Data in Brief 14 (2017) 538–542

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Prediction of transcription factor bindings sites affected by SNPs located at the osteopontin promoter

Marco Antonio Briones-Orta [a], S. Eréndira Avendaño-Vázquez [b,*], Diana Ivette Aparicio-Bautista [b], Jason D. Coombes [a], Georg F. Weber [c], Wing-Kin Syn [a,d,e,**]

[a] *Regeneration and Repair, Institute of Hepatology, Foundation for Liver Research, London, United Kingdom*
[b] *Instituto Nacional de Medicina Genómica, INMEGEN, Periférico Sur 4809, Ciudad de México 14610,. México*
[c] *James L. Winkle College of Pharmacy, University of Cincinnati Academic Health Center, Cincinnati, OH, United States*
[d] *Division of Gastroenterology and Hepatology, Department of Medicine, Medical University of South Carolina, Charleston, SC, United States*
[e] *Section of Gastroenterology, Ralph H Johnson Veteran Affairs Medical Center, Charleston, SC, United States*

ARTICLE INFO

Article history:
Received 7 July 2017
Accepted 21 July 2017
Available online 2 August 2017

ABSTRACT

This data contains information related to the research article entitled “Osteopontin splice variants and polymorphisms in Cancer Progression and Prognosis” [1]. Here, we describe an in silico analysis of transcription factors that could have altered binding to their DNA target sequence as a result of SNPs in the osteopontin gene promoter. We concentrated on SNPs associated with cancer risk and development.

The analysis was performed with PROMO v3.0.2 software which incorporates TRANSFACT v6.4 of. We also present a figure depicting the putative transcription factor binding according to genotype.

DOI of original article: http://dx.doi.org/10.1016/j.bbcan.2017.02.005

* Corresponding author.

** Corresponding author at: Regeneration and Repair, Institute of Hepatology, Foundation for Liver Research, London, United Kingdom.

E-mail addresses: seavendano@inmegen.gob.mx (S.E. Avendaño-Vázquez), synw@musc.edu (W.-K. Syn).

http://dx.doi.org/10.1016/j.dib.2017.07.057

Specifications Table

Subject area	*Biology, Molecular Biology*
More specific subject area	*Effect of SNPs in binding of transcription factors for the gene osteopontin*
Type of data	*Table and figure*
How data was acquired	*Software PROMO 3.0.2* (using TRANSFAC v.6.4)
Data format	*Analyzed*
Experimental factors	*SNPs sequences were obtained from NCBI Single Nucleotide Polymorphism Database (dbSNP). PROMO parameters were chosen for human sequences and human sites.*
Experimental features	SNPs located in OPN promoter with an effect in cancer risk and prognosis were analyzed to compare which transcription factors are binding in the variant sequences.
Data source location	
Data accessibility	*The data is available in this article*

Value of the data

- These data describe how putative DNA-binding sites for transcriptional factors can be created or interrupted by the changes in sequences generated by SNPs in the promoter of osteopontin.
- Differential binding among SNPs genotypes can potentially explain why these SNPs have been associated with changes in the risk of cancer for a specific population.
- This analysis is an example of how important databases, such as those containing SNP genotypes and the predictive tools for DNA-binding sites for transcriptional factors in a specific sequence, could be used to try to select potential signaling pathways modulating the development of cancer.

1. Data

The table provided in this article is a list of the transcription factors predicted to bind a DNA sequence at the SNPs contained in the osteopontin promoter. We analyzed only those SNPs that statistically in a population have been shown to have an effect on cancer risk and prognosis for the carriers. For each SNP we present both sequences. Each analysis contains the rs ID and the nucleotide position in reference to the osteopontin promoter; a schematic representation of the binding of the transcription factor to their target sequence; and an analysis of how similar the binding site is compared to its canonical binding sequence.

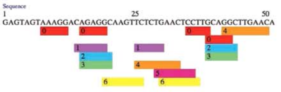
2. Experimental design, materials and methods

Analysis of SNP sequences was performed using software PROMO v3.0.2, (which utilizes TRANS FAC v6.4) [2,3] For each osteopontin gene promoter SNP, the sequences carrying each allele were loaded as the query sequence to search for potential binding sites. The prediction was carried out considering only sites and only human transcription factors. The output of this analysis is presented in Table 1. Each analysis contains the rs that corresponds to each SNP and its position relative to the transcription start site of osteopontin. For each SNP, we present the respective results for both sequences loaded as the query sequences. A schematic representation (boxes in color, also indicated with numbers) of the binding of the transcription factor to the target sequence, and a list of the putative transcription factors binding to the sequence. For each transcription factor site, several predicted parameters are reported. The *transcription*

Table 1
Transcription factors binding prediction to sequences associated to SNPs genotypes located in the promoter of the osteopontin gene.

rs11730582
GAGTAGTAAAGGACAGAGGCAAGTT[C/T]TCTGAACTCCTTGCAGGCTTGAACA

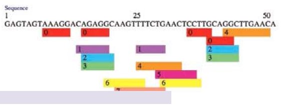
-443 C
GAGTAGTAAAGGACAGAGGCAAGTTCTCTGAACTCCTTGCAGGCTTGAACA



GR-alpha	0
RFX1	1
AP-2alphaA	2
AP-2alphaB	3
VDR	4
RXR-alpha	5
c-Myb	6

Factor Name	Start position	End position	Dissimilarity	Sequence	RE equally	RE query
GR-alpha [T00337]	7	11	0.207689	AAAGG	0.19922	0.23474
GR-alpha [T00337]	14	18	0.207689	AGAGG	0.19922	0.23474
GR-alpha [T00337]	34	38	8.281568	CCTTG	0.19922	0.20785
GR-alpha [T00337]	38	42	8.073878	GCAGG	0.19922	0.17535
RFX1 [T01673]	13	18	0	CAGAGG	0.03735	0.0389
RFX1 [T01673]	24	29	0	TCTCTG	0.03735	0.0389
AP-2alphaA [T00035]	14	19	2.551556	AGAGGC	0.01245	0.01263
AP-2alphaA [T00035]	38	43	0	GCAGGC	0.0249	0.01841
AP-2alphaB [T02466]	14	19	2.786893	AGAGGC	0.01245	0.01263
AP-2alphaB [T02466]	38	43	0.302681	GCAGGC	0.0249	0.01948
VDR [T00885]	24	32	8.079962	TCTCTGAAC	0.00623	0.00652
VDR [T00885]	41	49	5.771401	GGCTTGAAC	0.01089	0.01092
RXR-alpha [T01345]	28	35	1.688117	TGAACTCC	0.00311	0.00296
c-Myb [T00137]	18	25	4.840682	GCAAGTTC	0.00778	0.00731
c-Myb [T00137]	29	36	12.979731	GAACTCCT	0.01401	0.01446

-443 T
GAGTAGTAAAGGACAGAGGCAAGTTTTCTGAACTCCTTGCAGGCTTGAACA

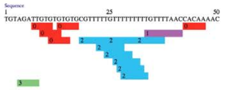


GR-alpha	0
RFX1	1
AP-2alphaA	2
AP-2alphaB	3
VDR	4
RXR-alpha	5
c-Myb	6
STAT6	7

Factor Name	Start position	End position	Dissimilarity	Sequence	RE equally	RE query
GR-alpha [T00337]	7	11	0.207689	AAAGG	0.19922	0.23309
GR-alpha [T00337]	14	18	0.207689	AGAGG	0.19922	0.23309
GR-alpha [T00337]	34	38	8.281568	CCTTG	0.19922	0.19462
GR-alpha [T00337]	38	42	8.073878	GCAGG	0.19922	0.16152
RFX1 [T01673]	13	18	0	CAGAGG	0.03735	0.03624
RFX1 [T01673]	24	29	9.512894	TTTCTG	0.18677	0.16869
AP-2alphaA [T00035]	14	19	2.551556	AGAGGC	0.01245	0.01129
AP-2alphaA [T00035]	38	43	0	GCAGGC	0.0249	0.01482
AP-2alphaB [T02466]	14	19	2.786893	AGAGGC	0.01245	0.01129
AP-2alphaB [T02466]	38	43	0.302681	GCAGGC	0.0249	0.01647
VDR [T00885]	24	32	9.234242	TTTCTGAAC	0.00311	0.00335
VDR [T00885]	41	49	5.771401	GGCTTGAAC	0.01089	0.01119
RXR-alpha [T01345]	28	35	1.688117	TGAACTCC	0.00311	0.00304
c-Myb [T00137]	18	25	6.12608	GCAAGTTT	0.00545	0.00577
c-Myb [T00137]	29	36	12.979731	GAACTCCT	0.01401	0.01552
STAT6 [T01580]	20	29	11.888653	AAGTTTTCTC	0.00389	0.00644

rs17524488
TGTAGATTGTGTGTGTGCGTTTTTG[-/G]TTTTTTTTTGTTTTAACCACAAAAC

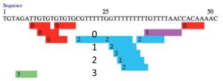
-156 G
TGTAGATTGTGTGTGTGCGTTTTTGTTTTTTTTTGTTTTAACCACAAAAC



USF1	0
HNF-1A	1
GR	2
GR-beta	3

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
USF1 [T00874]	6	10	7.629649	TTGTG	0.04883	0.14207
USF1 [T00874]	8	12	4.958698	GTGTG	0.04883	0.05489
USF1 [T00874]	10	14	4.958698	GTGTG	0.04883	0.05489
USF1 [T00874]	12	16	4.958698	GTGTG	0.04883	0.05489
USF1 [T00874]	42	46	7.629649	CACAA	0.04883	0.14207
HNF-1A [T00368]	33	41	8.337414	TGTTTTAAC	0.00687	0.00917
GR [T05076]	17	23	9.33358	CGTTTTT	0.04883	0.76283
GR [T05076]	21	27	9.33358	TTTGTTT	0.04883	0.76283
GR [T05076]	23	29	9.33358	TGTTTTT	0.04883	0.76283
GR [T05076]	24	30	8.033921	GTTTTTT	0.03662	0.41228
GR [T05076]	25	31	9.33358	TTTTTTT	0.04883	0.76283
GR [T05076]	26	32	9.33358	TTTTTTT	0.04883	0.76283
GR [T05076]	27	33	9.33358	TTTTTTT	0.04883	0.76283
GR [T05076]	31	37	9.33358	TTTGTTT	0.04883	0.76283
GR-beta [T01920]	3	7	3.361531	AGATT	0.09766	0.21902

-156 GG
TGTAGATTGTGTGTGTGCGTTTTTGGTTTTTTTTTGTTTTAACCACAAAAC

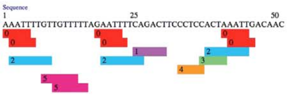


USF1	0
HNF-1A	1
GR	2
GR-beta	3

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
USF1 [T00874]	6	10	7.629649	TTGTG	0.0498	0.15853
USF1 [T00874]	8	12	4.958698	GTGTG	0.0498	0.06725
USF1 [T00874]	10	14	4.958698	GTGTG	0.0498	0.06725
USF1 [T00874]	12	16	4.958698	GTGTG	0.0498	0.06725
USF1 [T00874]	43	47	7.629649	CACAA	0.0498	0.15853
HNF-1A [T00368]	34	42	8.337414	TGTTTTAAC	0.007	0.00944
GR [T05076]	17	23	9.33358	CGTTTTT	0.0498	0.71543
GR [T05076]	24	30	8.033921	GGTTTTT	0.03735	0.4056
GR [T05076]	25	31	8.033921	GTTTTTT	0.03735	0.4056
GR [T05076]	26	32	9.33358	TTTTTTT	0.0498	0.71543
GR [T05076]	27	33	9.33358	TTTTTTT	0.0498	0.71543
GR [T05076]	28	34	9.33358	TTTTTTT	0.0498	0.71543
GR [T05076]	32	38	9.33358	TTTGTTT	0.0498	0.71543
GR-beta [T01920]	3	7	3.361531	AGATT	0.09961	0.20639

rs2728127
AAATTTTGTTGTTTTTAGAATTTTC[A/G]GACTTCCCTCCACTAAATTGACAAC

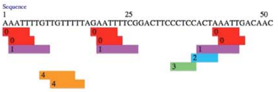
-1748 A
AAATTTTGTTGTTTTTAGAATTTTCAGACTTCCCTCCACTAAATTGACAAC



GR-beta	0
RFX1	1
HNF-3alpha	2
USF1	3
GR-alpha	4
GR	5

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
GR-beta [T01920]	0	4	0	AAATT	0.09961	0.37401
GR-beta [T01920]	1	5	0	AATTT	0.09961	0.37401
GR-beta [T01920]	17	21	1.680765	GAATT	0.09961	0.15191
GR-beta [T01920]	18	22	0	AATTT	0.09961	0.37401
GR-beta [T01920]	40	44	0	AAATT	0.09961	0.37401
GR-beta [T01920]	41	45	0.840383	AATTG	0.19922	0.56049
RFX1 [T01673]	24	29	9.512894	CAGACT	0.18677	0.09283
HNF-3alpha [T02512]	1	8	7.000129	AATTTTGT	0.02101	0.12839
HNF-3alpha [T02512]	18	25	3.500065	AATTTTCA	0.007	0.07181
HNF-3alpha [T02512]	37	44	10.500194	ACTAAATT	0.03035	0.13787
USF1 [T00874]	36	40	8.105784	CACTA	0.09961	0.06785
GR-alpha [T00337]	32	36	8.281568	CCTCC	0.19922	0.08049
GR [T05076]	7	13	8.033921	GTTGTTT	0.03735	0.1369
GR [T05076]	9	15	9.33358	TGTTTTT	0.0498	0.24844

-1748 G
AAATTTTGTTGTTTTTAGAATTTTCGGACTTCCCTCCACTAAATTGACAAC



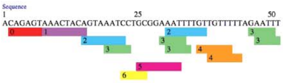
GR-beta	0
HNF-3alpha	1
USF1	2
GR-alpha	3
GR	4

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
GR-beta [T01920]	0	4	0	AAATT	0.09961	0.3258
GR-beta [T01920]	1	5	0	AATTT	0.09961	0.3258
GR-beta [T01920]	17	21	1.680765	GAATT	0.09961	0.14784
GR-beta [T01920]	18	22	0	AATTT	0.09961	0.3258
GR-beta [T01920]	40	44	0	AAATT	0.09961	0.3258
GR-beta [T01920]	41	45	0.840383	AATTG	0.19922	0.50376
HNF-3alpha [T02512]	1	8	7.000129	AATTTTGT	0.02101	0.11327
HNF-3alpha [T02512]	18	25	10.500194	AATTTTCG	0.03035	0.12527
HNF-3alpha [T02512]	37	44	10.500194	ACTAAATT	0.03035	0.12527
USF1 [T00874]	36	40	8.105784	CACTA	0.09961	0.06837
GR-alpha [T00337]	32	36	8.281568	CCTCC	0.19922	0.0889
GR [T05076]	7	13	8.033921	GTTGTTT	0.03735	0.13478
GR [T05076]	9	15	9.33358	TGTTTTT	0.0498	0.24569

Table 1 (*continued*)

rs29001511
ACAGAGTAAACTACAGTAAATCCTG[C/T]GGAAATTTTGTTGTTTTTAGAATTT

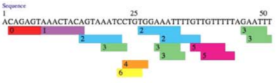
-1776 C
ACAGAGTAAACTACAGTAAATCCTGCGGAAATTTTGTTGTTTTTAGAATTT



RFX1	0
c-Myb	1
HNF-3alpha	2
GR-beta	3
GR	4
E2F-1	5
GR-alpha	6

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
RFX1 [T01673]	1	6	0	CAGAGT	0.03735	0.01628
c-Myb [T00137]	7	14	14.265129	AAACTACA	0.00934	0.01497
HNF-3alpha [T02512]	14	21	14.000258	AGTAAATC	0.04202	0.1033
HNF-3alpha [T02512]	29	36	7.000129	AATTTTGT	0.02101	0.12436
GR-beta [T01920]	18	22	5.042296	AATCC	0.09961	0.1703
GR-beta [T01920]	28	32	0	AAATT	0.09961	0.39325
GR-beta [T01920]	29	33	0	AATTT	0.09961	0.39325
GR-beta [T01920]	45	49	1.680765	GAATT	0.09961	0.16207
GR-beta [T01920]	46	50	0	AATTT	0.09961	0.39325
GR [T05076]	35	41	8.033921	GTTGTTT	0.03735	0.13533
GR [T05076]	37	43	9.33358	TGTTTTT	0.0498	0.20531
E2F-1 [T01542]	24	31	5.846171	GCGGAAAT	0.00467	0.0009
GR-alpha [T00337]	21	25	8.073878	CCTGC	0.19922	0.05788

-1776 T
ACAGAGTAAACTACAGTAAATCCTGTGGAAATTTTGTTGTTTTTAGAATTT

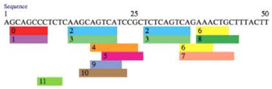


RFX1	0
c-Myb	1
HNF-3alpha	2
GR-beta	3
USF1	4
GR	5
GR-alpha	6

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
RFX1 [T01673]	1	6	0	CAGAGT	0.03735	0.01328
c-Myb [T00137]	7	14	14.265129	AAACTACA	0.00934	0.01515
HNF-3alpha [T02512]	14	21	14.000258	AGTAAATC	0.04202	0.10736
HNF-3alpha [T02512]	25	32	10.500194	TGGAAATT	0.03035	0.14956
HNF-3alpha [T02512]	29	36	7.000129	AATTTTGT	0.02101	0.14287
GR-beta [T01920]	18	22	5.042296	AATCC	0.09961	0.17816
GR-beta [T01920]	28	32	0	AAATT	0.09961	0.43573
GR-beta [T01920]	29	33	0	AATTT	0.09961	0.43573
GR-beta [T01920]	45	49	1.680765	GAATT	0.09961	0.16145
GR-beta [T01920]	46	50	0	AATTT	0.09961	0.43573
USF1 [T00874]	22	26	6.294173	CTGTG	0.09961	0.03356
GR [T05076]	35	41	8.033921	GTTGTTT	0.03735	0.15702
GR [T05076]	37	43	9.33358	TGTTTTT	0.0498	0.23853
GR-alpha [T00337]	21	25	0	CCTGT	0.19922	0.12593

rs2853744
AGCAGCCCTCTCAAGCAGTCATCCT[G/T]CTCTCAGTCAGAAACTGCTTTACTT

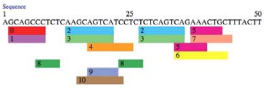
-719 G
AGCAGCCCTCTCAAGCAGTCATCCGCTCTCAGTCAGAAACTGCTTTACTT



Pax-5	0
p53	1
AP-1	2
c-Jun	3
PEA3	4
E2F-1	5
RFX1	6
STAT6	7
c-Myb	8
XBP-1	9
X2BP	10
GR-alpha	11

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
Pax-5 [T00070]	1	7	8.014558	GCAGCCC	0.05493	0.04514
p53 [T00671]	1	7	6.563521	GCAGCCC	0.01221	0.01015
AP-1 [T00029]	12	20	10.480716	AAGCAGTCA	0.01068	0.00925
AP-1 [T00029]	26	34	6.527374	TCTCAGTCA	0.00458	0.00406
c-Jun [T00133]	12	20	12.853308	AAGCAGTCA	0.0061	0.00524
c-Jun [T00133]	26	34	8.71589	TCTCAGTCA	0.00191	0.00171
PEA3 [T00685]	16	24	10.535275	AGTCATCCG	0.00687	0.00932
E2F-1 [T01542]	18	25	10.630946	TCATCCGC	0.01526	0.01801
RFX1 [T01673]	33	38	9.512894	CAGAAA	0.18311	0.16592
RFX1 [T01673]	36	41	14.26934	AAACTG	0.14648	0.12731
STAT6 [T01580]	33	42	0	CAGAAACTGC	0.00019	0.00014
c-Myb [T00137]	36	43	7.545286	AAACTGCT	0.01068	0.01095
GR-alpha [T00337]	6	10	0.207689	CCTCT	0.19531	0.23223
XBP-1 [T00902]	16	21	0	AGTCAT	0.02441	0.02135
X2BP [T01108]	14	22	13.110131	GCAGTCATC	0.01869	0.02019

-719 T
AGCAGCCCTCTCAAGCAGTCATCCTCTCTCAGTCAGAAACTGCTTTACTT

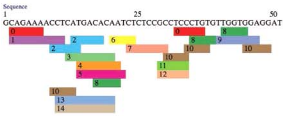


Pax-5	0
p53	1
AP-1	2
c-Jun	3
PEA3	4
RFX1	5
STAT6	6
c-Myb	7
GR-alpha	8
XBP-1	9
X2BP	10

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
Pax-5 [T00070]	1	7	8.014558	GCAGCCC	0.05493	0.0363
p53 [T00671]	1	7	6.563521	GCAGCCC	0.01221	0.00781
AP-1 [T00029]	12	20	10.480716	AAGCAGTCA	0.01068	0.00913
AP-1 [T00029]	26	34	6.527374	TCTCAGTCA	0.00458	0.0042
c-Jun [T00133]	12	20	12.853308	AAGCAGTCA	0.0061	0.00504
c-Jun [T00133]	26	34	8.71589	TCTCAGTCA	0.00191	0.00167
PEA3 [T00685]	16	24	4.30818	AGTCATCCT	0.00343	0.00502
RFX1 [T01673]	33	38	9.512894	CAGAAA	0.18311	0.1551
RFX1 [T01673]	36	41	14.26934	AAACTG	0.14648	0.11608
STAT6 [T01580]	33	42	0	CAGAAACTG	0.00019	0.00013
c-Myb [T00137]	36	43	7.545286	AAACTGCT	0.01068	0.01161
GR-alpha [T00337]	6	10	0.207689	CCTCT	0.19531	0.2548
GR-alpha [T00337]	22	26	0.207689	CCTCT	0.19531	0.2548
XBP-1 [T00902]	16	21	0	AGTCAT	0.02441	0.02105
X2BP [T01108]	14	22	13.110131	GCAGTCATC	0.01869	0.02252

rs28357094
GCAGAAAACCTCATGACACAATCTC[G/T]CCGCCTCCCTGTGTTGGTGGAGGAT

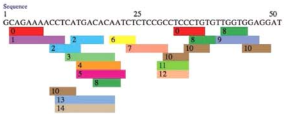
-66 T
GCAGAAAACCTCATGACACAATCTCTCCGCCTCCCTGTGTTGGTGGAGGAT



RFX1	0
STAT6	1
XBP-1	2
X2BP	3
ATF3	4
C/EBPalpha	5
GR-beta	6
E2F-1	7
USF1	8
NF-1	9
GR-alpha	10
AP-2alphaA	11
AP-2alphaB	12
Fra-1	13
JunB	14

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
RFX1 [T01673]	1	6	9.512894	CAGAAA	0.18677	0.19143
RFX1 [T01673]	31	36	9.512894	TCCCTG	0.18677	0.19143
STAT6 [T01580]	1	10	14.286065	CAGAAAACCT	0.00389	0.00308
XBP-1 [T00902]	8	13	7.365101	CCTCAT	0.14941	0.14023
XBP-1 [T00902]	12	17	1.626297	ATGACA	0.0249	0.02299
X2BP [T01108]	11	19	7.615488	CATGACACA	0.00545	0.00499
ATF3 [T01313]	13	20	6.744803	TGACACAA	0.007	0.00634
C/EBPalpha [T00105]	13	21	3.367013	TGACACAAT	0.00078	0.00065
GR-beta [T01920]	19	23	3.361531	AATCT	0.09961	0.07816
E2F-1 [T01542]	22	29	11.323028	CTCTCCGC	0.01245	0.01598
USF1 [T00874]	16	20	7.629649	CACAA	0.0498	0.04713
USF1 [T00874]	34	38	6.294173	CTGTG	0.09961	0.10345
USF1 [T00874]	40	44	8.105784	TGGTG	0.09961	0.10144
NF-1 [T00539]	39	46	13.670907	TTGGTGGA	0.0249	0.02574
GR-alpha [T00337]	8	12	6.263098	CCTCA	0.09961	0.10144
GR-alpha [T00337]	29	33	8.281568	CCTCC	0.19922	0.22823
GR-alpha [T00337]	33	37	0	CCTGT	0.19922	0.20028
GR-alpha [T00337]	44	48	8.281568	GGAGG	0.19922	0.22823
AP-2alphaA [T00035]	28	33	1.422205	GCCTCC	0.0249	0.03119
AP-2alphaB [T02466]	28	33	1.576169	GCCTCC	0.03735	0.04471
Fra-1 [T01462]	9	19	14.261792	CTCATGACAC	0.0019	0.00175
JunB [T01977]	9	19	14.261792	CTCATGACAC	0.0019	0.00175

-66 G
GCAGAAAACCTCATGACACAATCTCGCCGCCTCCCTGTGTTGGTGGAGGAT



RFX1	0
STAT6	1
XBP-1	2
X2BP	3
ATF3	4
C/EBPalpha	5
GR-beta	6
E2F-1	7
USF1	8
NF-1	9
GR-alpha	10
AP-2alphaA	11
AP-2alphaB	12
Fra-1	13
JunB	14

Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
RFX1 [T01673]	1	6	9.512894	CAGAAA	0.18677	0.19523
RFX1 [T01673]	31	36	9.512894	TCCCTG	0.18677	0.19523
STAT6 [T01580]	1	10	14.286065	CAGAAAACC	0.00389	0.0026
XBP-1 [T00902]	8	13	7.365101	CCTCAT	0.14941	0.13337
XBP-1 [T00902]	12	17	1.626297	ATGACA	0.0249	0.02193
X2BP [T01108]	11	19	7.615488	CATGACACA	0.00545	0.00472
ATF3 [T01313]	13	20	6.744803	TGACACAA	0.007	0.00595
C/EBPalpha [T00105]	13	21	3.367013	TGACACAAT	0.00078	0.00057
GR-beta [T01920]	19	23	3.361531	AATCT	0.09961	0.06568
E2F-1 [T01542]	22	29	11.323028	CTCGCCGC	0.01245	0.01898
USF1 [T00874]	16	20	7.629649	CACAA	0.0498	0.04536
USF1 [T00874]	34	38	6.294173	CTGTG	0.09961	0.1076
USF1 [T00874]	40	44	8.105784	TGGTG	0.09961	0.10004
NF-1 [T00539]	39	46	13.670907	TTGGTGGA	0.0249	0.02536
GR-alpha [T00337]	8	12	6.263098	CCTCA	0.09961	0.09829
GR-alpha [T00337]	29	33	8.281568	CCTCC	0.19922	0.23809
GR-alpha [T00337]	33	37	0	CCTGT	0.19922	0.19435
GR-alpha [T00337]	44	48	8.281568	GGAGG	0.19922	0.23809
AP-2alphaA [T00035]	28	33	1.422205	GCCTCC	0.0249	0.03563
AP-2alphaB [T02466]	28	33	1.576169	GCCTCC	0.03735	0.05023
Fra-1 [T01462]	9	19	14.261792	CTCATGACA	0.0019	0.00166
JunB [T01977]	9	19	14.261792	CTCATGACA	0.0019	0.00166

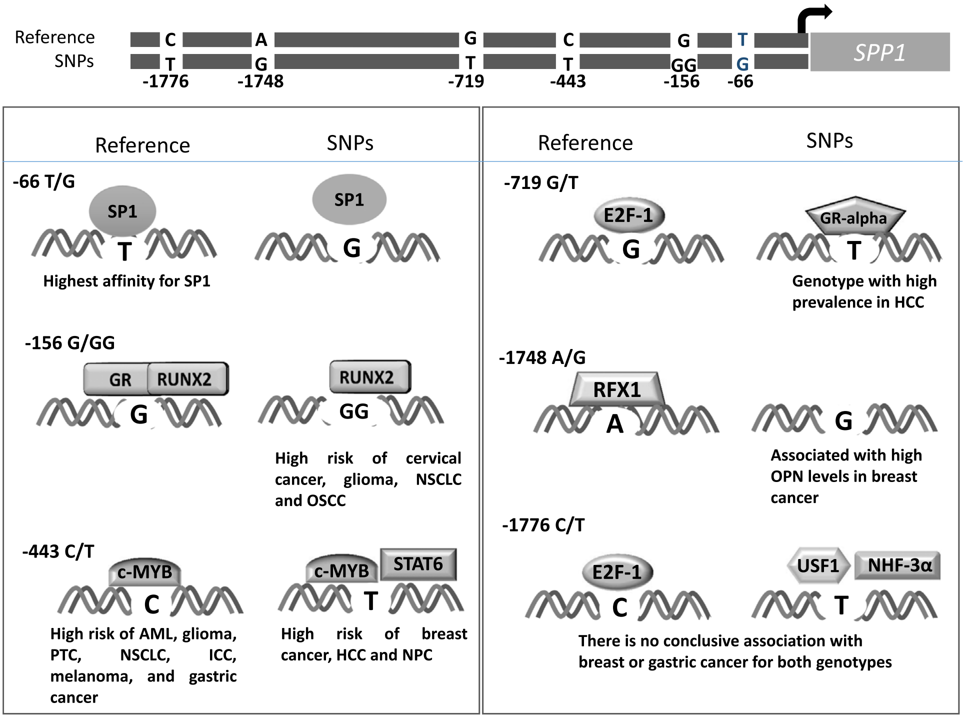

Reference
SNPs
C
T
-1776
A
G
-1748
G
T
-719
C
T
-443
G
GG
-156
T
G
-66
SPP1
Reference
SNPs
-66 T/G
SP1
T
SP1
G
Highest affinity for SP1
-156 G/GG
GR
RUNX2
G
RUNX2
GG
High risk of cervical cancer, glioma, NSCLC and OSCC
-443 C/T
c-MYB
C
c-MYB
STAT6
T
High risk of AML, glioma, PTC, NSCLC, ICC, melanoma, and gastric cancer
High risk of breast cancer, HCC and NPC
Reference
SNPs
-719 G/T
E2F-1
G
GR-alpha
T
Genotype with high prevalence in HCC
-1748 A/G
RFX1
A
G
Associated with high OPN levels in breast cancer
-1776 C/T
E2F-1
C
USF1
NHF-3α
T
There is no conclusive association with breast or gastric cancer for both genotypes

Fig. 1. Schematic representation of changes in transcription factors binding to SNPs located in the promoter of the osteopontin gene. At the top of the image there is a representation of the SNPs located in the osteopontin promoter that have been linked to variation in cancer risk in the carriers. The position of each SNP is given with respect to the transcription starting point. Below, for each SNP, the binding of the transcription factors and changes associated with altered genotype are exemplified.

factor name with the database accession number in brackets; the *start* and *end* positions of the putative binding sequences; *Dissimilarity* (%), which corresponds to the rate of dissimilarity between the putative and consensus sequences for a given transcription factor; *Sequence*, the nucleotide sequence of potential binding site; *Random Expectation (RE)* indicating the expected occurrences of the match in a random sequence of the same length as the query sequence according to the dissimilarity index, presented the *RE equally* (equi-probability for the four nucleotides) and *RE query* (nucleotide frequencies as in the query sequence). Markedly different changes are highlight in grey and the SNP is highlight in red. In Fig. 1 we depict the integration of information obtained from this predictive analysis and data previously reported for transcription factors binding to the osteopontin promoter.

References

[1] M.A. Briones-Orta, S.E. Avendaño-Vázquez, et al., Osteopontin Splice Varinats and Polymorphisms in Cancer Progression and Prognosis. *Biochim Biophys Acta.* **1868**, 2017, 93–108.

[2] D. Farré, R. Roset, M. Huerta, J.E. Adsuara, L. Roselló, M.M. Albà, X. Messeguer, Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN, Nucleic Acids Res. 31 (2003) 3651–3653.

[3] X. Messeguer, R. Escudero, D. Farré, O. Núñez, J. Martínez, M.M. Albà, PROMO: detection of known transcription regulatory elements using species-tailored searches, Bioinformatics 18 (2002) 333–334.