Zhang et al. Human Genomics (2017) 11:35
DOI 10.1186/s40246-017-0131-5

Human Genomics

PRIMARY RESEARCH Open Access

Overexpressed HSF1 cancer signature genes cluster in human chromosome 8q

Christopher Q. Zhang[1,4], Heinric Williams[3,4], Thomas L. Prince[3,4†] and Eric S. Ho[1,2*†]

Abstract

Background: HSF1 (heat shock factor 1) is a transcription factor that is found to facilitate malignant cancer development and proliferation. In cancer cells, HSF1 mediates a set of genes distinct from heat shock that contributes to malignancy. This set of genes is known as the HSF1 Cancer Signature genes or simply HSF1-CanSig genes. HSF1-CanSig genes function and operate differently than typical cancer-causing genes, yet it is involved in fundamental oncogenic processes.

Results: By utilizing expression data from 9241 cancer patients, we identified that human chromosome 8q21-24 is a location hotspot for the most frequently overexpressed HSF1-CanSig genes. Intriguingly, the strength of the HSF1 cancer program correlates with the number of overexpressed HSF1-CanSig genes in 8q, illuminating the essential role of HSF1 in mediating gene expression in different cancers. Chromosome 8q21-24 is found under selective pressure in preserving gene order as it exhibits strong synteny among human, mouse, rat, and bovine, although the biological significance remains unknown. Statistical modeling, hierarchical clustering, and gene ontology-based pathway analyses indicate crosstalk between HSF1-mediated responses and pre-mRNA 3′ processing in cancers.

Conclusions: Our results confirm the unique role of chromosome 8q mediated by the master regulator HSF1 in cancer cases. Additionally, this study highlights the connection between cellular processes triggered by HSF1 and pre-mRNA 3′ processing in cancers.

Keywords: Heat shock factor 1 (HSF1), HSF1 cancer signature (CanSig), Malignancy, Pre-mRNA 3′ processing, Chromosome 8q, Synteny

Introduction

Heat-shock factor 1 (HSF1) is a master transcription factor that initiates the expression of heat shock proteins (HSPs) and other genes in response to cellular stress, thus allowing cells to adapt and prolong survival. Cancer cells, however, hijack this protective mechanism to allow them to continue proliferating in their toxic microenvironments. Several studies have linked increased HSF1 activity to malignant cell growth [1]. This pro-malignant activity is associated with HSF1 binding to the promoters and initiating the expression of certain genes independent of heat shock [1]. Identified by Mendillo, Santagata, and coworkers, the HSF1 Cancer Signature (HSF1-CanSig) is a set of 475 genes overexpressed in a number of highly malignant cancer cells and primary tumors [1].

HSF1 itself is observed to be overexpressed in across different tumor types and to promote proliferation, migration, and invasion [2–5]. Frequently, overexpressed genes across different tumors or cancer types are understood to be likely oncogenic and contribute to tumorigenesis [6]. Consequently, this drives the overexpression of the HSF1-CanSig in tumor cells and possibly the surrounding stroma leading to metastasis and poor clinical outcomes [1, 2]. The biological processes and features that link the HSF1-CanSig to malignant cell growth, however, are still unclear.

To elucidate the role of the HSF1-CanSig in cancer, we mined and analyzed the overexpression in 9241 cancer cases from 27 unique primary tumor sites from The Cancer Genome Atlas (TCGA) hosted in cBioPortal [7]. We found that 27 of the top 100 most frequently overexpressed HSF1-CanSig genes are clustered in a highly syntenic region of chromosome 8q. We observed that

* Correspondence: hoe@lafayette.edu
†Equal contributors
[1]Department of Biology, Lafayette College, Easton, PA 18042, USA
[2]Department of Computer Science, Lafayette College, Easton, PA 18042, USA
Full list of author information is available at the end of the article

this HSF1-driven chromosome 8q gene is set to be overexpressed in a majority of cancer types. Furthermore, our gene ontology analysis indicates a link between HSF1-driven transcription and mRNA processing located on chromosome 8q.

Results

Top 100 high-scoring HSF1-CanSig genes are disproportionately located in chromosome 8q

We devised an overexpression score to quantify the prevalence of HSF1-CanSig genes being overexpressed among cancer cases (see the "Materials and methods" section for details). The overexpression score reflects the prevalence of a gene being overexpressed ≥ 2 standard deviations above the control reference as calculated by cBioPortal for different primary tumor sites [8]. A gene associated with a high overexpression score indicates that the gene is often upregulated in cancer cases. Overexpression scores of all 475 HSF1-CanSig genes from different primary tumor sites were aggregated, ranked, and examined by their percentage distribution per chromosome arm in tiers: top 100, 200, 300, and all HSF1-CanSig genes. If ranking has no effect on distribution, the distribution of genes in each tier should resemble the distribution of all HSF1-CanSig genes, meaning that they distribute like the result of random sampling of HSF1-CanSig genes. Intriguingly, 27% of the top 100 groups comprises of genes encoded in chromosome 8q (Fig. 1a, leftmost light blue bar) given that 8q consists of only 6% (29 of 475) of all HSF1-CanSig (leftmost gray bar). Such percentage declines continuously with from the top 200 group to top 300 group.

To assess the statistical significance of the distribution exhibited in Fig. 1a, we tested the probability of observing certain number of genes per chromosome arm in different tiers by hypergeometric test and Fisher's exact test (Additional file 1: Table S1). The four chromosome arms that encode the highest number of HSF1-CanSig genes are 1p (30), 8q (29), 11q (29), and 17q (30). When genes were examined in tiers by ranking, HSF1-CanSig genes encoded in 8q are included most disproportionately in different tiers but not for genes in the other three chromosome arms, i.e., 1p, 11q, and 17q. The group that attained the smallest p value of Fisher's test was the top 100 group ($5.41e^{-20}$), whereas hypergeometric test found that the top 50 group achieved the lowest p value ($3.13e^{-18}$) followed by the top 100 group ($1.09e^{-17}$). But the former captures only 22 8q genes out of 29 versus 27 8q genes by the latter. To balance statistical rigor and coverage, we decided to focus on the top 100 frequently overexpressed HSF1-CanSig genes. Importantly, regardless of which tier we choose, the most striking finding is that a large proportion of top-ranking genes are encoding in chromosome 8q as supported by Additional file 1: Table S1.

Next, we examine whether the skew distribution of 8q genes in the top 100 group is influenced by the intrinsic distribution of cancer or protein-coding genes. To rule out this possibility, the aggregated overexpression scores of the top 100 high-scoring HSF1-CanSig genes were compared with two null models or references: census cancer gene and all protein-coding genes, which consist of only 2–3% of 8q genes (gray and black bars in Fig. 1b, respectively). A high correlation between the two references based on the percentage of genes per chromosome arm was shown ($R = 0.92$, $p < 5.98e^{-8}$), suggesting that cancer genes exhibit similar distribution as protein coding genes in the human genome. To ascertain the bias of HSF1-CanSig 8q genes quantitatively, a hypergeometric test was used. The probability of observing 27 or more HSF1-CanSig 8q genes (listed in Table 1) in a sample of 100 genes by chance is 1.02×10^{-22}.

Overexpression of HSF1-CanSig 8q genes is primary tumor site-specific

Furthermore, we examined the distribution of HSF1-CanSig 8q genes by individual primary tumor sites. Skewed distribution of HSF1-CanSig genes on chromosome 8q, similar to that in Fig. 1b, was observed in 22 out of 27 primary sites with a variable degree of 8q bias (plots for individual primary site can be found in Additional file 2). The five primary tumor sites that showed no bias are the adrenal gland, kidney, nervous system, thymus, and thyroid. An exponential distribution ($p_i = \lambda e^{-\lambda i}$ or in linear form: $\ln p_i = -\lambda i + \ln \lambda$) was used to quantify the extent of skewness: p_i is the probability or proportion of the ith-ranked chromosome, λ is a coefficient, and i is the rank in the range 1 to 10. A large magnitude of λ indicates a high degree of skewness and vice versa. The λ values of primary sites span a wide spectrum of chromosome 8q bias as shown in Fig. 2, where higher 8q bias are observed in breast and liver cancers than in lymph nodes and soft tissue, for instance. Primary sites lacking 8q bias primary tumor sites tend to be cornered at a region with small λ value (labeled in gray). This result reveals that the overexpression of these HSF1-CanSig 8q genes is primary tumor site-specific.

HSF1-CanSig genes clustered at a 44-Mbp region near the end of chromosome 8q

The 27 HSF1-CanSig 8q genes ranked in the top 100 high-scoring HSF1-CanSig genes are tabulated in Table 1. Remarkably, all HSF1-CanSig 8q genes including HSF1 itself occupy the last three cytogenetic bands of chromosome 8q, i.e., 8q21.3 to 8q.24.3 (~ 44 Mbps). The 3′-

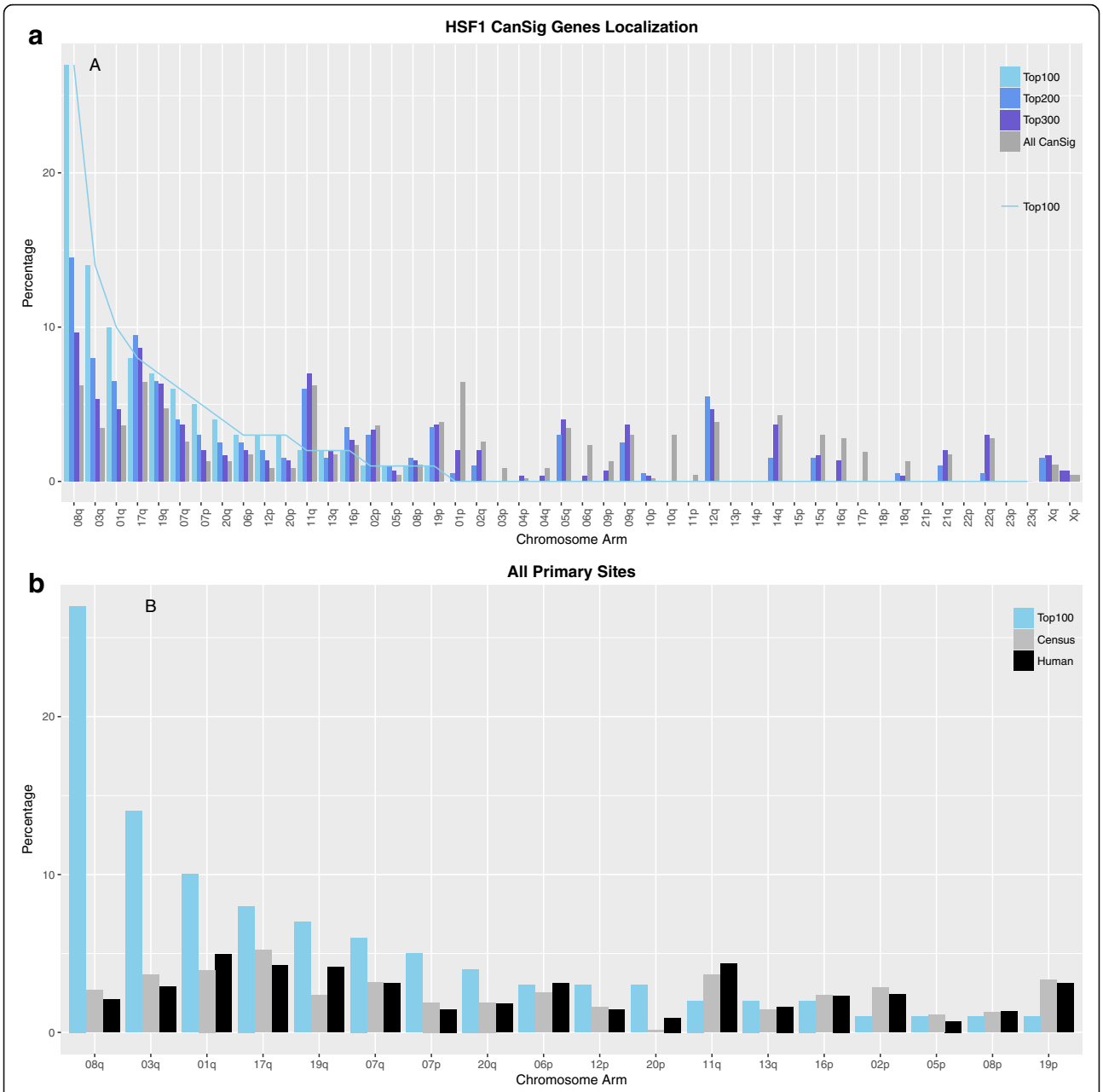

Fig. 1 **a** Percentage distribution of HSF1-CanSig genes by chromosome arms from different tiers: top 100, top 200, top 300, and all HSF1-CanSign genes. **b** Percentage distribution of the top 100 high-scoring HSF1-CanSig genes by chromosome arm. Two references were used for comparison: 633 cancer census genes (labeled census) from COSMIC [49] and 19,073 human protein-coding genes (labeled all) based on HuGO [48]

most gene ZNF250 is ~237 Kbps from the chromosome's 3′ end with only three genes encoded further downstream.

Frequently overexpressed HSF1 associates with the overexpression of HSF1-CanSig genes in 8q

Cancer is heterogeneous both within and between cancer types. Hence, we sought to rank individual HSF1-CanSig 8q genes in different primary tumor sites (Table 2). To cover the entire chromosomal segment encoding all 29 HSF1-CanSig 8q genes, two 8q genes, GPT and KLF10, were included even though they were not ranked in the top 100. The one-sample two-sided Student's *t* test was used to assess the ranking distribution of them (the second column from the right in Table 2) for each primary tumor site where the *p* values of the *t* test are shown in the last column of Table 2. At 95% confidence level, ranks of 8q genes in 20 primary tumor sites (from the ovary to pleura in Table 2) are detected with statistically significant deviation from the

Table 1 Twenty-seven out of the top 100 high-scoring HSF1-CanSig genes are clustered near the end of chromosome 8q

Gene symbol	Name	Location	Rank
PUF60	Poly(U) binding splicing factor 60	08q24.3	1
HSF1	Heat shock transcription factor 1	08q24.3	2
NUDCD1	Nudc domain containing 1	08q23.1	3
MRPL13	Mitochondrial ribosomal protein L13	08q24.12	4
ENY2	ENY2, transcription and export complex 2 subunit	08q23.1	5
CPSF1	Cleavage and polyadenylation specific factor 1	08q24.3	7
SHARPIN	SHANK associated RH domain interactor	08q24.3	8
AZIN1	Antizyme inhibitor 1	08q22.3	9
MAF1	MAF1 homolog, negative regulator of RNA polymerase III	08q24.3	10
PABPC1	Poly(A) binding protein cytoplasmic 1	08q22.3	11
TRAPPC9	Trafficking protein particle complex 9	08q24.3	12
ZNF250	Zinc finger protein 250	08q24.3	13
ZNF34	Zinc finger protein 34	08q24.3	15
NBN	Nibrin	08q21.3	16
JRK	Jrk helix-turn-helix protein	08q24.3	19
DPY19L4	Dpy-19 Like 4 (*C. elegans*)	08q22.1	22
CYHR1	Cysteine and histidine rich 1	08q24.3	23
TPD52	Tumor protein D52	08q21.13	28
PLEC	Plectin	08q24.3	32
MFSD3	Major facilitator superfamily domain containing 3	08q24.3	35
KIFC2	Kinesin family member C2	08q24.3	40
C8ORF37	Chromosome 8 open reading frame 37	08q22.1	49
NDRG1	N-Myc downstream regulated 1	08q24.22	54
SLC45A4	Solute carrier family 45 member 4	08q24.3	56
MROH6	Maestro heat like repeat family member 6	08q24.3	59
LY6K	Lymphocyte antigen 6 family member K	08q24.3	66
LRP12	LDL receptor related protein 12	08q22.3	69

Rank is based on the aggregated overexpression score of all primary tumor sites

mean rank of 233 (see the "Materials and methods" section for the determination of the mean rank). No single primary tumor site had all 29 HSF1-CanSig 8q genes overexpressed in the top 100 HSF1-CanSig genes. Surprisingly there were no primary tumor sites with significant low-average rank (the lowest two are the thymus, $p = 0.896$, and the thyroid, $p = 0.283$, in Table 2), suggesting the absence of a strong suppression among HSF1-CanSig 8q genes in those primary tumor sites.

While no single HSF1-CanSig 8q gene is ranked in the top 100 across all primary sites, HSF1 is frequently overexpressed in most primary tumor sites with kidney as the only exception (26 out of 27, shown at the bottom of HSF1 column in Table 2). HSF1 is also the gene with the highest average rank (33) followed by CPSF1 (52). Additionally, the rank of HSF1 correlates with the average rank of primary sites ($R = 0.71$), but the most correlated 8q gene is CYHR1, not HSF1. Importantly, HSF1 is the only gene in which its rank is always lower than the average rank (Additional file 3: Figure S1).

Lastly, primary tumor sites associated with a small number of high-scoring HSF1-CanSig 8q genes, such as the lymph nodes, soft tissue, thymus, and thyroid, tend to exhibit lower skewness (see Fig. 2). These results suggest a minor role of HSF1-mediated activities in those primary tumor sites.

Primary tumor sites are segregated by ranks

We next asked if the ranks of HSF1-CanSig 8q genes in different primary tumor sites share similar patterns. By using hierarchical clustering with two distinct distance methods, primary sites were clustered using the ranks of HSF1-CanSig 8q genes as shown in Fig. 3. Both panels in Fig. 3 show that primary tumor sites can be reliably split into two groups by the average rank, namely, the top-ranking and the low-ranking groups. The medians

Fig. 2 Quantification of 8q skewness. The percentages of the top 10 chromosomes were fitted to an exponential model. The model outputs (λ = slope, ln(λ) = intercept) are plotted. Primary tumor sites with 8q and without 8q clustering are in blue and gray, respectively. The red triangle represents the fitting of data when 8q genes from all primary sites are ranked together

of the top-ranking groups produced by the two distance methods reflect higher consistency as they vary within a narrow range of 65–68. But the medians of the low-ranking groups show a larger difference from 148 to 198.

Regarding the range of average rank within each group, if the outlier, i.e., soft tissue with average rank 213, is taken out from the top-ranking correlation-distance group in Fig. 3b, the range of average ranks in both becomes identical, 34 to 102. A cutoff value of 102 seems to divide the two groups. Moreover, top-ranking groups created by the two distance methods share 13 out of 14 to 17 primary sites (marked by asterisks in Fig. 3). The low-ranking groups however display a larger variation in rank medians 146–198, but 9 out of 10 to 13 primary sites are clustered together by both methods (marked by plus signs in Fig. 3). This result suggests the feasibility of stratifying primary tumor sites by the over-expression of HSF1-CanSig 8q genes.

HSF1-CanSig 8q genes are highly syntenic in mammals

Since the proximity of HSF1-CanSig 8q genes seems to be a contributing factor for their cohesive expression pattern, we asked whether their proximity reveals biological significance. We sought answers in the light of evolution by examining the synteny of the genomic region spanning the 5′- and 3′-most HSF1-CanSig 8q genes, i.e., TPD53 and ZNF250, respectively. It means that other non-HSF1-CanSig genes were also included in the syntenic analysis. This region consists of 267 genes with known genomic coordinates in human where 29 of them are HSF1-CanSig genes. We aligned these human genes with homologs from bovine, mouse, and rat. Their pairwise percentage of synteny can be found in Table 3A, which varies from the lowest 50% (mouse and bovine) to the highest 84% (human and mouse). As our focus is on the HSF1-CanSig genes, we highlighted their genomic order in Table 3B. Syntenic information of all 267 genes can be found in Additional file 1: Table S5.

Twenty-three out of 29 human HSF1-CanSig 8q genes are syntenic to mouse and rat. Although three non-syntenic genes (Nbn, Dpy19l4, and 2610301B20Rik in mouse and Nbn, Dpy19l4, and MGC94199 in rat) reside in other chromosomes, the order of their locations still agrees with that in humans. If these three genes are included in calculating the percentage of synteny, human-

Table 2 Ranks of HSF1-CanSig 8q genes by primary tumor site. The bracketed number next to the name of a primary site denotes the number of cases recruited in the study. Genes displayed in the table heading are arranged in a syntenic order. The number inside a cell denotes the rank of the gene among HSF1-CanSig genes with the most frequently overexpressed genes ranked 1. Ranks ≤ 100 are highlighted in solid black. Some genes may share the same rank if they have the same overexpression score. Count includes only genes with rank ≤ 100. Average rank is the arithmetic mean of ranks. HSF1-CanSig genes GPT and LRP12 are not part of the top 100 high-scoring genes. *p* value is determined by one-sample two-sided Student's *t* test. The table is sorted by average rank and count in ascending and descending order, respectively

Primary Site	TPD52	NBN	DPY19L4	C8ORF37	PABPC1	KLF10	AZIN1	LRP12	NUDCD1	ENY2	MRPL13	NDRG1	TRAPPC9	SLC45A4	JRK	LY6K	MROH6	PUF60	PLEC	SHARPIN	MAF1	HSF1	CPSF1	CYHR1	KIFC2	GPT	MFSD3	ZNF34	ZNF250	Count	Average Rank	p-value
ovary (558)	30	18	49	85	35	60	11	40	23	25	28	40	6	32	33	136	70	2	9	1	5	3	29	16	50	120	15	4	7	27	34	1.47E-23
eye (80)	29	10	23	27	332	245	7	21	17	2	4	25	16	12	6	42	42	3	23	8	14	21	6	9	16	37	2	11	21	27	36	9.29E-15
prostate (498)	2	11	28	45	1	120	5	30	5	3	8	154	16	87	22	135	59	11	28	15	9	6	18	28	13	135	35	30	7	25	37	4.70E-20
breast (1100)	25	34	23	43	9	187	7	58	2	3	1	46	12	92	35	27	64	4	56	22	17	5	29	104	40	163	59	13	10	26	41	1.44E-19
colorectal (382)	55	35	8	62	13	34	7	60	12	17	30	104	22	113	41	177	60	9	132	18	12	20	23	24	120	342	42	19	22	23	56	7.09E-14
skin (472)	129	63	32	93	25	250	32	164	65	17	16	129	27	38	15	173	74	5	108	53	39	15	6	3	53	59	42	11	38	23	61	2.26E-15
bladder (408)	41	16	16	29	3	125	5	125	11	4	9	83	32	117	67	37	63	6	146	44	12	1	27	83	202	188	117	146	21	21	61	2.92E-15
liver (373)	29	19	25	113	2	376	14	113	3	6	1	74	13	72	20	415	21	4	30	5	7	11	9	17	33	356	23	15	10	24	63	1.08E-08
lung (517)	26	29	32	59	7	257	19	111	6	13	48	43	20	304	39	39	59	1	155	22	5	4	11	18	39	461	48	9	15	24	65	2.23E-09
head and neck (522)	209	45	70	70	21	26	31	75	16	21	51	16	84	280	112	23	152	3	23	14	19	2	10	33	28	447	84	29	53	24	71	7.05E-10
esophagus (185)	50	69	32	91	75	270	26	188	14	31	9	38	34	75	50	20	170	4	22	19	15	2	26	19	57	299	188	69	99	24	71	7.89E-12
pancreas (179)	39	97	31	153	21	383	4	97	2	3	1	465	50	12	12	60	39	21	9	18	50	24	9	78	60	153	18	153	19	24	72	9.73E-09
stomach (415)	13	25	22	97	12	450	4	161	9	10	7	31	33	180	92	128	55	3	142	45	59	8	14	17	59	148	31	274	110	21	77	2.62E-09
uterus (177)	83	223	16	104	12	386	27	262	4	23	56	27	71	18	35	188	43	56	35	27	7	7	10	47	65	310	65	23	16	23	77	3.21E-09
cervix (306)	167	53	67	167	16	149	34	118	34	41	19	96	149	135	96	41	268	18	268	36	21	20	28	46	53	393	82	62	72	20	95	5.10E-09
testis (156)	51	5	152	17	8	101	16	127	183	183	51	86	26	101	21	338	40	51	9	56	152	35	14	295	56	127	463	6	127	17	100	4.14E-07
brain (530)	241	22	114	262	15	319	319	202	28	13	23	163	68	114	18	35	449	22	42	42	10	8	13	35	96	37	163	28	30	19	101	1.46E-06
blood (179)	20	33	177	25	25	128	15	100	37	249	33	100	9	33	2	462	3	79	47	100	79	79	177	47	289	464	58	79	9	22	102	3.56E-06
bile duct (36)	9	14	41	74	414	144	414	414	74	144	41	414	6	74	9	22	74	74	269	74	144	41	41	41	144	269	269	41	22	18	131	4.44E-04
pleura (87)	230	230	230	300	58	386	91	230	58	18	91	230	58	58	143	300	386	143	143	91	58	5	91	143	18	29	91	91	230	15	146	1.71E-04
adrenal gland (79)	138	82	69	82	232	329	393	138	23	82	113	232	138	329	100	466	466	100	82	69	159	46	35	192	232	138	393	439	82	12	185	7.45E-02
nervous system (528)	245	145	272	272	145	190	272	26	166	99	190	409	125	215	112	442	332	99	215	59	59	22	59	442	426	190	190	39	22	9	189	7.60E-02
lymph nodes (48)	106	56	358	56	106	211	358	358	211	106	106	358	106	211	11	56	358	358	358	358	211	56	26	211	458	211	211	358	56	7	207	3.06E-01
soft tissue (263)	334	38	110	285	72	257	68	215	4	22	2	215	285	309	196	426	387	72	110	285	232	93	82	410	447	460	447	126	196	9	213	4.75E-01
kidney (534)	70	196	437	344	63	131	80	175	217	30	429	241	344	450	241	450	48	70	149	48	58	196	63	407	217	437	283	283	379	9	225	7.80E-01
thymus (120)	29	90	133	350	213	350	213	281	133	406	281	213	90	281	350	406	52	90	350	52	281	52	213	281	281	281	461	350	281	7	236	8.96E-01
thyroid (509)	443	452	421	207	352	278	421	124	400	57	180	352	96	452	400	421	57	278	96	36	43	96	324	443	207	207	180	352	236	7	262	2.83E-01
Count	17	22	17	16	20	3	20	9	21	22	21	13	21	12	20	11	18	24	14	25	21	26	24	17	16	4	15	18	20	17\19		
Average Rank	105	78	111	130	85	227	107	149	65	60	68	162	72	155	84	202	144	59	113	60	66	33	52	129	139	239	150	113	81		112	

rat and human-mouse share 72 and 84% of synteny, respectively. In comparison between bovine and human, all 8q homologs in bovine are found in a single chromosome, i.e., chromosome 14, but in the complementary strand of the bovine's reference genome. Twenty-five of the 29 8q genes are syntenic. If the entire 8q region is considered, bovine's genes share only 57% of synteny with human versus 61–71% in rat and mouse, indicating bovine genome has accumulated a larger scale of genome rearrangement than human, mouse, and rat since divergence. Regardless, the gene order is largely preserved between human and bovine. This finding is congruent with the estimated divergence time among these four mammals [9]. According to the report, bovine and human and mouse and human diverge in about 71–113 million years ago (MYA) and 62–101 MYA, respectively. Furthermore, our findings concur with the synteny analyses from the genome sequencing project of human (Figure 46 of [10]), Brown Norway rat (Figure 4 of [11]), and Taurine cattle (Figure S12 of [12]). To further investigate the synteny of human 8q in mouse, rat, and bovine, the tool Cinteny [13] was used to visualize the extent of synteny and they can be found in Additional file 3: Figure S2.

There are two possible explanations to account for the observed syntenic conservation among these genes: (i) the divergence time between taxa is insufficient to observe significant rearrangement or (ii) the proximity of genes is essential for biological functions. Classic examples are the Hox and globin gene clusters. As the estimation from [9] is adequate to disregard the former, we were intrigued to explore the latter, i.e., the biological function shared among these genes.

Synteny does not explain coexpression

To determine whether or not the overexpression of HSF1-CanSig 8q genes in cancers is related to synteny, we analyzed the expression pattern of genes flanking HSF1 (genomic coordinates of HSF1's neighboring genes can be found in Additional file 1: Table S7) in each cancer case to shed light on the clustering of HSF1-CanSig 8q genes. We reasoned that if the expression level of HSF1's neighboring genes correlates with their distance from HSF1, synteny may be a factor accounting for the clustering of CanSig 8q genes; otherwise reasons other than synteny may shape the clustering.

We selected 1682 cancer cases with HSF1 overexpressed two or more standard deviations above the reference. As shown in Fig. 4a, expression of HSF1 and its neighboring genes do not correlate with the distance between them. Genes close to HSF1 such as MROH1, BOP1, and DGAT1 do not show higher correlation (in terms of overexpression with HSF1) than HSF1-CanSig 8q genes that are farther from them. For example,

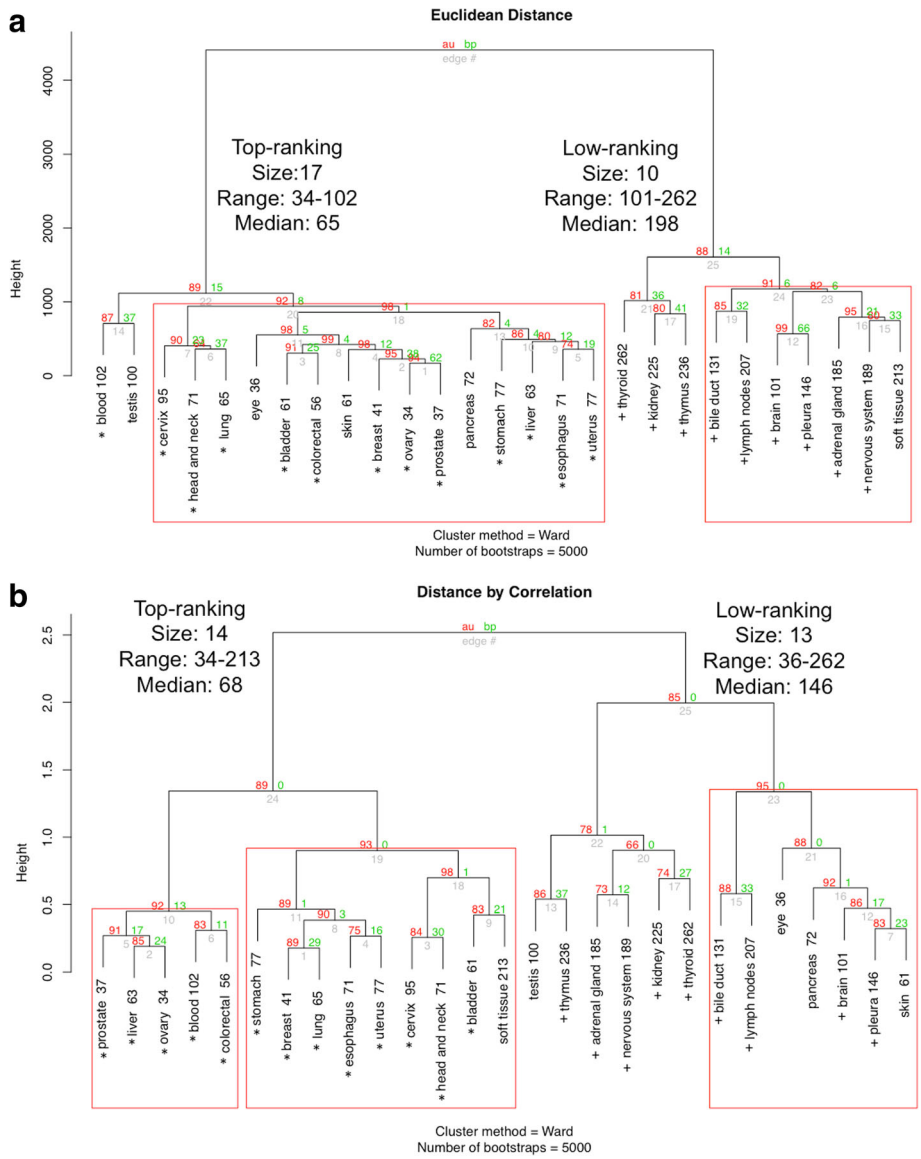

Fig. 3 Rank analysis by hierarchical clustering with bootstrapping. The number beside each primary tumor site is the average rank. Subtrees with high bootstrap value (≥ 90) are highlighted in red boxes. Each dendrogram was built after bootstrapping 5000 times. The red and green numbers at the branch point represent approximately unbias (AU) *p* value and bootstrap probability (BP), respectively. AU is recommended as a better parameter than BP to assess the clustering reliability at each branch point [54]. Cluster subtree highlighted by red box indicates AU ≥ 90. For the top-ranking groups, primary tumor sites shared between two distance methods are marked by asterisks. Similarly, for the low-ranking groups, primary sites are marked by plus signs. **a** Cluster dendrogram based on Euclidean distance. **b** Same input data as panel **a** but correlation was used as the distance calculation method

although 8q genes SHARPIN, MAF1, and CPSF1 are far apart from HSF1 than MROH1, BOP1, and DGAT1, they exhibit higher correlation with HSF1's expression. Similar variability is also observed when correlations are examined by individual primary tumor site (Additional file 4). Thus, our data suggests that synteny is not a contributing factor for the co-overexpression of HSF1-CanSig genes clustered in chromosome 8q21-24. It is however intriguing to see that several overexpressed non-HSF1-CanSig genes are correlated with HSF1 such as MAF1 ($R = 0.7$), CYC1 ($R = 0.6$), and SLC52A2 ($R = 0.6$) (see Fig. 4b).

Network analysis links HSF1-mediated responses to pre-mRNA 3′ processing

Next, we took a systems approach to determine if HSF1-CanSig 8q genes work cooperatively in biological networks. WebGestalt gene network (or pathway) analysis was performed to determine whether shared biological processes are common to these genes. The basic

Table 3 Synteny analysis of HSF1-CanSig 8q genes. A. Percentage of syntenic genes between species. Percentages are based on the number of genes in the column's species. The percentages in lower triangle excluded the separated chr4q in mouse and chr5q in rat. Whereas, the percentages in the upper triangle (shaded) include the separated chr4q in mouse and chr5q in rat. B. Order of human genes is used as the reference. Genes are ordered in 5′-to-3′ direction according to respective reference genomes from top to bottom except for bovine, which is in reverse order. Non-syntenic genes are marked in boxes. Superscript + indicates the chromosomal location of the gene was determined by BLAT. Exact genomic coordinates of these genes can be found in Additional file 1: Table S6

A	Human	Mouse	Rat	Bovine
Human		84	72	57
Mouse	71		82	59
Rat	61	69		65
Bovine	57	50	55	

B			
Bovine (bosTau8) Chr14	Human (hg38) Chr8q	Mouse (mm10) Chr15q	Rat (rn6) Chr7q
MRPL13			
	TPD52	*Chr4q#*	*Chr5q#*
NBN	NBN	*Nbn#*	*Nbn#*
DPY19L4+	DPY19L4	*Dpy19l4#*	*Dpy19l4#*
C14H8orf37+	C8orf37	*2610301B20Rik#*	*MGC94199#*
PABPC1	PABPC1	Pabpc1	Pabpc1
KLF10	KLF10	Klf10	Klf10
AZIN1	AZIN1	Azin1	Azin1
LRP12	LRP12	Lrp12	Lrp12
NUDCD1	NUDCD1	Nudcd1	Nudcd1
ENY2	ENY2	Eny2	Eny2
	MRPL13	Mrpl13	Mrpl13
TPD52 SLC45A4+			
NDRG1	NDRG1	Ndrg1	Ndrg1
TRAPPC9+	TRAPPC9	Trappc9	Trappc9
	SLC45A4	Slc45a4	Slc45a4+
JRK+	JRK	Jrk	Jrk
N.A.	LY6K	N.A.	N.A.
MROH6+	MROH6	Mroh6	Mroh6+
PUF60	PUF60	Puf60	Puf60
PLEC+	PLEC	Plec	Plec
SHARPIN	SHARPIN	Sharpin	Sharpin
MAF1	MAF1	Maf1	Maf1+
HSF1	HSF1	Hsf1	Hsf1
CPSF1	CPSF1	Cpsf1	Cpsf1
CYHR1	CYHR1	Cyhr1	Cyhr1
KIFC2	KIFC2	Kifc2	Kifc2
GPT	GPT	Gpt	Gpt
MFSD3+	MFSD3	Mfsd3	Mfsd3
ZNF34	ZNF34	N.A.	N.A.
ZNF250+	ZNF250	Zfp647+	Zfp647+

mechanism of WebGestalt is to detect the enrichment of Gene Ontology (GO) terms associated with a set of genes with statistical support. We used 29 HSF1-CanSig 8q genes to query 33 TCGA RNA-Seq studies in various primary tumor sites (the list of studies queried can be found in Additional file 1: Table S8). A sample of result generated by WebGestalt can be found in Additional file 3.

Out of the 33 TCGA datasets, eight studies were found to be enriched by HSF1-CanSig 8q genes: adrenocortical carcinoma, colon adenocarcinoma, esophageal carcinoma, kidney chromophobe, hepatocellular carcinoma,

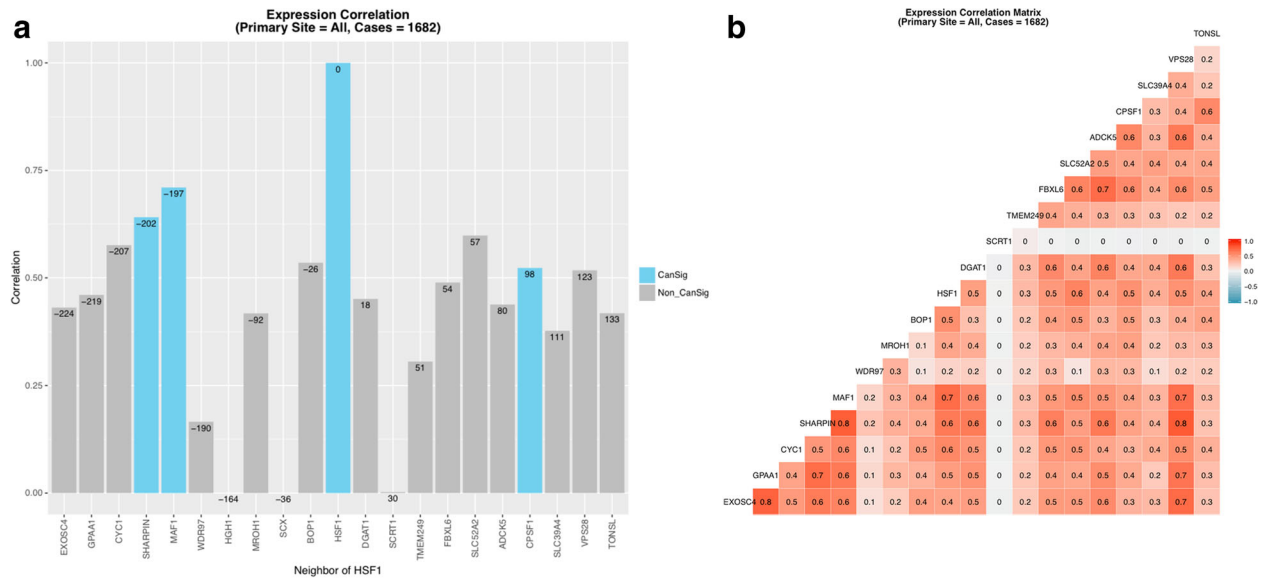

Fig. 4 Expression correlation between genes in the neighborhood of HSF1. **a** HSF1 is set as the focal gene. No expression data can be obtained for HGH1 and SCX in cBioPortal; therefore, their correlations were set to zero. The numbers on the top of the bars represent the distance of the gene from HSF1 in Kbps. Negative signs indicate genes are located upstream of HSF1. **b** Correlation matrix of all 21 genes. Blank boxes represent insufficient number of cases to determine the correlation

lung adenocarcinoma, ovarian serous cystadenocarcinoma, and stomach adenocarcinoma. Intriguingly, all results pointed to the enrichment of RNA polyadenylation (GO:0043631, p values 2.40e-3 to 4.34e-2), mRNA polyadenylation (GO:0006378, p values 2.40e-3 to 4.34e-2), and 3′ end processing (GO:0031124, p values 3.03–4.30e-2). With no exception, the CanSig 8q genes that contributed to these hits were HSF1, PABPC1, and CPSF1.

To corroborate WebGestalt's results, we repeated the hierarchical clustering analysis of rank data (Table 2) but using only HSF1 and three pre-mRNA 3′ processing genes: PABPC1, CPSF1, and PUF60. (Note that the clustering method requires a minimum of four genes; therefore, PUF60, a RNA processing factor, was included here even though WebGestalt did not highlight it.) If these four genes dominate the average rank of HSF1-CanSig 8q genes in different primary tumor sites, the clustering results produced by them should highly resemble to the results in Fig. 3. Figure 5a, b shows two cluster dendrograms produced by the two distance methods which share high similarity with Fig. 3a, b, respectively. When the two cluster dendrograms generated by the Euclidean distance (Figs. 3a and 5a) are compared, the top-ranking and low-ranking groups are found to share 15 and 9 common primary sites, respectively, which means 24 out of 27 primary sites. Similarly, the two dendrograms using correlation distance method (Figs. 3b and 5b) share 12 and 9 common primary sites, respectively. It means that these four genes are representative for all HSF1-CanSig 8q genes, indicating influential role played by them. It also suggests that HSF1 mediates pre-mRNA 3′ processing in cancer development.

Discussion

We have undertaken a cancer bioinformatics approach to study the clustering of HSF1-CanSig genes in chromosome 8q. Based on the TCGA expression data obtained from cancer cases, we found that HSF1-CanSig genes are clustered at chromosome 8q substantiated by compelling statistical evidence under two null models: all protein-coding genes and all HSF1-CanSig genes. λ value analysis further reveals variation of 8q prevalence in 27 primary tumor sites (Fig. 2). This result may support the use of λ to characterize the strength of HSF1 cancer program in various cancers: bone [14], breast [1], colon [1], esophagus [3], head and neck [15], liver [16], lung [1], ovarian [5], pancreas [17], prostate [18], uterine [4], and skin cancer [19]. Crucially, this perspective has been confirmed by the hierarchical clustering analysis as all of the above cancers, except bone, are members of the top-ranking group (Figs. 3 and 5).

The rank analysis discovers the association between HSF1 and HSF1-CanSig genes encoded in chromosome 8q in 27 primary tumor sites. HSF1's rank correlates with and always falls below the average rank across primary sites (Additional file 3: Figure S1). The ranks of 20 out of 27 primary tumor sites are asymmetrically skewed to high ranks with statistical significance, suggesting the overexpression of HSF1-CanSig 8q genes is coordinated

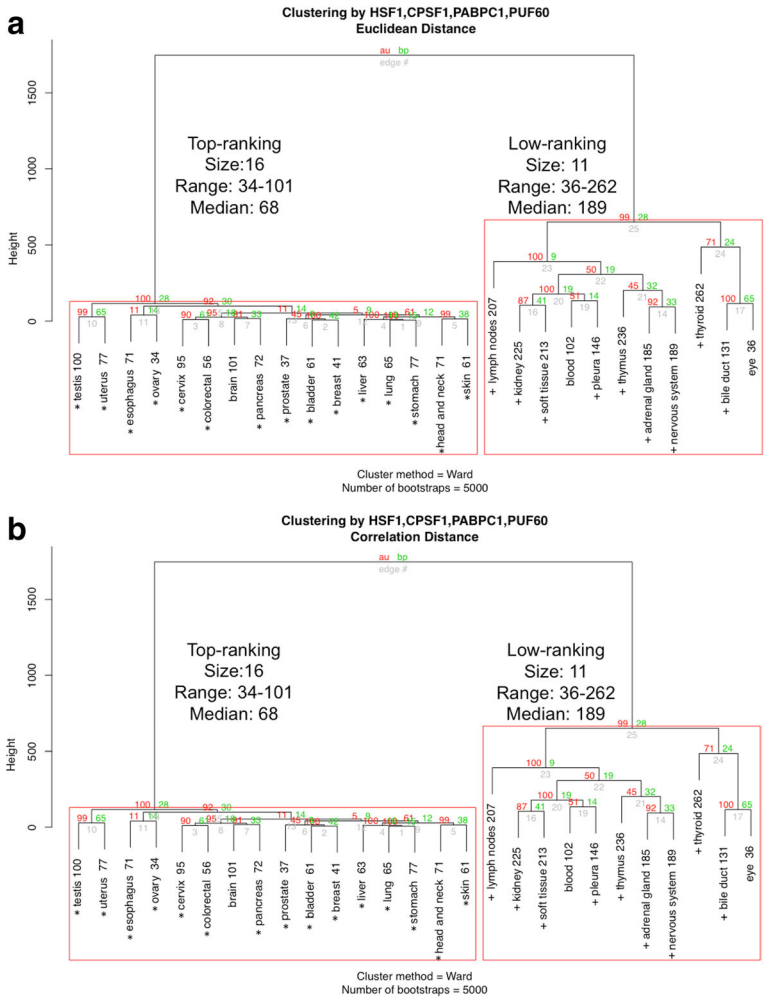

Fig. 5 a Hierarchical clustering of rank data (Table 2) by HSF1 and three pre-mRNA 3′ processing factors: PABPC1, CPSF1, and PUF60. Distance method is based on the Euclidean distance. In the top-ranking group, the asterisk denotes the primary tumor site shared with Fig. 3a. In the low-ranking group, the plus sign indicates the primary site shared with Fig. 3a. **b** Clustering by correlation distance

in those tumor sites. In contrast, when the rank of HSF1 is greater than the cutoff rank ~ 102, more diverse overexpression patterns are observed among HSF1-CanSig 8q genes, implying the diminishing control of HSF1 in those cancers. These observations engender HSF1 to be an important target for the development of cancer treatment.

Clustered HSF1-CanSig 8q genes co-express at a high level in several primary tumor sites. Is the clustering of HSF1-CanSig 8q genes a contributing factor for co-expression biologically? As "Nothing in biology makes sense except in the light of evolution" [20], we sought answers from synteny. Our results (Table 3B) indicate the majority (22–23) of HSF1-CanSig 8q genes are syntenic among human, mouse, rat, and bovine. Despite the neighboring gene model suggesting an evolutionary explanation of this co-expression [21, 22], our results do not align with such a model. When the expressions of genes flanking HSF1 were examined, our results demonstrate incoherent overexpression levels attributed to distance. Thus, other forces such as the transcriptional activation by HSF1 or its downstream products may be the cause of the co-expression of HSF1-CanSig 8q genes in tumors. It still remains as a question whether this is common for other cancer-related transcription factors. Related to this analysis is the finding that a few non-HSF1-CanSig genes encoded in chromosome 8q whose expressions correlate with HSF1: GPAA1 ($R \sim 0.8$) and EXOSC4 ($R \sim 0.75$) in the ovary and FBXL6 ($R \sim 0.8$) in the prostate (Additional file 4). These findings concur with previous studies that genes in regions 8q21, 8q22, and 8q24 are notable for cancers [23–25].

Hierarchical clustering and network analysis have fostered the connection between HSF1 and pre-mRNA 3′ processing. Three essential pre-mRNA processing factors are discovered in relation to the overexpression of

HSF1-CanSig genes through network analysis: PABPC1, CPSF1, and PUF60. CPSF1 is an essential polyadenylation factor. To date, no biochemical evidence supports any association between HSF1 and CPSF1. However, the interaction between the two processes, heat shock and polyadenylation, has been confirmed by two previous studies [26, 27], suggesting intermediaries may be involved in the crosstalk between these two processes. The poly(A) tail of a mRNA serves as a checkpoint for mRNA nuclear export. Intriguingly, HSF1-TRP interaction had been reported to facilitate the export of HSP70 mRNA under stress [28]. HSF1 had also been shown to form complexes with two core polyadenylation factors: symplekin and CstF64 [29, 30]. Alternative polyadenylation (APA) is ubiquitous and tissue-specific in mammals [31–33]. Importantly, it has been confirmed to play an essential role in many tumor types [34–43]. In rare cases, APA interferes with the protein encoded by the gene. Two classic, non-cancer examples are calcitonin-related polypeptide-alpha gene (CALCA) [44] and immunoglobulin M (IgM) [45]. But more often, APA alters the 3′ untranslated regions (UTRs) by either shortening or lengthening the constitutive 3′ UTRs. 3′ UTR is known to harbor binding sites of regulatory proteins, and miRNAs, in addition to secondary structures [46]. Their binding can either attenuate or bolster mRNA stability, affecting downstream protein synthesis and subsequently the proteome of cells.

It is noteworthy that two other HSF1-CanSig 8q genes are associated with pre-mRNA 3′-end processing: PUF60 and PABPC1. PUF60 is a member of the highly conserved nucleic acid-binding protein family: pumilio and FBF homology protein (PUF). PUF60 involves in pre-mRNA splicing and transcriptional regulation. Notably, PABPC1 binds to poly(A) tail of eukaryotic mRNA, promoting mRNA translatability. A proteomic study had identified the presence of PUF60 and PABPC1 in the polyadenylation complex [47], ascertaining the linkage between heat shock and polyadenylation.

Conclusion

The role of HSF1 played in cancer proliferation and malignancy is critical and well established. The cancer bioinformatics approach taken by us provides new information about the overexpression of HSF1-CanSig 8q genes mediated by HSF1 in different tumor types, illuminating the connection between malignancy progression driven by HSF1 and pre-mRNA 3′ processing. However, the true underlying biological mechanisms that drives HSF1 relationship with chromosome 8q is multifaceted and largely unknown but may hold the key to understanding tumor development in certain tissues and organs. As the activation of the master regulator HSF1 varies among different primary tumor sites, HSF1-CanSig 8q genes may be developed as prognosis biomarkers for improving clinical outcomes.

Materials and methods

HSF1-CanSig genes and cancer expression data

The list of 456 HSF1 CanSig genes was downloaded from Table S5 of [1]. To facilitate searching by gene symbols, they were adhered to the standard HUGO Gene Nomenclature [48]. As a result, 87 non-HuGO gene symbols from [1] were replaced by the official HuGO gene symbols. Six of them were converted to multiple gene symbols, expanding the original list to 475 HSF1-CanSig genes. The complete list of HSF1-CanSig genes and their full names and chromosome locations can be found in Additional file 1: Table S2.

mRNA expression (in fold change) of the HSF1-CanSig genes in different cancer primary sites were retrieved from cBioPortal via its Web API [7]. At the time of writing, cBioPortal contains 159 cancer studies conducted in 29 primary tumor sites (Additional file 1: Table S3). Our goal is to include one RNA-Seq study per primary site. The following criteria were used to select expression data for our analysis:

1. The study is focused on a specific primary tumor.
2. Transcriptome data is generated by the study.
3. Only one study is selected per primary tumor site. If multiple transcriptome datasets are found for a primary site, the study recruited the largest number of cancer cases is chosen.
4. Only expressions of HSF1-CanSig genes are considered.

Transcriptome data were found in the majority, but not all, of cancer studies. In this report, 27 studies were selected according to the above criteria and they are tabulated in Additional file 1: Table S9.

Overexpression score of HSF1-CanSig genes

A cancer study consists of a group of cancer cases or patients, and each cancer case is associated with a set of mRNA expression z-scores of genes. But in this report, our focus is solely on the 475 HSF1-CanSig genes mentioned above. cBioPortal pre-computes a z-score for each gene, representing the number of standard deviations of its expression deviates from the reference gene population [8]. For example, a z-score value 2.8318 for gene ANAT in the case TCGA-OR-A5J1-01 from the Adrenocortical Carcinoma study indicates that the gene ANAT is overexpressed by 2.8318 standard deviations above the reference.

To quantify the prevalence of overexpression of HSF1-CanSig genes among cancer cases, we devised the

following scoring scheme to assess the magnitude of a gene's overexpression as:

$$Overexpression\ Score = Number\ of\ overexpressed\ cases / Total\ number\ of\ cases$$

A gene is considered overexpressed if its cBioPortal's expression z-score is ≥ 2, the same threshold used in cBioPortal website. For instance, the Adrenocortical Carcinoma study consists of 79 cancer cases (see Table 3), UBE2B gene was detected to overexpress in 33 cases. The overexpression score of UBE2B is 33/79 = .4177. The full list of genes' overexpression scores by primary site can be found in Additional file 1: Table S4.

Chromosome localization of top 100 HSF1-CanSig genes

When expression data of different primary tumor sites was combined for analysis, the overexpression score of an HSF1-CanSig gene was calculated by adding the overexpression scores of the gene in all primary sites. The top 100 highest scoring genes were used for the chromosome localization analysis. The complete list of the top 100 HSF1-CanSig genes can be found in Additional file 1: Table S5. Two sets of genes were chosen as the reference or null model: all human protein-coding genes based on HuGO [48] and genes implicated in cancer when mutated according to the COSMIC Cancer Gene Census [49]. In total, 19,073 and 633 of protein-coding genes and cancer census genes were used as references, respectively.

A hypergeometric model was used to assess the probability of localization. $P(i \geq k) = 1 - \Sigma_{i=0}^{i=k-1} \binom{K}{i} \left(\frac{N-K}{n-i}\right) / \binom{N}{n}$, where K is the number of genes encoded in the chromosome 8q, k is the number of 8q genes among the top 100 high-scoring HSF1-CanSig genes, n is the sample size, and N is the total number of protein-coding genes. For $K = 401$, $k = 27$, $n = 100$, and $N = 19{,}073$, the probability of observing 27 or more HSF1-CanSig 8q genes in a random sample of 100 genes by chance is 1.02×10^{-22}.

There are 29 HSF1-CanSig genes in 8q. The probability of finding 27 out of 29 HSF1-CanSig 8q genes in a sample of 100 HSF1-CanSig genes can be determined by hypergeometric model. Using the same formula above: K (= 29) is the number of HSF1-CanSig genes encoded in 8q, k (= 27) is the number of 8q genes in the top 100, n (= 100) is the sample size, and N (= 475) is the total number of HSF1-CanSig genes. The probability is 6.85×10^{-18}.

λ value analysis was performed by fitting of the linear form of the exponential distribution, i.e., $\ln p_i = -\lambda i + \ln \lambda$, to the rank data in Additional file 1: Table S5. Only the percentages of the top 10 high-ranking chromosome arms were used for fitting. Fitting was done by lm() function in *R*.

Synteny analysis

The genomic segment encoding human HSF1-CanSig 8q genes were analyzed for syntenic conservation. This segment spans from 80,034,869 (TPD52) to 144,901,461 (ZNF250) in chromosome 8 of human. By using the homology information from NCBI HomoloGene database (build 68) [50], 267, 165, 237, and 199 homologous genes were identified with genomic coordinates in human (hg38), bovine (bosTau8), mouse (mm10), and rat (rn6), respectively. Their genomic locations were retrieved automatically from the UCSC Genome Browser [51] through the Python package CruzDb [52]. For genomic locations that cannot be obtained by this approach, they were either determined manually by aligning the cDNA sequences on the genome by BLAT [53] or discarded. Genomic coordinates of the genes can be found in Additional file 1: Table S6.

Coexpression analysis

Ten genes immediately flanking the upstream and downstream regions of HSF1 were identified in the human genome (hg38) using the UCSC Genome Browser [51]. This region spans about 365 kbps. The ten immediate upstream genes of HSF1 are EXOSC4, GPAA1, CYC1, SHARPIN*, MAF1*, WDR97, HGH1, MROH1, BOP1, and SCX. The ten downstream genes are DGAT1, SCRT1, TMEM249, FBXL6, SLC52A2, ADCK5, CPSF1*, SLC39A4, VPS28, and TONSL (HSF1-CanSig genes are suffixed by an asterisk). Genomic coordinates of them can be found in Additional file 1: Table S7. Expressions (in fold changes) of these 21 genes, including HSF1, were retrieved from cBioPortal if HSF1's expression of a cancer case is ≥ 2 standard deviations higher than the reference. Correlation of expression fold change was computed between each of the 20 genes and HSF1 using the function cor.test in *R* where the method was Pearson.

Student's *t* test and hierarchical clustering

Our goal is to assess the statistical significance of the average rank of HSF1-CanSig 8q genes per primary site (listed in the second column from the right in Table 2). By assuming ranks are independent and uniformly distributed, the average rank of a sample consisting of 29 ranks randomly drawn from 1 to 466 without replacement does distribute normally with mean 233. (Note that the largest rank is 466 instead of 475 because of ties.) Thus, t test is suitable for this analysis. *R*'s t.test() function was used with mean 233 (null hypothesis). The mean was 233 as the range of rank is 1 to 466. p values

of the t test can be found in the rightmost column in Table 2.

Hierarchical clustering was performed using *R*'s package pvclust [54]. One distinct feature offered by pvclust is the assessment of cluster subtrees by resampling bootstrap. The input to pvclust was the gene ranks in Table 2. The cluster method and the number of bootstrapping were "Ward" and 5000, respectively. Two distance calculation methods were used for cross-validation purpose: Euclidean distance and correlation.

WebGestalt analysis

We used WebGestalt to uncover possible gene networks commonly perturbed by HSF1-CanSig 8q genes [55]. At WebGestalt website [56], we selected "hsapiens" as the organism, "Network Topology-based Analysis (NTA)" as the method of interest, "network" as the functional database, and "genesymbol" as gene ID type. Based on these parameters, WebGestalt automatically shortlisted 33 cancer studies from The Cancer Genome Atlas (TCGA) that were available for searching. We used 27 HSF1-CanSig 8q genes, including HSF1, identified from the top 100 most frequently overexpressed genes to query WebGestalt while values of other parameters remained in default settings. We repeated the search for all 33 TCGA cancer studies (Additional file 1: Table S8).

Additional files

Additional file 1: Table S1. Statistical tests of observing HSF1-CanSig genes by chromosome arm. **Table S2.** List of 475 HSF1-CanSig genes. **Table S3.** List of 29 primary tumor sites. **Table S4.** Overexpression scores of HSF1-CanSig genes in each primary site. **Table S5.** Top 100 most frequently overexpressed HSF1-CanSig genes. **Table S6.** Genomic coordinates of syntenic genes. **Table S7.** Genomic coordinates of HSF1's flanking genes. **Table S8.** List of TCGA studies queried in WebGestalt website. (XLSX 643 kb)

Additional file 2: Percentage distribution of HSF1-CanSig genes by chromosome arm for each primary tumor site. For each primary site, two plots are included, one using all protein-coding genes and cancer census genes as references. The other uses HSF1-CanSig genes as the reference. (PDF 186 kb)

Additional file 3: Rank of HSF1 among primary sites. Syntenty visualization by Cinteny. An example showing the output web page of WebGestalt (PDF 738 kb)

Additional file 4: Expression correlation analysis of syntenic genes for each primary tumor sites. (PDF 240 kb)

Abbreviations

GO: Gene Ontology; HSF1-CanSig Gene: Heat Shock Factor 1 Cancer Signature Gene; HuGO: Human Genome Organization; MYA: Million years ago; TCGA: The Cancer Genome Atlas; UTR: Untranslated region

Acknowledgements

The authors thank the patient participants in TCGA and the cBioPortal support team at the Memorial Sloan Kettering Cancer Center.

Funding

ESH is kindly supposed by the Biology Department at Lafayette College. CQZ is supported by the Weis Center for Research. TLP and HW are supported by Mowad Endowment for New Discoveries, Department of Urology, Geisinger Clinic.

Availability of data and materials

The data used in this report can be found in Additional file 1.

Authors' contributions

All authors conceived the study. CQZ, TLP, and ESH conducted the data analysis. ESH performed the statistical analysis and programming. CQZ, HW, TLP, and ESH wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Biology, Lafayette College, Easton, PA 18042, USA. [2]Department of Computer Science, Lafayette College, Easton, PA 18042, USA. [3]Urology Department, Geisinger Medical Center, Danville, PA 17822, USA. [4]Weis Research Center, Geisinger Medical Center, Danville, PA 17822, USA.

Received: 11 August 2017 Accepted: 11 December 2017
Published online: 21 December 2017

References

1. Mendillo ML, Santagata S, Koeva M, Bell GW, Hu R, Tamimi RM, Fraenkel E, Ince TA, Whitesell L, Lindquist S. HSF1 drives a transcriptional program distinct from heat shock to support highly malignant human cancers. Cell. 2012;150(3):549–62.
2. Scherz-Shouval R, Santagata S, Mendillo ML, Sholl LM, Ben-Aharon I, Beck AH, Dias-Santagata D, Koeva M, Stemmer SM, Whitesell L, et al. The reprogramming of tumor stroma by HSF1 is a potent enabler of malignancy. Cell. 2014;158(3):564–78.
3. Liao Y, Xue Y, Zhang L, Feng X, Liu W, Zhang G. Higher heat shock factor 1 expression in tumor stroma predicts poor prognosis in esophageal squamous cell carcinoma patients. J Transl Med. 2015;13:338.
4. Engerud H, Tangen IL, Berg A, Kusonmano K, Halle MK, Oyan AM, Kalland KH, Stefansson I, Trovik J, Salvesen HB, et al. High level of HSF1 associates with aggressive endometrial carcinoma and suggests potential for HSP90 inhibitors. Brit J Cancer. 2014;111(1):78–84.
5. Powell CD, Paullin TR, Aoisa C, Menzie CJ, Ubaldini A, Westerheide SD. The heat shock transcription factor HSF1 induces ovarian cancer epithelial-mesenchymal transition in a 3D spheroid growth model. PLoS One. 2016;11(12)
6. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. EPIGENETICS AND GENETICS: a census of amplified and overexpressed human cancer genes. Nat Rev Cancer. 2010;10(1):59–64.
7. cBioPortal Web API. http://www.cbioportal.org/web_api.jsp. Accessed 1 May 2017.
8. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.
9. Donoghue PC, Benton MJ. Rocks and clocks: calibrating the tree of life using fossils and molecules. Trends Ecol Evol. 2007;22(8):424–31.
10. Lander ES, Consortium IHGS, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
11. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004;428(6982):493–521.

12. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Abatepaulo ARR, Abbey CA, Adelson DL, Aerts J, Ahola V, Alexander L, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 2009; 324(5926):522–8.
13. Sinha AU, Meller J. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. BMC Bioinf. 2007;8
14. Zhou Z, Li Y, Jia Q, Wang Z, Wang X, Hu J, Xiao J. Heat shock transcription factor 1 promotes the proliferation, migration and invasion of osteosarcoma cells. Cell Prolif. 2017;
15. Ishiwata J, Kasamatsu A, Sakuma K, Iyoda M, Yamatoji M, Usukura K, Ishige S, Shimizu T, Yamano Y, Ogawara K, et al. State of heat shock factor 1 expression as a putative diagnostic marker for oral squamous cell carcinoma. Int J Oncol. 2012;40(1):47–52.
16. Fang F, Chang RM, Yang LY. Heat shock factor 1 promotes invasion and metastasis of hepatocellular carcinoma in vitro and in vivo. Cancer-Am Cancer Soc. 2012;118(7):1782–94.
17. Dudeja V, Chugh RK, Sangwan V, Skube SJ, Mujumdar NR, Antonoff MB, Dawra RK, Vickers SM, Saluja AK. Prosurvival role of heat shock factor 1 in the pathogenesis of pancreatobiliary tumors. Am J Physiol-Gastr L. 2011; 300(6):G948–55.
18. El Gammal AT, Bruchmann M, Zustin J, Isbarn H, Hellwinkel OJC, Kollermann J, Sauter G, Simon R, Wilczak W, Schwarz J, et al. Chromosome 8p deletions and 8q gains are associated with tumor progression and poor prognosis in prostate cancer. Clin Cancer Res. 2010;16(1):56–64.
19. Toma-Jonik A, Widlak W, Korfanty J, Cichon T, Smolarczyk R, Gogler-Piglowska A, Widlak P, Vydra N. Active heat shock transcription factor 1 supports migration of the melanoma cells via vinculin down-regulation. Cell Signal. 2015;27(2):394–401.
20. Dobzhansky T. Nothing in biology makes sense except in light of evolution. Am Biol Teach. 1973;35(3):125–9.
21. Ghanbarian AT, Hurst LD. Neighboring genes show correlated evolution in gene expression. Mol Biol Evol. 2015;32(7):1748–66.
22. Liao BY, Zhang JZ. Coexpression of linked genes in mammalian genomes is generally disadvantageous. Mol Biol Evol. 2008;25(8):1555–65.
23. van Duin M, van Marion R, Vissers K, Watson JEV, van Weerden WM, Schroder FH, Hop WCJ, van der Kwast TH, Collins C, van Dekken H. High-resolution array comparative genomic hybridization of chromosome arm 8q: evaluation of genetic progression markers for prostate cancer. Genes Chromosomes Cancer. 2005;44(4):438–49.
24. Salinas CA, Kwon E, Carlson CS, Koopmeiners JS, Feng Z, Karyadi DM, Ostrander EA, Stanford JL. Multiple independent genetic variants in the 8q24 region are associated with prostate cancer risk. Cancer Epidemiol Biomarkers. 2008;17(5):1203–13.
25. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet. 2007; 39(5):638–44.
26. Di Giammartino DC, Shi YS, Manley JL. PARP1 represses PAP and inhibits polyadenylation during heat shock. Mol Cell. 2013;49(1):7–17.
27. Hollerer I, Curk T, Haase B, Benes V, Hauer C, Neu-Yilik G, Bhuvanagiri M, Hentze MW, Kulozik AE. The differential expression of alternatively polyadenylated transcripts is a common stress-induced response mechanism that modulates mammalian mRNA expression in a quantitative and qualitative fashion. RNA. 2016;22(9):1441–53.
28. Skaggs HS, Xing H, Wilkerson DC, Murphy LA, Hong Y, Mayhew CN, Sarge KD. HSF1-TPR interaction facilitates export of stress-induced HSP70 mRNA. J Biol Chem. 2007;282(47):33902–7.
29. Nagaike T, Manley JL. Transcriptional activators enhance polyadenylation of mRNA precursors. RNA Biol. 2011;8(6):964–7.
30. Xing HY, Mayhew CN, Cullen KE, Park-Sarge OK, Sarge KD. HSF1 modulation of hsp70 mRNA polyadenylation via interaction with symplekin. J Biol Chem. 2004;279(11):10551–5.
31. Beaudoing E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res. 2001; 11(9):1520–6.
32. Zhang HB, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. Genome Biol. 2005;6(12)
33. Hoque M, Ji Z, Zheng DH, Luo WT, Li WC, You B, Park JY, Yehia G, Tian B. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. Nat Methods. 2013;10(2):133–9.
34. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun. 2014;5
35. Singh P, Alley TL, Wright SM, Kamdar S, Schott W, Wilpan RY, Mills KD, Graber JH. Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. Cancer Res. 2009;69(24): 9422–30.
36. Mayr C, Bartel DP. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009;138(4): 673–84.
37. Morris AR, Bos A, Diosdado B, Rooijers K, Elkon R, Bolijn AS, Carvalho B, Meijer GA, Agami R. Alternative cleavage and polyadenylation during colorectal cancer development. Clin Cancer Res. 2012;18(19):5256–66.
38. Masamha CP, Xia Z, Yang JX, Albrecht TR, Li M, Shyu AB, Li W, Wagner EJ. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature. 2014;510(7505):412-+.
39. YG F, Sun Y, Li YX, Li J, Rao XQ, Chen C, AL X. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res. 2011;21(5):741–7.
40. Liaw HH, Lin CC, Juan HF, Huang HC. Differential MicroRNA regulation correlates with alternative polyadenylation pattern between breast cancer and normal cells. PLoS One. 2013;8(2)
41. Elkon R, Drost J, van Haaften G, Jenal M, Schrier M, Vrielink JAFO, Agami R. E2F mediates enhanced alternative polyadenylation in proliferation. Genome Biol. 2012;13(7)
42. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science. 2008;320(5883):1643–7.
43. Buchert M, Papin M, Bonnans C, Darido C, Raye WS, Garambois V, Pelegrine A, Bourgaux JF, Pannequin J, Joubert D, et al. Symplekin promotes tumorigenicity by up-regulating claudin-2 expression. Proc Natl Acad Sci U S A. 2010;107(6):2628–33.
44. Amara SG, Jonas V, Rosenfeld MG, Ong ES, Evans RM. Alternative RNA processing in calcitonin gene-expression generates messenger-RNAs encoding different polypeptide products. Nature. 1982;298(5871):240–4.
45. Alt FW, Bothwell ALM, Knapp M, Siden E, Mather E, Koshland M, Baltimore D. Synthesis of secreted and membrane-bound immunoglobulin-mu heavy-chains is directed by messenger-RNAs that differ at their 3′ ends. Cell. 1980; 20(2):293–301.
46. Wu XB, Bartel DP. Widespread influence of 3′-end structures on mammalian mRNA processing and stability. Cell. 2017;169(5):905-+.
47. Shi YS, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR, Frank J, Manley JL. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol Cell. 2009;33(3):365–76.
48. HuGO Gene Nomenclature Committee. ftp://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/hgnc_complete_set.txt. Accessed 26 Apr 2017.
49. COSMIC Cancer Gene Census. http://cancer.sanger.ac.uk/census. Accessed 1 May 2017.
50. NCBI HomoloGene. ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build68/. Accessed 1 May 2017.
51. UCSC Genome Browser. http://genome.ucsc.edu. Accessed 1 May 2017.
52. Pedersen BS, Yang IV, De S. CruzDB: software for annotation of genomic intervals with UCSC genome-browser database. Bioinformatics. 2013;29(23): 3003–6.
53. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4): 656–64.
54. Suzuki R, Shimodaira H. Pvclust: an *R* package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2.
55. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017;
56. WebGestalt 2017. http://www.webgestalt.org/option.php. Accessed May 2017.