization at the genomic level (Figure 2).

Undoubtedly, several large-scale NGS studies of soft tissue tumours are under way, and numerous previously unsuspected genes will be implicated in tumour development. For obvious reasons, the focus will initially be on highly malignant lesions, for which current treatment protocols are insufficient. However, it will be of great interest to also determine the 'true' mutation spectrum in less aggressive sarcomas and in benign soft tissue lesions, as this information could be highly relevant for understanding why some tumours remain localised and others disseminate rapidly.

The future

NGS technologies have already made a huge contribution to oncological research.[3] No doubt they will be used increasingly to study tumour genomes and transcriptomes, and with time they will completely replace many of the methods currently used for mutation detection, copy number evaluation and gene expression profiling. Future challenges will concern technical and bioinformatic aspects, such as how to obtain even deeper sequencing, how to best use small cell samples for both DNA and RNA level analyses, how to make bioinformatic tools more user-friendly, and how to store the excessive amounts of data. These problems will be solved and, bearing in mind not only the diagnostic information inherent in the genetic profiles of neoplasms but also the rapidly increasing availability of drugs targeting specific molecules or signalling pathways, it seems highly likely that NGS will soon also be the method of choice in clinical molecular pathology.

Although many of the technologies used in clinical practice today, such as chromosome banding analysis, RT–PCR or FISH, for detecting gene fusions and other genetic aberrations of diagnostic importance will thus be replaced, information already available from such studies should not be discarded immediately. For instance, abnormal karyotypes from 2280 soft tissue tumours have been reported in the literature (and many more have not been published), and a survey of these cases reveals more than 750 different balanced translocations.[10] Taking into account that gene fusions can result also from unbalanced structural rearrangements, it seems fair to assume that only a small minority of the gene fusions present in soft tissue tumours have been identified. Currently available cytogenetic data might offer some shortcuts in this context, as it might be expected that tumours with known balanced rearrangements and/or recurrent chromosomal breakpoints are more likely to yield novel gene fusions at NGS analysis.

Indeed, several studies have already shown that this is a highly efficient strategy for selecting cases for deep sequencing and for guiding the bioinformatic analysis.[39,53,54,61] However, most types of soft tissue tumour are uncommon, and some fusion events causing their development might be present in only minute proportion of cases. A comparison with what is known about acute myeloid leukaemia (AML) is sobering: there are currently close to 200 known gene fusions in AML, but the tenth most common one is present in fewer than 0.1% of cases.[4,10] As soft tissue tumours constitute a decidedly more heterogeneous group of neoplasms, with many subtypes being exceedingly rare, it seems safe to conclude that it will take many years before all pathogenetically important gene fusions have been identified.

Acknowledgements

The grant support from The Swedish Cancer Foundation, the National Research Council of Sweden, the Gunnar Nilsson Cancer Foundation, and the Swedish Childhood Cancer Foundation is gratefully acknowledged.

Conflict of interests

The authors declare that there are no conflicts of interest.

References

1. Boveri T. *Zur frage der entstehung maligner tumoren*. Jena: Gustav Fisher Verlag, 1914.
2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**; 719–724.
3. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science* 2013; **339**; 1546–1558.
4. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* 2007; **7**; 233–245.
5. Straessler KM, Jones KB, Hu H *et al.* Modeling clear cell sarcomagenesis in the mouse: cell of origin differentiation state impacts tumor characteristics. *Cancer Cell* 2013; **23**; 215–227.
6. Druker BJ. Translation of the Philadelphia chromosome into therapy for CML. *Blood* 2008; **112**; 4808–4817.
7. Mano H. ALKoma: a cancer subtype with a shared target. *Cancer Discov.* 2012; **2**; 495–502.
8. Gisselsson D. Cytogenetic methods. In Heim S, Mitelman F, eds. *Cancer cytogenetics*, 3rd edn. Hoboken, NJ: Wiley-Blackwell, 2009; 9–16.
9. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.* 2004; **36**; 331–334.
10. Mitelman F, Johansson B, Mertens F eds. *Mitelman database of chromosome aberrations and gene fusions in cancer* (2013). http://cgap.nci.nih.gov/Chromosomes/Mitelman (accessed 30 May 2013).
11. Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* 2003; **12**; R145–R152.
12. Sinclair PB, Nacheva EP, Leversha M *et al.* Large deletions at the t(9;22) breakpoint are common and may identify a poor-prognosis subgroup of patients with chronic myeloid leukemia. *Blood* 2000; **95**; 738–744.
13. Tomlins SA, Rhodes DR, Perner S *et al.* Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* 2005; **310**; 644–648.
14. Wang L, Motoi T, Khanin R *et al.* Identification of a novel, recurrent *HEY1–NCOA2* fusion in mesenchymal chondrosarcoma based on a genome-wide screen of exon-level expression data. *Genes Chromosom. Cancer* 2012; **51**; 127–139.
15. Lovf M, Thomassen GOS, Bakken AC *et al.* Fusion gene microarray reveals cancer type-specificity among fusion genes. *Genes Chromosomes Cancer* 2011; **50**; 348–357.
16. Shendure J, Ji H. Next-generation DNA sequencing. *Nat. Biotechnol.* 2008; **26**; 1135–1145.
17. Fullwood MJ, Wei C-L, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* 2009; **19**; 521–532.
18. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 2009; **10**; 57–63.
19. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 2011; **12**; 87–98.
20. Sanger F, Air GM, Barrell BG *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977; **265**; 687–695.
21. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* 1977; **74**; 5463–5467.
22. Margulies M, Egholm M, Altman WE *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**; 376–380.
23. Shendure J, Porreca GJ, Reppas NB *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005; **309**; 1728–1732.
24. Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med.* 2010; **2**; 84.
25. Volik S, Zhao S, Chin K *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA* 2003; **100**; 7696–7701.
26. Volik S, Raphael BJ, Huang G *et al.* Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.* 2006; **16**; 394–404.
27. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics* 2012; **28**; 3169–3177.
28. Carrara M, Beccuti M, Lazzarato F *et al.* State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed. Res. Int.* 2013; **2013**; 340620.
29. Ruan Y, Ooi HS, Choo SW *et al.* Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* 2007; **17**; 828–838.
30. Raphael BJ, Volik S, Yu P *et al.* A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biol.* 2008; **9**; R59.
31. Campbell PJ, Stephens PJ, Pleasance ED *et al.* Identification of somatically acquired rearrangements in cancer using genome-

wide massively parallel paired-end sequencing. *Nat. Genet.* 2008; **40**; 722–729.
32. Maher CA, Kumar-Sinha C, Cao X *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458**; 97–101.
33. Maher CA, Palanisamy N, Brenner JC *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA* 2009; **106**; 12353–12358.
34. Stephens PJ, McBride DJ, Lin M-L *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009; **462**; 1005–1010.
35. Asmann YW, Necela BM, Kalari KR *et al.* Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res.* 2012; **72**; 1921–1928.
36. Banerji S, Cibulskis K, Rangel-Escareno C *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* 2012; **486**; 405–409.
37. Bass AJ, Lawrence MS, Brace LE *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A–TCF7L2 fusion. *Nat. Genet.* 2011; **43**; 964–968.
38. Imielinski M, Berger AH, Hammerman PS *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 2012; **150**; 1107–1120.
39. Mohajeri A, Tayebwa J, Collin A *et al.* Comprehensive genetic analysis identifies a pathognomonic *NAB2/STAT6* fusion gene, non-random secondary genomic imbalances, and a characteristic gene expression profile in solitary fibrous tumor. *Genes Chromosomes Cancer* 2013; **52**; 873–886.
40. Ågerstam H, Lilljebjörn H, Lassen C *et al.* Fusion gene-mediated truncation of *RUNX1* as a potential mechanism underlying disease progression in the 8p11 myeloproliferative syndrome. *Genes Chromosomes Cancer* 2007; **46**; 635–643.
41. Akiva P, Toporik A, Edelheit S *et al.* Transcription-mediated gene fusion in the human genome. *Genome Res.* 2006; **16**; 30–36.
42. Nacu S, Yuan W, Kan Z *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genomics* 2011; **4**; 11.
43. Gingeras TR. Implications of chimaeric non-co-linear transcripts. *Nature* 2009; **461**; 206–211.
44. Zaphiropoulos PG. *Trans*-splicing in higher eukaryotes: implications for cancer development? *Front. Genet.* 2011; **2**; 1–4.
45. Li H, Wang J, Mor G, Sklar J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* 2008; **321**; 1357–1361.
46. Rickman DS, Pflueger D, Moss B *et al. SLC45A3–ELK4* is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res.* 2009; **69**; 2734–2738.
47. Panagopoulos I. Absence of the JAZF1/SUZ12 chimeric transcript in the immortalized non-neoplastic endometrial stromal cell line T HESCs. *Oncol. Lett.* 2010; **1**; 947–950.
48. Mandahl N, Mertens F. Soft tissue tumors. In Heim S, Mitelman F eds. *Cancer cytogenetics*, 3rd edn. Hoboken, NJ: Wiley-Blackwell, 2009; 675–711.
49. Delattre O, Zucman J, Plougastel B *et al.* Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* 1992; **359**; 162–165.
50. Wachtel M, Dettling M, Koscielniak E *et al.* Gene expression signatures identify rhabdomyosarcoma subtypes and detect a novel t(2;2)(q35;p23) translocation fusing *PAX3* to *NCOA1*. *Cancer Res.* 2004; **64**; 5539–5545.
51. McPherson A, Hormozdiari F, Zayed A *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* 2011; **7**; e1001138.
52. Taylor BS, DeCarolis PL, Angeles CV *et al.* Frequent alterations and epigenetic silencing of differentiation pathway genes in structurally rearranged liposarcomas. *Cancer Discov.* 2011; **1**; 587–597.
53. Lee C-H, Ou W-B, Marino-Enriquez A *et al.* 14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma. *Proc. Natl Acad. Sci. USA* 2012; **109**; 929–934.
54. Nyquist KB, Panagopoulos I, Thorsen J *et al.* Whole-transcriptome sequencing identifies novel *IRF2BP2–CDX1* fusion gene brought about by translocation t(1;5)(q42;q32) in mesenchymal chondrosarcoma. *PLoS ONE* 2012; **7**; e49705.
55. Pierron G, Tirode F, Lucchesi C *et al.* A new subtype of bone sarcoma defined by *BCOR–CCNB3* gene fusion. *Nat. Genet.* 2012; **44**; 461–466.
56. Antonescu CR, Le Loarer F, Mosquera J-M *et al.* Novel *YAP1–TFE3* fusion defines a distinct subset of epithelioid hemangioendothelioma. *Genes Chromosom. Cancer* 2013; **52**; 775–784.
57. Chmielecki J, Crago AM, Rosenberg M *et al.* Whole-exome sequencing identifies a recurrent *NAB2–STAT6* fusion in solitary fibrous tumors. *Nat. Genet.* 2013; **45**; 131–132.
58. Mosquera JM, Sboner A, Zhang L *et al.* Recurrent *NCOA2* gene rearrangements in congenital/infantile spindle cell rhabdomyosarcoma. *Genes Chromosomes Cancer* 2013; **52**; 538–550.
59. Panagopoulos I, Thorsen J, Gorunova L *et al.* Fusion of the *ZC3H7B* and *BCOR* genes in endometrial stromal sarcomas carrying an X;22-translocation. *Genes Chromosomes Cancer* 2013; **52**; 610–618.
60. Robinson DR, Wu Y-M, Kalyana-Sundaram S *et al.* Identification of recurrent *NAB2–STAT6* gene fusions in solitary fibrous tumor by integrative sequencing. *Nat. Genet.* 2013; **45**; 180–185.
61. Panagopoulos I, Thorsen J, Gorunova L *et al.* RNA sequencing identifies fusion of the *EWSR1* and *YY1* genes in mesothelioma with t(14;22)(q32;q12). *Genes Chromosomes Cancer* 2013; **52**; 733–740.