BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 22 2013, pages 2822–2825
doi:10.1093/bioinformatics/btt497

Genome analysis

Advance Access publication August 27, 2013

The code structure of the p53 DNA-binding domain and the prognosis of breast cancer patients

Keiko Sato*, Toshihide Hara* and Masanori Ohya

Department of Information Science, Tokyo University of Science, Noda, Chiba 278-8510, Japan

Associate Editor: John Hancock

ABSTRACT

Motivation: The tumor-suppressor gene TP53 mutations are diverse in the central region encoding the DNA-binding domain. It has not been clear whether the prognostic significance for survival in breast cancer patients is the same for all types of mutations. Are there specific types of mutations carrying a worse prognosis? To understand the correlation between the mutations in the gene encoding the DNA-binding domain and the prognosis of breast cancer, we studied the code structure of the DNA-binding domain of breast cancer patients by using various artificial codes in information transmission.

Results: We indicated that the prognostic significance of all types of mutations in the DNA-binding domain is not the same, and that the DNA-binding domain having a certain code structure is important for estimating the prognosis of breast cancer patients.

Contact: keiko@is.noda.tus.ac.jp or hara@is.noda.tus.ac.jp

Received on July 1, 2013; revised on July 30, 2013; accepted on August 16, 2013

1 INTRODUCTION

Breast cancer is the most common cancer among women. Approximately 1.38 million people around the world are diagnosed with breast cancer each year, and ~458 000 people die from this disease, according to the report of International Agency for Research on Cancer.

The tumor-suppressor gene TP53 is one of the most frequently mutated genes in human cancer including breast cancer (Makwane and Saxena, 2009; Suzuki and Matsubara, 2011; Takahashi *et al.*, 2000). The protein p53 encoded by the TP53 gene is a DNA sequence-specific transcription factor involved in the induction of diverse effects such as cell cycle arrest, apoptosis, repair of DNA lesions, senescence and angiogenesis (Takahashi *et al.*, 2000; Varna *et al.*, 2011). Mutant p53 not only loses the wild-type function but also gains new abilities to promote tumorigenesis (Brosh and Rotter, 2009; Murphy and Rosen, 2000; Suzuki and Matsubara, 2011). Most studies have reported that mutations in the TP53 gene confer a worse overall survival and disease-free survival in breast cancer cases, and this effect is independent of other risk factors (Olivier *et al.*, 2006; Olivier *et al.*, 2010; Petitjean *et al.*, 2007; Pharoah *et al.*, 1999). Langerød *et al.* (2007) have reported that the breast cancer death rate for patients with TP53 mutations is four to five times higher than that for those without mutations. TP53 mutations are mostly missense point mutations and concentrate in the central region encoding the DNA-binding domain (Kucera *et al.*, 1999; Suzuki and Matsubara, 2011; Varna *et al.*, 2011). Several studies have shown that mutations in the DNA-binding domain including L2 and L3 loops are associated with poor disease-free and overall survival (Gentile *et al.*, 1999; Lothe *et al.*, 1995; Young *et al.*, 2007). On the other hand, Powell *et al.* (2000) have indicated that the survival of patients with mutations in the L2/L3 loops is not significantly different to the survival of patients with mutations outside of these domains, and Caleffi *et al.* (1994) have indicated that there is no significant difference in the overall survival rates between patients with mutant and wild-type p53 tumors.

There are many patterns for p53 mutations in breast cancer. Various mutations in the TP53 gene cause single amino acid changes at many different positions. It is not clear whether the prognostic significance is the same for all types of mutations. Are there specific types of mutations carrying a worse prognosis? We need further work to understand the significance concerning the different mutations in breast cancer prognosis.

In information theory, the concept of information has two aspects, one of which expresses the amount of complexity of the whole system such as a sequence itself, and the other concerns the structure of the system such as the rule stored in the order of sequence (Ingarden *et al.*, 1997).

Concerning the first concept of information, according to Shannon's philosophy, if a system has larger complexity, then the system carries larger information, expressed through the entropy (Ohya and Sato, 2000; Ohya and Volovich, 2011). Here, the idea of Kolmogorov, the complexity of sequences, might be used (Chaitin, 1969). Concerning the second concept of information, one looks for the structure of sequences (in genome). For this purpose, coding theory might be useful for the sequence analysis. The code structure of TP53 gene is studied by applying artificial codes in coding theory. The aim of this study is to investigate whether the gene encoding the DNA-binding domain in breast cancer case can be characterized by the artificial codes in information transmission, and to understand the correlation between various mutations in the gene and breast cancer prognosis.

2 METHODS

2.1 Patient characteristics

This study includes a total of 117 primary breast cancer patients with locally advanced (stage III) or distant metastasis (stage IV) (Berns *et al.*, 1998; Bertheau *et al.*, 2007; Chrisanthar *et al.*, 2008; Chrisanthar *et al.*, 2011; Geisler *et al.*, 2001; Geisler *et al.*, 2003; Langerød *et al.*, 2007; Powell *et al.*, 2000; Takahashi *et al.*, 2000). Of these patients, 23 patients

*To whom correspondence should be addressed.

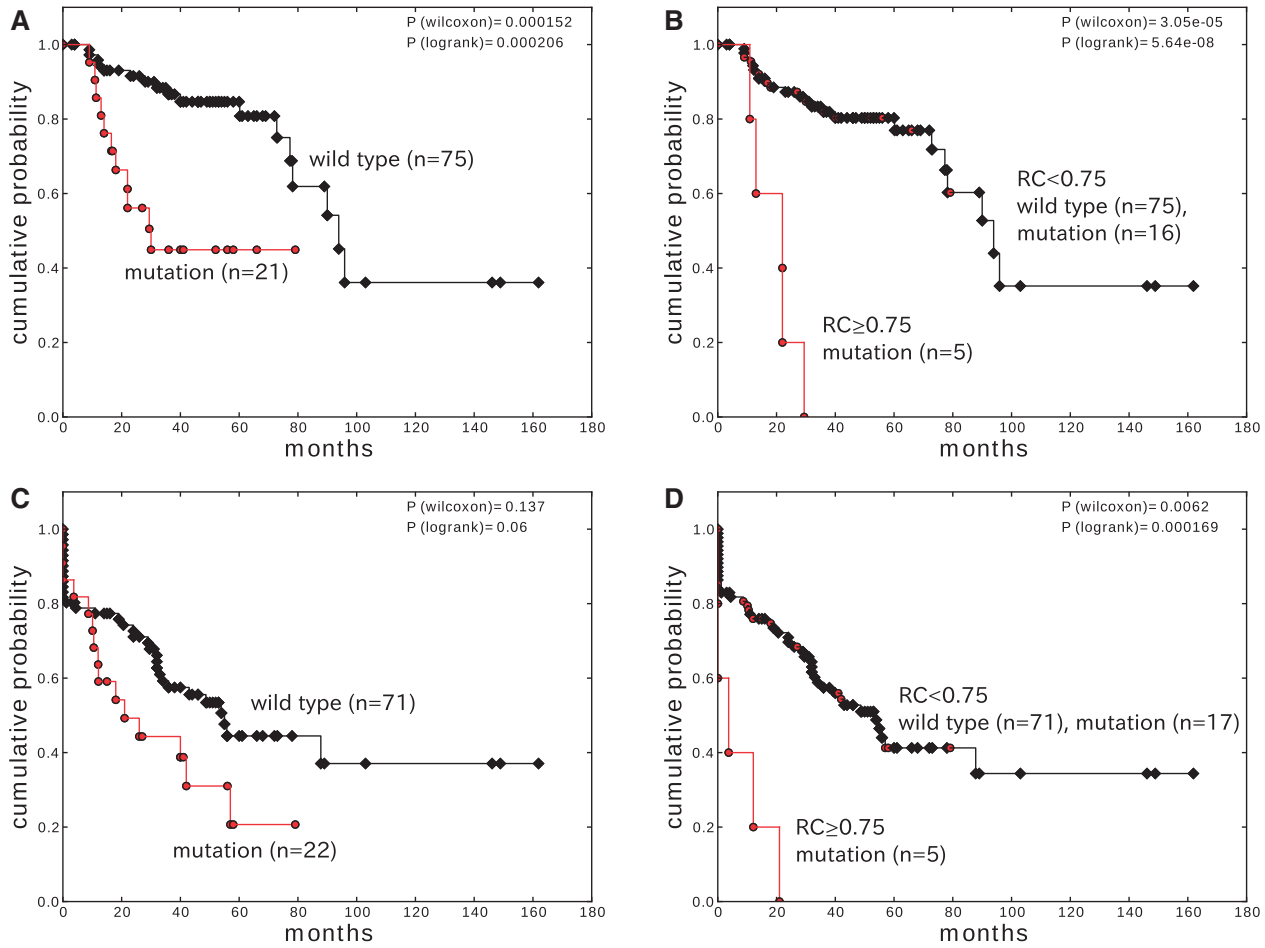

Fig. 1. Survival curves in the L2 loop. Kaplan–Meier survival curves in breast cancer patients stratified by mutation status in the L2 loop (left panels; A and C), and the RC values in the L2 loop (right panels; B and D) are shown: (**A**) Overall survival of patients with mutation in the L2 loop versus those with wild-type. (**B**) Overall survival of patients with the L2 loop that is close to the (33, 22)-BCH code versus those with the L2 loop that is far from that code. (**C**) Relapse-free survival of patients with mutation in the L2 loop versus those with wild-type. (**D**) Relapse free survival of patients with the L2 loop that is close to the (33, 22)-BCH code versus those with the L2 loop that is far from that code

had missense point mutations affecting L2 loop (codons 163–195) of DNA-binding domain, 18 patients had missense point mutations affecting L3 loop (codons 236–251) of DNA-binding domain and 76 patients had wild-type. We selected patients with at least the following clinical data: TP53 mutation status affecting L2/L3 loops identified based on biopsy or surgery, TNM (tumor, node, metastasis) classification and/or histological stage and relapse-free survival and/or overall survival.

2.2 The code structure of the p53 DNA-binding domain

Because the Galois Field $GF(4)$ consists of four elements, 0, 1, α and α^2 such that $\alpha^2 + \alpha + 1 = 0$, the four bases can be expressed in 24 combinations (four bases at each of four elements). We rewrite an important part of the nucleotide sequence encoding the DNA-binding domain by that of these four elements, and we make the error-correcting/detecting code by using an artificial code. The total length of such a code is multiples of three, and the length of the information symbols is multiples of two.

1. Remove the third nucleotide of each codon, corresponding to the check symbol, and transform the remainder (information symbol) into the elements of the Galois Field.
2. Calculate the check symbols from the information symbols by means of the code rule.
3. Put the calculated check symbols into the corresponding position of the third nucleotide of codon.
4. Rewrite the encoded symbol sequence back into the encoded nucleotide sequence.

Artificial codes used for our study are linear codes, cyclic codes, Bose-Chaudhuri-Hocquenghem (BCH) codes, self-orthogonal codes and Iwadare codes.

Let X be an amino acid sequence. We encoded the nucleotide sequence using a code C, and then we got the encoded amino acid sequences X^C by the code C. If the code structure of the sequence X is the same as the structure of the code C, then $X^C = X$. Therefore, we compute the similarity between X and X^C, denoted by rate of coincidence (RC) below, and we claim that this similarity becomes larger if the code structure of X is closer to that of C (Ohya and Sato, 2000; Ohya and Volovich, 2011). The degree RC measuring the similarity between an artificial code of X^C and the intrinsic code of amino acid sequence X before coding by C is defined by

$$RC(X, X^C) = 1 - \frac{a}{t} \quad (0 \le RC(X, X^C) \le 1),$$

where a is the numbers of sites for which two amino acid sequences differ from each other and t is the total number of sites compared. $RC(X, X^C)$ is close to 0 for poorly related sequences, and $RC(X, X^C)$ is close to 1 for

similar sequences. We call $RC(X, X^C)$ the rate of coincidence (RC for short) for the code C. By calculating the RC for various codes, we can find a code structure characterizing the DNA-binding domain. If $RC(X, X^C)$ for a code C is bigger than that of any other codes, then we can infer that the DNA-binding domain has the property that the code C owns.

2.3 Survival analysis

The Kaplan–Meier method (Kaplan and Meier, 1958) was used to estimate the relapse-free survival rate and the overall survival rate of breast cancer patients with stage III or stage IV. The patients were stratified by mutation status in the L2 loop, the values of the RC in the L2 loop, mutation status in the L3 loop and the values of the RC in the L3 loop. The Wilcoxon test and the log-rank test were used to assess the difference between survival curves. For the two statistical tests, the P-values <0.05 were considered as statistically significant.

3 RESULTS

The Kaplan–Meier plots of breast cancer patients stratified by mutation status in the L2 loop are shown in Figure 1A and C. There was a statistically significant difference in the overall survival between the breast cancer patients with mutation in the L2 loop and those with wild-type (5-year overall survival: 44.9 versus 84.6%, Fig. 1A). In contrast, there was no significant difference in recurrence-free survival between two cases mentioned previously ($P = 0.137$ for the Wilcoxon test, $P = 0.06$ for the log-rank test, Fig. 1C).

The Kaplan–Meier plots of breast cancer patients stratified by the RC values in the L2 loop are shown in Figure 1B and D. The overall survival and recurrence-free survival were statistically significantly worse in patients with the L2 loop that is close to the (33, 22)-BCH code than patients with the L2 loop that is far from that code (5-year overall survival: 0 versus 80.3%, Fig. 1B; 5-year recurrence-free survival: 0 versus 41.3%, Fig. 1D). In this analysis, we made a correspondence between the four bases and the elements in $GF(4)$ as $\mathrm{A} \to 0$, $\mathrm{C} \to 1$, $\mathrm{G} \to a$ and $\mathrm{T} \to a^2$. The generator polynomial of the (33, 22)-BCH code was $G(x) = x^{11} + x^{10} + x^8 + x^7 + \alpha x^6 + \alpha x^5 + \alpha^2 x^4 + \alpha^2 x^3 + \alpha^2 x + \alpha^2$.

Moreover, the breast cancer survival curves according to mutation status in the L3 loop and the RC values in the L3

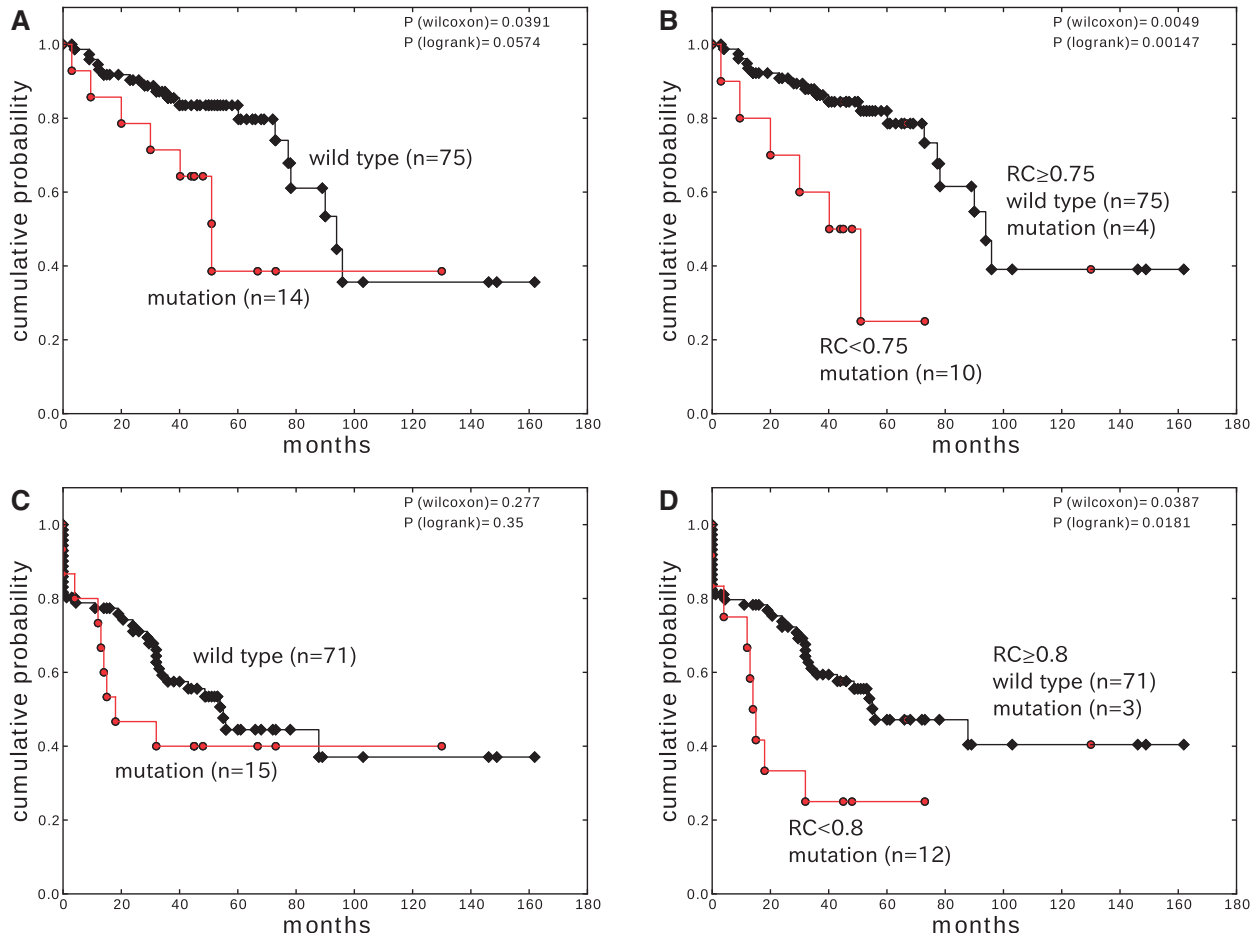

Fig. 2. Survival curves in the L3 loop. Kaplan–Meier survival curves in breast cancer patients stratified by mutation status in the L3 loop (left panels; A and C), and the RC values in the L3 loop (right panels; B and D) are shown: (**A**) Overall survival of patients with mutation in the L3 loop versus those with wild-type. (**B**) Overall survival of patients with the L3 loop that is close to the (24, 16)-BCH code versus those with the L3 loop that is far from that code. (**C**) Relapse-free survival of patients with mutation in the L3 loop versus those with wild-type. (**D**) Relapse-free survival of patients with the L3 loop that is close to the (6, 4)-BCH code versus those with the L3 loop that is far from that code

loop are shown in Figure 2. There was no statistically significant difference in overall survival and recurrence-free survival between the patients with mutation in the L3 loop and those with wild-type ($P=0.0574$ for the log-rank test, Fig. 2A; $P=0.277$ for the Wilcoxon test, $P=0.35$ for the log-rank test, Fig. 2C). In contrast, there was a significant difference in the overall survival between the patients with the L3 loop that is close to the (24, 16)-BCH code and the patients with the L3 loop that is far from that code (5-year overall survival: 82 versus 25%, Fig. 2B). The generator polynomial of the (24, 16)-BCH code was $G(x)=x^8+\alpha x^6+\alpha x^4+\alpha x^2+\alpha$, provided that the four bases were T→ 0, G→ 1, A→ a and C→ a^2.

In addition, there was a statistically significant difference in the recurrence-free survival between the patients with the L3 loop that is close to the (6, 4)-BCH code and the patients with the L3 loop that is far from that code (5-year overall survival: 47.2 versus 25%, Fig. 2D). The generator polynomial of the (6, 4)-BCH code was $G(x)=x^2+1$, provided that the four bases were G→ 0, A→ 1, T→ a and C→ a^2.

4 DISCUSSION

The tumor-suppressor gene TP53 mutations are diverse in the central region encoding the DNA-binding domain. It has not been clear whether the prognostic significance for survival in breast cancer patients is the same for all types of mutations. Are there specific types of mutations carrying a worse prognosis? We studied the code structure of the DNA-binding domain of breast cancer patients by using various artificial codes in information transmission.

In this study, we found a significant relation between the code structure of the DNA-binding domain and the breast cancer prognosis. Various types of mutations in the DNA-binding domain were classified by the code structure. We indicated that the prognostic significance of all types of mutations in the DNA-binding domain is not the same, and that the DNA-binding domain having a certain code structure is important for estimating the prognosis of breast cancer patients. We are convinced that the classification according to the code structure of the DNA-binding domain is useful for predicting patients who have a high mortality.

Although novel TP53 mutations are continuously reported, the biological function is incompletely understood for many mutant p53 proteins (Berge *et al*., 2013). Our classification may be also useful for establishing effect of individual mutation on protein properties such as cell cycle arrest, apoptosis and senescence.

Funding: Ministry of Education, Culture, Sports, Science and Technology Grants-in-Aid for Scientific Research [22740071].

Conflict of Interest: none declared.

REFERENCES

Berge,E.O. *et al*. (2013) Functional characterization of p53 mutants identified in breast cancers with suboptimal responses to anthracyclines or mitomycin. *Biochim. Biophys. Acta*, **1830**, 2790–2797.

Berns,E.M. *et al*. (1998) Mutations in residues of TP53 that directly contact DNA predict poor outcome in human primary breast cancer. *Br. J. Cancer*, **77**, 1130–1136.

Bertheau,P. *et al*. (2007) Exquisite sensitivity of TP53 mutant and basal breast cancers to a dose-dense epirubicin-cyclophosphamide regimen. *PLoS Med*., **4**, e90.

Brosh,R. and Rotter,V. (2009) When mutants gain new powers: news from the mutant p53 field. *Nat. Rev. Cancer*, **9**, 701–713.

Caleffi,M. *et al*. (1994) p53 gene mutations and steroid receptor status in breast cancer. Clinicopathologic correlations and prognostic assessment. *Cancer*, **73**, 2147–2156.

Chaitin,G.J. (1969) On the length of programs for computing finite binary sequences: statistical considerations. *J. ACM*, **16**, 145–159.

Chrisanthar,R. *et al*. (2008) CHEK2 mutations affecting kinase activity together with mutations in TP53 indicate a functional pathway associated with resistance to epirubicin in primary breast cancer. *PLoS One*, **3**, e3062.

Chrisanthar,R. *et al*. (2011) Predictive and prognostic impact of TP53 mutations and MDM2 promoter genotype in primary breast cancer patients treated with epirubicin or paclitaxel. *PLoS One*, **6**, e19249.

Geisler,S. *et al*. (2001) Influence of TP53 gene alterations and c-erbB-2 expression on the response to treatment with doxorubicin in locally advanced breast cancer. *Cancer Res*., **61**, 2505–2512.

Geisler,S. *et al*. (2003) TP53 gene mutations predict the response to neoadjuvant treatment with 5-fluorouracil and mitomycin in locally advanced breast cancer. *Clin. Cancer Res*., **9**, 5582–5588.

Gentile,M. *et al*. (1999) p53 and survival in early onset breast cancer: analysis of gene mutations, loss of heterozygosity and protein accumulation. *Eur. J. Cancer*, **35**, 1202–1207.

Ingarden,R.S. *et al*. (1997) *Information Dynamics and Open Systems: Classical and Quantum Approach*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Kaplan,E.L. and Meier,P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc*., **53**, 457–481.

Kucera,E. *et al*. (1999) Prognostic significance of mutations in the p53 gene, particularly in the zinc-bindingdomains, in lymph node- and steroid receptor positive breast cancer patients. Austrian Breast Cancer Study Group. *Eur. J. Cancer*, **35**, 398–405.

Langerød,A. *et al*. (2007) TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res*., **9**, R30.

Lothe,R.A. *et al*. (1995) Deletion of 1p loci and microsatellite instability in colorectal polyps. *Genes Chromosomes Cancer*, **14**, 182–188.

Makwane,N. and Saxena,A. (2009) Study of mutations in p53 tumour suppressor gene in human sporadic breast cancers. *Indian J. Clin. Biochem*., **24**, 223–228.

Murphy,K.L. and Rosen,J.M. (2000) Mutant p53 and genomic instability in a transgenic mouse model of breast cancer. *Oncogene*, **19**, 1045–1051.

Ohya,M. and Sato,K. (2000) Use of information theory to study genome sequences. *Rep. Math. Phys*., **46**, 419–428.

Ohya,M. and Volovich,I. (2011) *Mathematical Foundations of Quantum Information and Computation and Its Applications to Nano- and Bio-systems*. Springer Dordrecht Heidelberg London, New York.

Olivier,M. *et al*. (2006) The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. *Clin. Cancer Res*., **12**, 1157–1167.

Olivier,M. *et al*. (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol*., **2**, a001008.

Petitjean,A. *et al*. (2007) TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, **26**, 2157–2165.

Pharoah,P.D. *et al*. (1999) Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *Br. J. Cancer*, **80**, 1968–1973.

Powell,B. *et al*. (2000) Prognostic significance of mutations to different structural and functional regions of the p53 gene in breast cancer. *Clin. Cancer Res*., **6**, 443–451.

Suzuki,K. and Matsubara,H. (2011) Recent advances in p53 research and cancer treatment. *J. Biomed. Biotechnol*., **2011**, 978312.

Takahashi,M. *et al*. (2000) Distinct prognostic values of p53 mutations and loss of estrogen receptor and their cumulative effect in primary breast cancers. *Int. J. Cancer*, **89**, 92–99.

Varna,M. *et al*. (2011) TP53 status and response to treatment in breast cancers. *J. Biomed. Biotechnol*., **2011**, 284584.

Young,K.H. *et al*. (2007) Mutations in the DNA-binding codons of TP53, which are associated with decreased expression of TRAILreceptor-2, predict for poor survival in diffuse large B-cell lymphoma. *Blood*, **110**, 4396–4405.