Smestad and Maher *BMC Medical Genetics* (2015) 16:91
DOI 10.1186/s12881-015-0236-4

BMC Medical Genetics

RESEARCH ARTICLE Open Access

Relationships between putative G-quadruplex-forming sequences, RecQ helicases, and transcription

John A. Smestad[1] and L. James Maher III[2*]

Abstract

Background: Putative G-quadruplex-forming sequences (PQS) have long been implicated in regulation of transcription, though the actual mechanisms are not well understood. One proposed mechanism involves the activity of PQS-specific helicases belonging to the RecQ helicase family. However, patterns of PQS that correlate with transcriptional sensitivity to RecQ helicases are not well studied, and no adequate transcriptional model exists to account for PQS effects.

Methods: To better understand PQS transcriptional effects, we analyze PQS motifs in genes differentially-transcribed in Bloom Syndrome (BS) and Werner Syndrome (WS), two disorders resulting in loss of PQS-interacting RecQ helicases. We also correlate PQS genome-wide with transcription in multiple human cells lines while controlling for epigenetic status. Finally, we perform neural network clustering of PQS motifs to assess whether certain motifs are over-represented in genes sensitive to RecQ helicase loss.

Results: By analyzing PQS motifs in promoters of genes differentially-transcribed in BS and WS, we demonstrate that abundance of promoter PQS is generally higher in down-regulated genes and lower in up-regulated genes, and show that these effects are position-dependent. To interpret these correlations we determined genome-wide PQS correlations with transcription while controlling for epigenetic status. Our results identify multiple discrete transcription start site-proximal positions where PQS are correlated with either increased or decreased transcription. Finally, we report neural network clustering analysis of PQS motifs demonstrating that genes differentially-expressed in BS and WS are significantly biased in PQS motif composition.

Conclusions: Our findings unveil unappreciated detail in the relationship between PQS, RecQ helicases, and transcription. We show that promoter PQS are generally correlated with reduced gene expression, and that this effect is relieved by RecQ helicases. We also show that PQS at certain positions on the downstream sense strand are correlated with increased transcription. We therefore propose a new transcriptional model in which promoter PQS have at least two distinct types of transcriptional regulatory effects.

Background

Bloom Syndrome (BS) and Werner Syndrome (WS) are human genetic disorders resulting from loss-of-function mutations in DNA helicases belonging to the RecQ helicase family [1–6]. BS patients present with short stature, immunodeficiency, and photo-sensitivity. WS is classically associated with progeria. Both syndromes result in decreased fertility and increased carcinogenesis. The BLM and WRN helicases implicated in BS and WS, respectively, have both been shown to unfold G-quadruplex structures assembled *in vitro* [7, 8], in addition to other types of DNA structures. It has been proposed that BLM and WRN helicases function *in vivo* by resolving DNA structures, including intrastrand and interstrand G-quadruplexes hypothesized to form at putative G-quadruplex-forming sequences (PQS) during homologous recombination [9], base-excision repair (WRN only) [10], in telomeres during cellular replication [11–13], and in regulation of gene transcription [14]. Here PQS are identified by an algorithm (see Methods) analyzing

* Correspondence: maher@mayo.edu
[2]Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55905, USA
Full list of author information is available at the end of the article

the potential to form intrastrand G-quadruplexes under proper conditions if DNA strands were separated [15].

Analysis of transcriptional perturbations in BS and WS has identified effects on various cellular signaling pathways including control of growth/proliferation, death/survival, protein synthesis, gene expression, and development [16, 17]. Interestingly, it has also been noted that genes differentially expressed in BS and WS have increased PQS abundance, suggesting that transcriptional changes upon loss of RecQ helicases could result from failure to properly suppress G-quadruplex structures [14, 16]. Despite the absence of conclusive biochemical evidence for G-quadruplex structures at PQS in these genes, sequence-expression correlations are compelling.

Little is yet known concerning specific PQS motifs in the promoters of genes differentially sensitive to loss of BLM and WRN helicases. Such knowledge might provide insight into how PQS /RecQ helicase interactions modulate gene expression. In the current work, we therefore analyze the abundance and sequences of PQS within 2 kbp of transcription start sites (TSS) for genes differentially expressed in BS and WS. We elucidate statistically significant PQS abundance patterns in these genes vs. genes whose transcription is not altered by RecQ helicase loss. We also use a new approach to correlate PQS location with transcriptional activation or repression. The method applies epigenetic information to predict gene expression, with subsequent analysis of the modeling error and correlation with PQS position. These two methods map intrinsic PQS transcription regulatory effects, and predict how PQS abundance at discrete positions correlates with transcriptional changes upon BLM or WRN helicase loss, and lead us to propose a new transcriptional model in which promoter PQS have at least two distinct types of transcriptional regulatory effects. Finally, we analyze PQS motifs using neural network clustering to demonstrate that genes differentially-expressed in BS and WS are significantly biased in their PQS motifs. This suggests an unappreciated biological relationship between PQS, RecQ helicases, and transcription.

Methods

PQS mapping near TSS

GENCODE version 7 gene annotations (produced using GRCh37) were downloaded from www.gencodegenes.org/releases/7.html. The list of annotations was filtered to remove all entries with a tag of "cds_start_NF" or "low_sequence_quality", leaving a set of 156,253 GENCODE v7 transcripts. From this list of transcripts, 139,758 unique transcription start sites (TSS) were identified, representing 59,005 unique genes. PQS were then mapped to the 2 kbp upstream and downstream of identified TSS using PQS midpoint reference genomic coordinates (Additional file 1). R code for this procedure is provided in Additional file 2. A total of 88,058 unique PQS were mapped to within 2 kbp of a TSS. Since many genes have multiple TSS located within a few kbp of each other, some PQS were counted multiple times, with a total of 251,386 PQS mapping assignments to known TSS (Additional file 3).

BS and WS gene expression data

Differential gene expression data for Bloom Syndrome patient fibroblasts was obtained from the supplemental information section of reference [16]. Genes in this dataset had been identified by Affymetrix GeneChip Human Exon 1.0 ST arrays using manufacturer-recommended protocols, with the criterion for differential expression set at ≥ 1.5 absolute fold expression change with an adjusted p-value < 0.05 and FDR < 0.1. Datasets for genes differentially expressed in Werner Syndrome patient fibroblasts were obtained from reference [17] uploaded to GEO accession GSE48761. Genes in this dataset had been identified using Human Gene 1.0 ST Array (Affymetrix) using standard procedures. Datasets were downloaded from the GEO repository and processed in R using microarray analysis packages available from Bioconductor. Packages included hgu95av2cdf, hgu95av2.db, limma, marray, affy, affyQCReport, and affyPLM. WS datasets were normalized using the robust multiple-array average (RMA) algorithm, and genes differentially-expressed in WS patient samples were identified with the criterion of ≥ 1.5 absolute fold expression change compared to controls, with an adjusted p-value of < 0.05 and FDR < 0.05 (Additional file 4).

Analysis of PQS in genes differentially expressed in BS and WS

Genes differentially-expressed in BS and WS were divided into up- and down-regulated gene sets. All TSS were identified in genes differentially-expressed in BS and WS, together with all PQS positioned within 2 kbp of these TSS. PQS were annotated for sense or antisense strand and positions assigned in 200 bp bins repeated at a 10 bp interval from −2 kbp to +2kbp relative to TSS. This approach was also applied to map all PQS in the human genome, and an additional 100 times in a statistical bootstrapping method using a randomly-generated collection of genes of the same size as the test dataset. Test datasets and randomly-generated datasets were compared to the genome-wide dataset using a p-value generated from the prop.test function in R. P-values and the ratio of mean PQS per TSS for the comparison between the test datasets and genome-wide controls were plotted as a function of bin position relative to the TSS. The threshold for statistical significance was picked to be the p-value below which a data point in the randomly generated dataset has a 1 % chance of false rejection of

the null hypothesis. FRD were calculated as the ratio of predicted number of false positives data points (3.81 per 381 data points) to the number of data points in the test dataset that pass the threshold *p*-value. R code for this analysis can be found in Additional file 5.

Epigenetic prediction of gene expression

The generation of predictive models for gene expression based on epigenetic data followed a method similar to that previously described [19]. Epigenetic and gene expression data were obtained from the ENCODE project through the NCBI Gene Expression Omnibus online repository (accession GSE34448, GSE32970, and GSE29611). Cell lines used in the epigenetic modeling of gene expression include H1-hESC (embryonic stem cell), HeLa-S3 (cervical cancer), K562 (immortalized myelogenous leukemia), HUVEC (human umbilical vein endothelial cell), HepG2 (hepatocellular carcinoma), NHEK (normal human epidermal keratinocyte), and GM12878 (EBV-transformed lymphoblastoid cell). For each of these cell lines, genome-wide tracks for histone modifications H3K9ac, H3K4me3, H3K4me2, H3K27ac, H3K79me2, H3K36me3, H3K4me1, H3K27me3, H4K20me1, histone variant H2A.Z, and chromatin accessibility (quantified by digital DNAse I hypersensitivity) were obtained, along with gene expression datasets quantified by cap analysis of gene expression (CAGE) technology.

GENCODE version 7 transcript annotations were used to map individual CAGE-quantified transcript levels to known TSS. Total transcriptional activity for each TSS was calculated as the sum of the expression values for all transcripts that originate from that TSS. Since many genes contain multiple TSS, only the TSS with the highest aggregate expression level for each of the 59,005 unique genes in the genome was retained for analysis. R code for these data manipulations can be found in Additional files 6, 7, and 8.

The GC fraction for the 5 kbp upstream of each of the identified strongest TSS was calculated using tools in the R package Biostrings produced by the open source Bioconductor software project. Computation for this part of the analysis was run on the Mayo Research Computing Facility (RCF) shared-resource, Beowulf-style Linux cluster using R version 3.0.2, with scripts written to accommodate batch mode execution managed by the Open Grid Engine open-source batch-queuing system.

ENCODE data for DNAse I hypersensitivity, histone modification, and histone variant tracks were obtained from the GEO repository as described above. Track signal files were downloaded in .bigwig format and processed using utilities in the R package rtracklayer from Bioconductor. Track signals within 1 kbp upstream and downstream of each TSS were extracted and the signal within this region was split and averaged into twenty 100-bp bins, spanning from 1 kbp upstream to 1 kbp downstream of each TSS. R code for this procedure can be found in Additional files 9 and 10. For each of the 20 bins constructed for each genomic track, a correlation coefficient was calculated between the $\log_2$-transformed bin signal (with a small pseudocount added to avoid the $\log_2(0)$ issue) and $\log_2$-transformed gene expression signal quantified via CAGE, excluding TSS with a expression value of 0. For each genomic feature, the bin that had the highest correlation with expression was selected as the "best bin" for analysis. The optimal pseudocount value for each genomic track was determined by repeating the correlation analysis described above, but with pseudocounts ranging from 0.25-5 % of the maximal binned signal for that track feature. For each genomic track, the best pseudocount and "bestbin" combination was selected by identifying the bin and pseudocount combination that resulted in the highest absolute correlation with gene expression. R code for this procedure can be found in Additional file 11.

"Bestbin" and pseudocount combinations for each epigenetic track, along with the GC% 5 kbp upstream of each TSS were then used to construct predictive models of gene expression. TSS with missing values for any of the genomic tracks were excluded, leaving between 4,011 and 9,614 TSS, depending on the cell line. Datasets for the cell lines used in the analysis can be found in Additional files 12, 13, 14, 15, 16, 17, and 18. The subset of TSS containing no PQS within 2 kbp of the TSS was then divided into equal training and test datasets using a random sampling approach. The subset of TSS containing at least one PQS was included in the test dataset. All data channels in training and test datasets were normalized to be mean-centered at 0 and to have a standard deviation of 1. A Bayesian linear regression model for predicting gene expression from epigenetic parameters and GC fraction was then generated using the training dataset and the R package tgp. This model was then applied to the test dataset (not used in the training of the model) to generate gene expression predictions.

Modeling PQS correlation with transcription

For each TSS in the epigenetic modeling of gene expression test dataset, a prediction error was calculated as the CAGE-determined TSS-specific expression value minus the Bayesian linear regression model prediction from the epigenetic data. The portion of the test dataset TSS containing at least one PQS was extracted and the position and strand of the nearest PQS to each TSS was determined. Using an iterative approach, TSS-specific prediction error data for sense and antisense strands were aggregated based on the location of the nearest PQS, using a 200 bp bin size, and running 2 kbp upstream to 2 kbp downstream at an iteration interval of 10 bp. For

each bin, the distribution of prediction errors was compared to the prediction errors of the TSS in the test dataset containing no PQS using the oneway.test() function in R for one-way ANOVA. As a statistical control to assess the likelihood of obtaining a low *p*-value in the comparison by random chance, a statistical bootstrapping approach was employed in which the prediction errors in the PQS dataset were replaced with random prediction errors from the test dataset, and the *p*-values were calculated as above. This comparison was repeated 100 times and the threshold *p*-value for determining statistical significance of the test dataset comparison was picked to be the *p*-value below which a data point in the control dataset has a 1 % chance of false rejection of the null hypothesis. *P*-values in the test dataset below this threshold value were selected as statistically significant. FDR were calculated as the ratio of significant *p*-value data points in the test dataset compared to the average number of false positives in the control datasets (3.81 false discoveries per 381 data points). R code for this entire analysis can be found in Additional file 19.

PQS motif analysis

The number of G stacks containing at least 3 contiguous guanine nucleotides in each PQS within 2 kbp of a TSS was calculated for all human genes using a pattern-matching algorithm. The nucleotide fractions of adenine, thymine, and cytosine in each PQS were calculated using the package Biostrings. For the subset of PQS containing 4 G stacks, with each G stack containing 3 bases, lengths of loop sequences between G stacks were also calculated. In order to avoid ambiguity in this analysis, it was required that the first and last base in a loop sequence not be guanine. Computation was performed on the Mayo Research Computing Facility (RCF) shared-resource, Beowulf-style Linux cluster using R version 3.0.2, with scripts written to accommodate batch mode execution managed by the Open Grid Engine open-source batch-queuing system.

PQS were binned based on position, using a 200 bp bin width calculated at a 50 bp interval, spanning from 2 kbp upstream to 2 kbp downstream of the TSS. Subsets of PQS data for genes up- and down-regulated in BS and WS were extracted, and PQS features (total length, loop lengths, number of G-stacks, fractions adenine, cytosine, and thymine) were compared individually between genes differentially expressed in BS and WS and the remainder of the genome, bin by bin, using one-way ANOVA. Comparisons were done independently for PQS in the sense and antisense DNA strands. This bin-based comparison was repeated 100 times in a statistical bootstrapping method using a randomly-generated collection of genes of the same size as the test dataset. The threshold for statistical significance was picked to be the *p*-value below which a data point in the randomly generated dataset has a 1 % chance of false rejection of the null hypothesis. False discovery rates were calculated as the ratio of predicted number of false positives data points to the number of data points in the test dataset that pass the threshold *p*-value. R code for this analysis can be found in Additional file 20.

Self-organizing map multidimensional classification of PQS

From the entire human genome, the subset of PQS with 4 G-stacks and 3 loops within 2 kbp of a TSS was selected, excluding PQS in which the first or last base in a loop was guanine. This selection criterion reduced the number of surveyed PQS from 88,058 to 17,795. Each PQS in this selected dataset was classified for total sequence length, loop lengths, and fractions of adenine, cytosine, and thymine. Self-organizing map artificial neural network analysis was implemented in R using the package kohonen. Maps consisting of 25 nodes in a hexagonal grid were trained using the PQS dataset, 100 iterative presentations of the data to the model, and a learning rate with linear decline from 0.05 to 0.01. The average number of PQS per gene per node was calculated and compared to the same calculation repeated for the PQS and gene subset for genes up- and down-regulated in BS and WS. Log_2 enrichment ratios were calculated for each node in the BS and WS datasets. As a means to ascertain whether the enrichment of PQS in certain nodes could arise by chance, a statistical bootstrapping method was employed in which enrichment ratios of nodes on the PQS map were calculated for 100 randomly-generated gene sets of the same size as the BS and WS datasets. Mean and 2σ values were calculated node by node for these random distributions. Statistical significance for enrichment ratios of nodes in the BS and WS datasets was assigned on the basis of lying outside of the 95 % CI for the random distributions. R code for this analysis can be found in Additional file 21. Motif classifications of all genomic PQS can be found in Additional file 22.

Results

PQS abundance in genes differentially expressed in BS and WS

We hypothesized that PQS abundance in promoters of genes differentially expressed in BS and WS reflects transcriptional sensitivity to RecQ helicase activity. We used published gene expression array datasets comparing BS and WS patient fibroblasts to normal control fibroblasts to identify genes that are significantly up- and down-regulated in BS and WS [absolute fold expression change ≥ 1.5 with an adjusted *p*-value < 0.05 and false discovery rate (FDR) < 0.1 for BS genes and FDR < 0.05 for

WS genes] [16, 17]. The BS dataset consisted of 1012 up-regulated genes and 141 down-regulated genes [16], and the WS dataset consisted of 1046 up-regulated genes and 540 down-regulated genes. Comparing genes identified as differentially-expressed in each syndrome, we find them to be largely non-overlapping (Additional file 23: Figure S1). Notably, the fraction of these differentially-expressed genes that contains at least one PQS within 2 kbp of a TSS (BS: 84 % of up-regulated genes and 90 % of down-regulated genes, WS: 74 % of up-regulated genes and 84 % of down-regulated genes), is high compared to the genomic average of 55 % (Additional file 23: Table S1).

We analyzed how PQS abundance varies as a function of position and strand (sense or antisense) near promoters of genes differentially expressed in BS and WS, compared to other genes. Histograms showing PQS abundance (raw counts) as a function of position upstream and downstream of the TSS on sense and antisense strands are shown in Additional file 23: Figure S2. We then compared PQS abundance between genes differentially expressed in BS and WS vs. all other genes. A heat map showing PQS abundance (raw counts) as a function of position, normalized to total number of TSS in the respective datasets, is presented in Fig. 1a. We also conducted quantitative analysis of PQS abundance in genes differentially expressed in BS and WS vs. all other genes, using 200 bp bins, repeated at a 10-bp interval, spanning 2 kbp upstream and downstream of known TSS on both strands. For each bin, 1-way ANOVA *p*-values were calculated comparing PQS abundance in the differentially-expressed gene dataset to

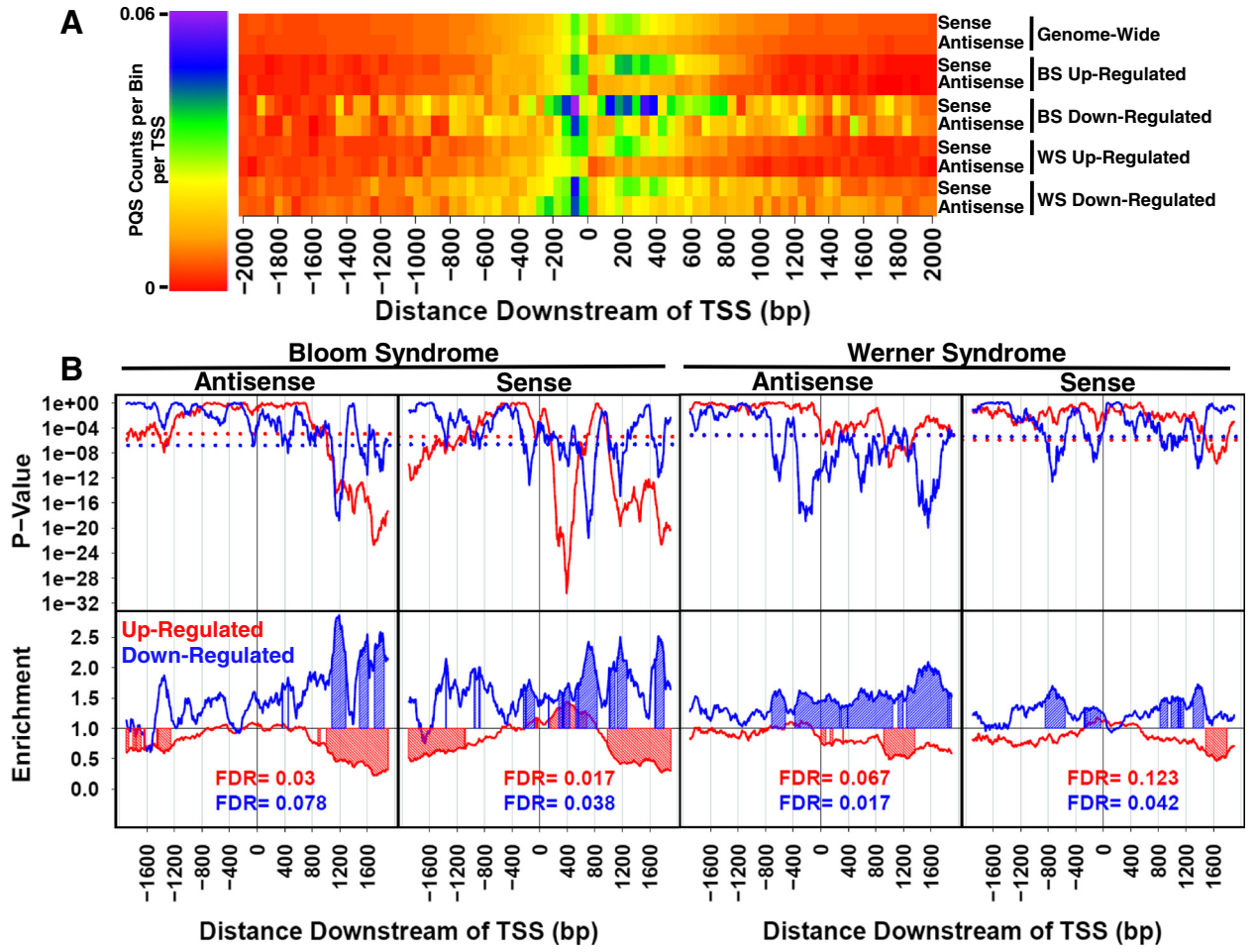

Fig. 1 PQS occurrence in genes differentially expressed in BS and WS. PQS occurrence in sense (S) and antisense (AS) strands was analyzed independently. **a** Heat map showing PQS occurrence genome-wide and in genes differentially-expressed in BS and WS, represented as raw counts normalized to total number of TSS in each dataset. **b** Top panels show *p*-values for comparing PQS abundance per TSS between genes differentially expressed in BS and WS and all other genes. Dotted lines represent p-value cutoffs for determining statistical significance, with less than 1 % of data points from a random gene dataset of the same size as the test dataset having a *p*-value below this threshold. Bottom panels show PQS enrichment ratio in genes differentially expressed in BS and WS, with values > 1 indicating that PQS are more abundant in the differentially-expressed gene set. Both *p*-values and enrichment ratios were calculated using a 200 bp bin value repeated at a 10 bp interval. Regions with shaded peaks represent locations of statistically-significant PQS excess or scarcity

all other genes. The statistically-significant result, shown in Fig. 1b, is intriguing. Genes up- and down-regulated in BS and WS tend to have opposite patterns of PQS abundance near TSS. Relative to the remainder of the genome, genes up-regulated in BS and WS tend to have fewer PQS, while genes down-regulated in BS and WS tend to have more PQS. An exception to this pattern is that BS up-regulated genes have more PQS between 160 and 680 bp downstream of the TSS on the sense strand. Other PQS abundance patterns are listed in Additional file 23: Table S2. Threshold *p*-values and FDR for the analysis, determined by statistical simulation with randomly-generated gene sets, are presented in Additional file 23: Table S3.

Correlation of PQS patterns and transcription

To interpret PQS abundance for genes differentially expressed in BS and WS, we studied how PQS location relative to the TSS correlates with gene expression for all human genes. A previous study correlated PQS position and gene transcription while controlling for gene family, function, and promoter similarity [18]. In that work, PQS from 0 to 500 bp downstream of TSS were correlated with increased gene expression. PQS from 0 to 500 bp upstream of the TSS were not significantly correlated with gene expression. In our analysis, positions found to have altered PQS abundance in genes differentially expressed in BS and WS often occur more than 500 bp from TSS (86 % for BS, 71 % for WS). No data were available for genome-wide correlation between transcription and PQS in such positions.

We therefore implemented a novel analysis combining published epigenetic and gene expression data from 7 human cell lines (H1-hESC, HeLa-S3, K562, HUVEC, HepG2, NHEK, and GM12878) to correlate PQS positions within 2 kbp of TSS genome-wide with gene expression. The method successfully predicts gene expression using epigenetic data, and sorts calculated prediction errors (actual expression minus predicted expression) by PQS position near TSS. The method is unique in its robust statistics that control for gene epigenetic signature to isolate regulatory effects not detectable by standard association-based methods. Our approach to modeling gene expression was inspired by a prior publication [19]. Bayesian linear regression models were trained on TSS-specific cap analysis of gene expression (CAGE) measurements, epigenetic data including abundance of local histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H4K20me1), histone variant H2A.Z, and GC fraction 5 kbp upstream of the TSS. For each histone modification and variant, we identified the position near the TSS where the epigenetic signature best correlated with gene expression (Fig. 2a, Additional file 23: Figure S3). Model training was conducted using half of the PQS-free TSS (i.e.no PQS within 2 kbp of the TSS) for each cell line. Resulting models were then applied to the remaining TSS in the dataset (the remaining half of all PQS-free TSS as well as all PQS-containing TSS) to generate gene expression predictions (Additional file 23: Tables S4, S5, and S6). Thus, each TSS in the test dataset has both predicted and measured transcription values. A representative plot of predicted vs. measured transcription is shown in Fig. 2b. The difference between measured and predicted transcription is the prediction error. Prediction error values > 0 indicate greater-than-predicted transcription, and values < 0 indicate less-than-predicted transcription. Prediction errors for genes with ≥ 1 PQS within 2 kbp of the TSS were sorted based on PQS position and this distribution compared to the distribution of prediction errors for PQS-free TSS using 1-way ANOVA. Statistical methods were used to determine threshold *p*-values and to estimate the FDR by repetition of the analysis after replacement of the prediction error values with randomly sampled values from the control distribution. A representative plot for this type of analysis is shown in Fig. 2c.

This analysis was repeated for each of 7 human cell lines, generating a composite analysis of all statistically-significant PQS positions correlated with gene expression (Fig. 2d). False discovery rates (FDR) for the analysis of sense and antisense strands were below 10 % for four of the seven cell lines tested, indicating that the observed correlations have very low probability of being statistical noise (Additional file 23: Tables S4). The GM12878 cell line notably had a very high FDR for both DNA strands (47 % for antisense strand and 54 % for sense strand). The reason for this is unclear, but the high FDR reflects the presence of fewer identified positions where PQS significantly correlated with altered transcription. This is the first PQS correlation analysis incorporating epigenetic data from multiple cell lines and extending the study region to 2 kbp upstream and downstream of TSS. Interestingly, the general agreement for data from different cell lines suggests that PQS position correlates with gene expression regardless of epigenetic context. The analysis correlates PQS position in the antisense strand with lower gene expression, regardless of PQS position upstream or downstream of the TSS. In contrast, PQS in the sense strand show different correlations depending on position. PQS in the sense strand are correlated with lower gene expression, except for PQS positioned downstream of the TSS between 140–270 bp, 1750–1770 bp, and 1900 bp. PQS at these three sense strand positions were correlated with increased gene expression. This analysis did not find PQS correlations with gene expression at all locations. In total, 33 % and 35 % of analyzed PQS positions in the antisense and sense strands, respectively, displayed statistically-significant PQS abundance correlated with gene expression. These data

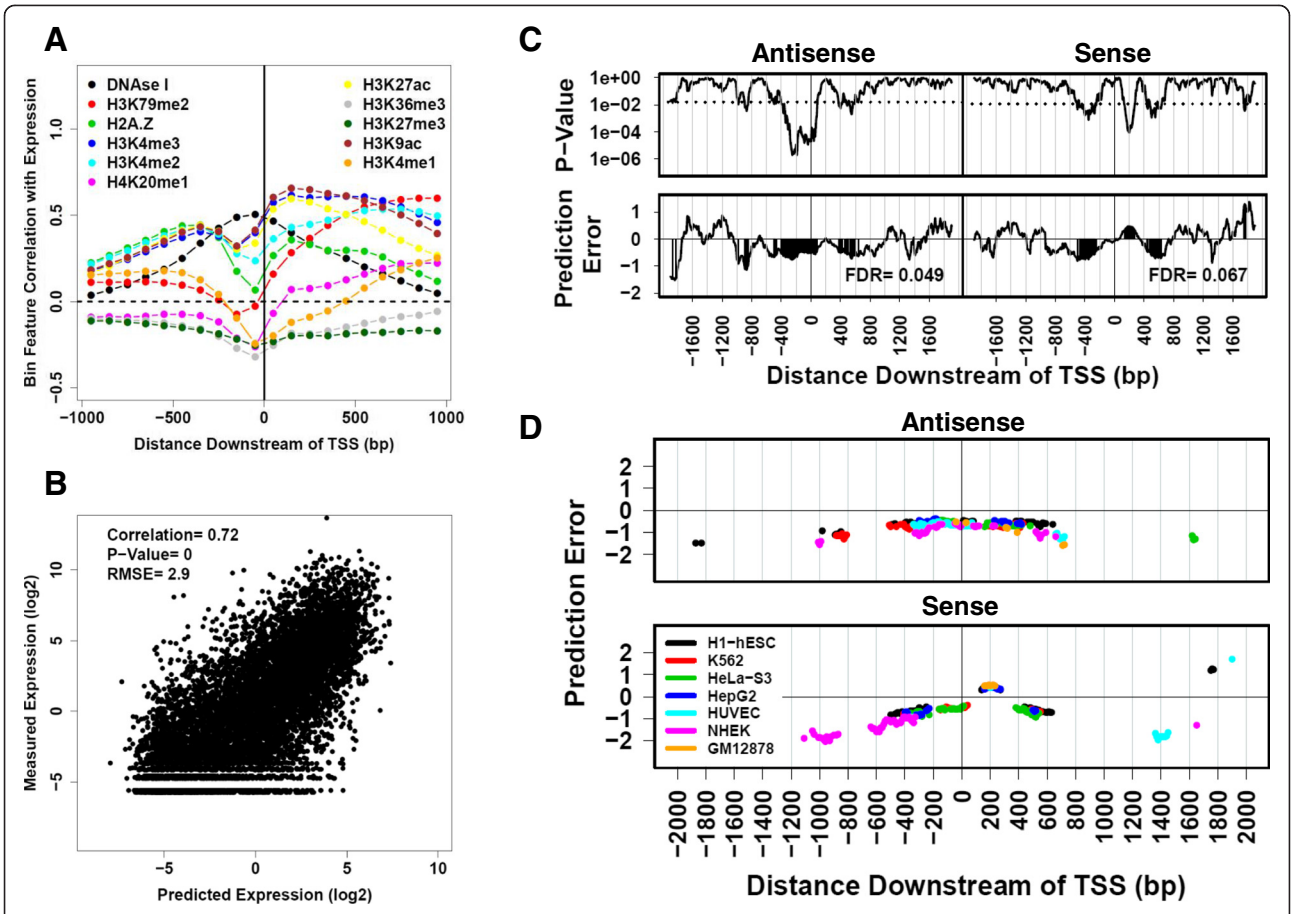

Fig. 2 Epigenetic prediction of gene expression to identify PQS positions correlated with altered gene expression. **a** Correlation of $\log_2$-transformed epigenomic track signals with $\log_2$-transformed gene expression represented as a function of location 1 kbp upstream and downstream of TSS. Bin size is 100 bp. Prior to $\log_2$-transformation, 0.25 % of the maximal bin value was added to each bin feature to avoid the $\log_2(0)$ issue. **b** Representative correlation of CAGE TSS-specific gene expression measurements with epigenetic model predictions generated through Bayesian linear regression. Models were trained on datasets containing half of all PQS-free TSS in the human genome. Gene expression predictions were then generated for the remainder of the PQS-free TSS. **c** Analysis of model prediction error as a function of PQS position, assigned to sense (S) and antisense (AS) strand effects. Prediction error estimates were sorted into 200-bp bins iterated at a 10-bp interval, based on the position of the PQS nearest to the TSS, and then compared to the prediction error estimates for PQS-free TSS using 1-way ANOVA. The top panels show the p-value from this analysis as a function of position with respect to the TSS. The dashed line represents the *p*-value threshold used for determining statistical significance in the analysis. The bottom panels show prediction error averages for the binned values. Areas with shaded peaks represent locations where the *p*-value is statistically significant for rejection of the null hypothesis. Prediction error value > 0 represent positions where PQS presence correlates with higher gene expression. Prediction error values < 0 represent positions where PQS presence correlates with lower gene expression. **d** Aggregate map of all statistically-significant positions from prediction error analysis based on seven human cell lines

represent a large improvement over previous studies in terms of resolution and coverage.

It was then possible to compare position-dependent PQS correlations with gene expression for all genes to data for genes differentially expressed in BS and WS (Fig. 3). For both DNA strands of genes up-regulated in BS and WS, TSS-proximal PQS are under-represented at positions correlated with low gene expression. Similarly, for both DNA strands of genes down-regulated in BS and WS, TSS-proximal PQS are over-represented at positions correlated with low gene expression. These observations are together consistent with a model in which the BLM and WRN helicases suppress some inhibitory effect of PQS during transcription. It is tempting to speculate that this suppression involves destabilizing non-B-DNA structures at PQS. Interestingly, genes up-regulated in BS tend to feature a region of increased PQS abundance 160–680 bp downstream of the TSS on the sense strand, overlapping a region 140–270 bp downstream of the TSS correlated with increased gene expression. It is difficult to account for this observation, and it suggests a possible BLM helicase-independent correlation between gene

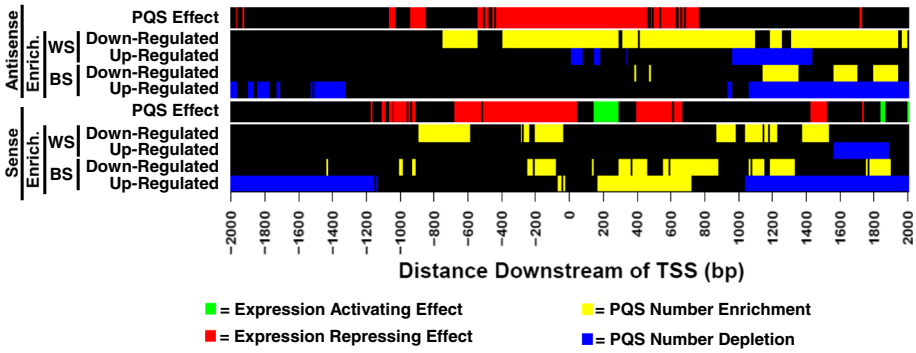

Fig. 3 Analysis of PQS abundance in BS and WS along with position-dependent correlations with transcription. Heat map showing PQS excess or scarcity in genes differentially expressed in BS and WS (yellow = excess, blue = scarcity), as well as the correlation of PQS at these positions with transcription (green = high transcription, red = low transcription)

expression and PQS positioned 140–270 bp downstream of the TSS on the DNA sense strand.

PQS motifs in genes differentially expressed in BS and WS

To examine whether particular PQS motifs are correlated with transcription effects upon RecQ helicase loss, we analyzed PQS length, number of G-stacks, base composition, and loop lengths for genes differentially expressed in BS and WS, comparing these results to all other PQS. Loop lengths were calculated only for PQS containing four G-stacks, and lacking guanine at the first or last position of any loop sequence. PQS motifs in genes differentially-expressed in BS and WS were then compared to all other genes as a function of position upstream or downstream of the TSS (Fig. 4). The results illustrate the interesting finding that PQS motifs in genes differentially expressed in BS and WS are statistically different from PQS in the remainder of the genome. Interestingly, these differences are consistent regardless of PQS position upstream or downstream of the TSS. FDR for this analysis was determined by statistical modeling using randomly-generated gene datasets, and is presented in Additional file 23:

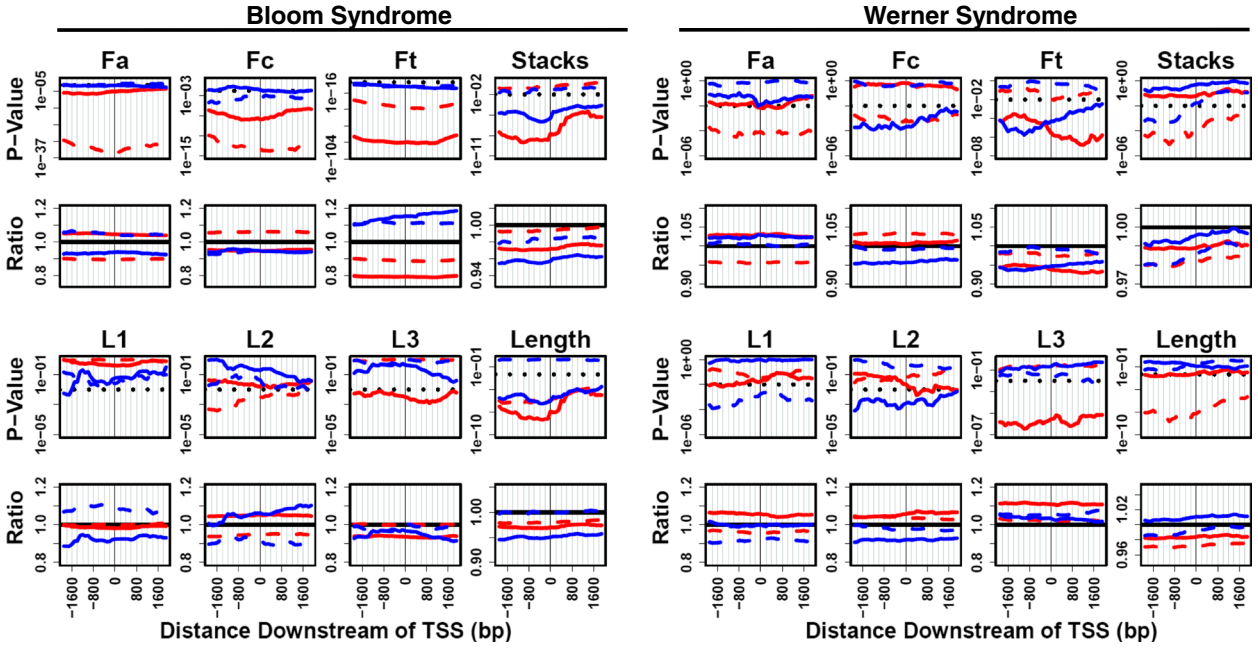

Fig. 4 PQS motif characteristics in genes differentially expressed in BS and WS analyzed via ANOVA. PQS motif characteristics (total sequence length, loop lengths, base fractions, and number of G-stacks) in genes differentially expressed in BS and WS. *P*-values and ratios for comparison to the genome-wide distribution are calculated using a 200 bp bin selection, repeated at a 50 bp interval, and in reference to corresponding PQS position on antisense (AS) strand or sense (S) strand for all genes. Red = genes up-regulated in BS and WS. Blue = genes down-regulated in BS and WS. Solid line = AS. Dashed line = S

Table S7A. The position-independent similarity of PQS motifs in genes differentially expressed in BS and WS allowed an aggregate analysis of these motifs in comparison with all other PQS motifs near promoters of other genes (Additional file 23: Table S7B).

Multidimensional PQS motif clustering

We further analyzed PQS motifs present in genes differentially expressed in BS and WS using a multidimensional self-organizing map (SOM) neural network classification. SOMs were historically designed to represent a high-dimensional space as a simple two-dimensional topological map [20]. We implemented SOM clustering of all PQS containing four G stacks. Using a 5X5 input matrix, SOM clustering was conducted using PQS total motif length, loop lengths, and base composition as the input dataset for training. PQS features for motif centroids of the 25 nodes resulting from the clustering protocol are presented in Fig. 5a. The number of PQS counts per node, calculated on the basis of shortest multidimensional distance, is presented in Fig. 5b. The multidimensional distance between nodes is represented in Fig. 5c. To ascertain whether particular PQS motifs are associated with genes differentially expressed in BS and WS, we extracted these PQS motifs for these genes and calculated the average number of PQS per node, normalized to the total number of genes in the dataset. The same gene number-normalized calculation was repeated for the genome-wide dataset. A 95 % confidence interval for normal statistical variability in cluster composition was also determined by statistical modeling conducted by repeating the same calculation on randomly-generated gene sets of the same size as the differential expression gene datasets. This comparison of PQS motif clusters from genes differentially expressed in BS and WS confirmed that certain PQS motifs are enriched in a

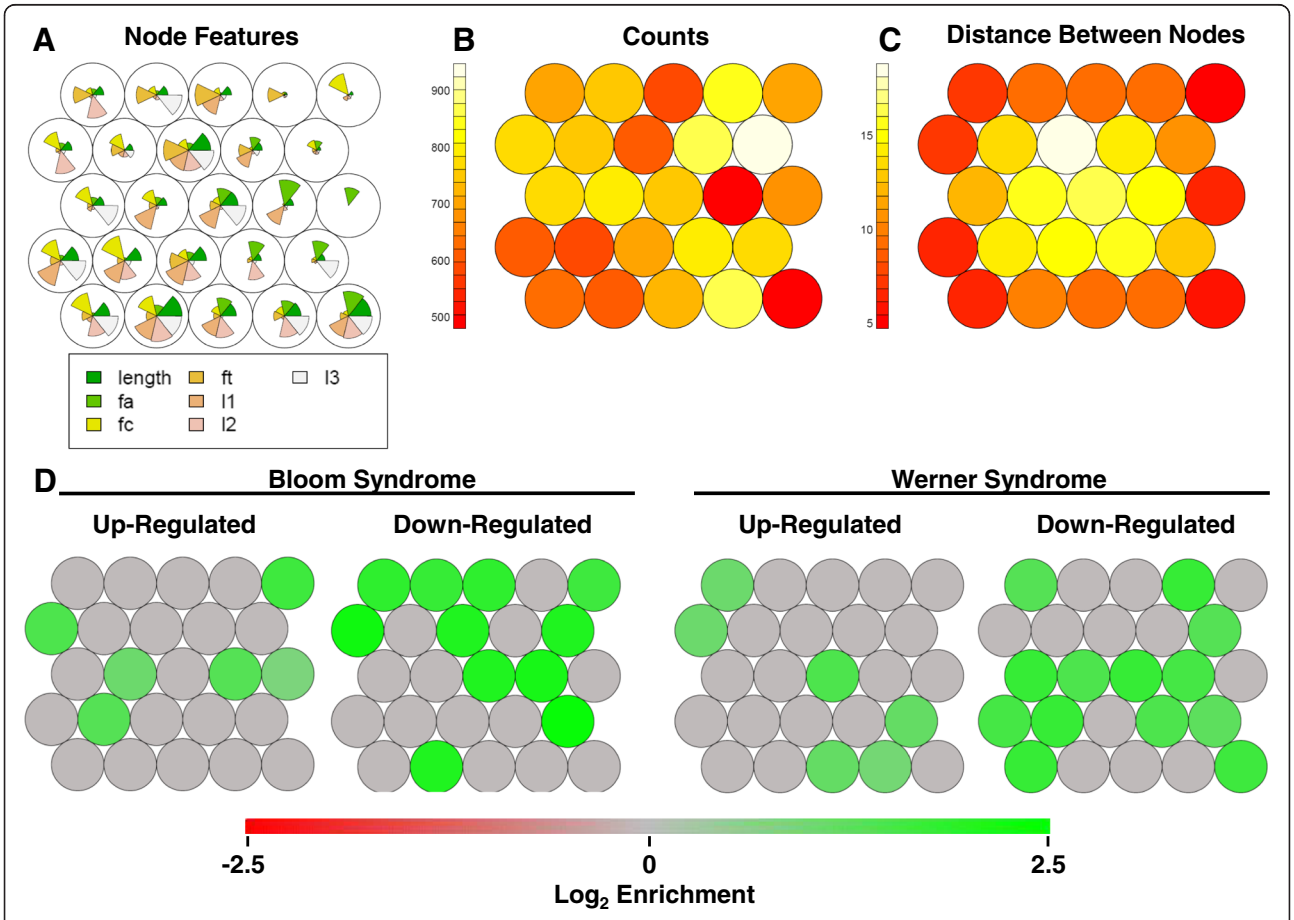

Fig. 5 Self-organizing map analysis of PQS motifs in genes differentially-expressed in BS and WS. Multidimensional self-organizing map clustering of all human genome PQS within 2 kbp of TSS, based on total sequence length, loop lengths, number of G-stacks, and nucleotide base fractions. **a** PQS parameters for node centroids. **b** Counts of PQS binned by distance in multidimensional space to closest node. **c** Distance between nodes. **d** PQS node bias in genes differentially expressed in BS and WS. Bias represents number of unique PQS per node, normalized to dataset gene number, compared to the same calculation for all genes genome-wide. Green and red nodes represent positive (excess) or negative (scarcity) bias values that are outside of the 95 % CI for randomly-generated gene datasets of the same size. Green = excess. Red = scarcity. Gray = no statistically-significant difference

statistically-meaningful fashion (Fig. 5d). This provocative finding reveals a PQS motif bias in genes differentially expressed upon loss of RecQ helicases.

Discussion

We hypothesized that PQS patterns in genes differentially transcribed in BS and WS reveal transcriptional effects of PQS. In essence, BS and WS are natural RecQ helicase knockout experiments where transcriptional effects of persistent PQS non-B DNA structures are revealed. We further hypothesize that PQS in genes differentially-expressed in BS and WS are biased in their composition. We report analyses that support these hypotheses.

Promoter PQS abundance correlates with transcriptional effect upon RecQ helicase loss

Regarding the first hypothesis, our analysis of PQS in promoters of genes differentially-expressed in BS and WS shows that patterns of PQS abundance in up- and down-regulated genes are opposite (Fig. 1). Genes up-regulated upon RecQ helicase loss have a scarcity of promoter PQS and down-regulated genes have an abundance of PQS. This is reflected in the higher or lower numbers of PQS per TSS calculated for these gene sets compared to the genomic average (BS up-regulated: 1.61, BS down-regulated: 2.80, WS up-regulated: 1.58, WS down-regulated: 2.42, genome-wide average: 1.80; Additional file 23: Table S3). Interestingly, up- and down-regulated genes in BS and WS are more likely to have at least one PQS within 2 kbp of a TSS (BS: 84 % of up-regulated genes and 90 % of down-regulated genes, WS: 74 % of up-regulated genes and 84 % of down-regulated genes, genome-wide average: 55 %; Additional file 23: Table S1). The implication of these results is that genes up-regulated in BS and WS have fewer PQS per TSS than the genomic average, but, surprisingly, are unlikely to have zero PQS motifs. Our results support a model in which promoter PQS generally repress transcription and RecQ helicases tend to moderate this repression.

Promoter PQS abundance is altered at discrete positions in genes sensitive to RecQ helicase loss

Beyond PQS abundance, we find that PQS position is important. We have demonstrated that PQS abundance in genes differentially expressed in BS and WS is not randomly distributed across promoters, but discrete positions of abundance/scarcity are detected. The most striking example of this position-dependence is on the sense strand of genes up-regulated in BS (Fig. 1b). Here PQS are scarce > 1 kbp upstream and downstream, but abundant 160–680 bp downstream of the TSS. This subtle and perplexing PQS position-dependence is missed when evaluating only aggregate PQS numbers per TSS.

Our analysis expands the current state of knowledge regarding PQS abundance in genes differentially expressed in BS and WS. Johnson et al. have shown that PQS are more abundant in non-coding regions of genes up-regulated in BS and WS [14]. This study, however, did not find statistically-significant correlation between PQS and genes down-regulated in BS and WS. Recent work improved upon this analysis by demonstrating increased PQS abundance in genes down-regulated in BS up to 250 bp upstream of the TSS on the antisense strand and at the 5′ end of the first intron on the sense strand [16]. Genes up-regulated in BS were also found to have increased PQS abundance flanking the TSS (within 250 bp) on the antisense strand and at the 5′ end of the first intron on both sense and antisense strands. Interestingly, our analysis does not corroborate the findings of Nguyen et al. that PQS abundance is increased up to 250 bp upstream of the TSS on the antisense strand in genes down-regulated in BS, nor the finding that PQS abundance is increased flanking the TSS on the antisense strand for up-regulated genes. We suggest that the discrepancy is due to the rigorous statistical criteria in our analysis, and the higher resolution (smaller bin size) than was used by Nguyen et al. Thus, our data analysis examines PQS patterns further upstream and downstream of the TSS (a full 4 kbp window), and at a higher resolution (200 bp) than any previous study. We provide the first evidence that PQS abundance is significantly lower in the promoter-proximal region of genes up-regulated in BS and WS. The finding that PQS abundance is generally higher in genes down-regulated in BS and WS and lower in genes up-regulated genes in BS and WS, compared to the genomic average, provides an important insight. Again, these results support a model in which promoter PQS are generally repressive of transcription and RecQ helicases tend to modulate this repression.

PQS position correlates with transcriptional effect genome-wide

We have also correlated PQS position with gene expression while controlling for epigenetic status to isolate intrinsic PQS transcriptional effects (Fig. 2). For 33 % and 35 % of analyzed PQS positions in the antisense and sense strands, respectively, there was a statistically-significant correlation between PQS abundance and gene expression. Of these statistically-significant regions, 100 % and 84 % on antisense and sense strands, respectively were correlated with reduced expression. This suggests that transcriptional repression may be the dominant effect of promoter-proximal PQS, although certain PQS positions are correlated with increased gene expression (e.g. 140–270 bp downstream of TSS on the sense strand). The finding that PQS are correlated with increased or decreased

transcription in a position-dependent manner is extremely intriguing and has never been reported before.

While there is general agreement with results of previous work, our genome-wide correlation of promoter PQS with gene expression provides a significant improvement over previous analyses. A prior analysis [18] also correlated PQS in the sense strand up to 500 bp downstream of the TSS with higher gene expression. This prior analysis, which controls for attributes of gene family, function, and promoter similarity, did not find statistically significant correlation of PQS upstream of the TSS with gene expression. In contrast, the robust statistical analysis reported here, which controls for epigenetic status, shows that PQS upstream of the TSS on both DNA strand are correlated with lower gene expression. In addition to this greater sensitivity, our genome-wide correlation of promoter PQS and gene expression covers a larger genomic window (2 kbp upstream and downstream of each TSS) at a higher resolution (200 bp) than previously reported by Zhuo et al. (500 bp). For many of the genes differentially expressed in BS and WS, important PQS positions are > 500 bp distal from the TSS (86 % for BS, 71 % for WS). The study by Zhuo et al. was not able to provide insight into gene expression correlations for PQS in these positions. Additionally, the methodology of our study utilizing simultaneous analysis of PQS abundance in genes differentially expressed in BS and WS, and genome-wide correlation of promoter PQS with gene expression are unique (Fig. 3).

PQS motifs are biased in genes sensitive to RecQ helicase loss

Regarding the second hypothesis of this work, we show that PQS motifs in genes differentially expressed in BS and WS are significantly biased in their composition (Figs. 4 and 5). This indicates that specific sequence patterns may be important in PQS biological function. The reason for this is not clear, but it could reflect specificity in the PQS-helicase interaction or else selection of certain motif patterns for yet unknown reasons. This is the first demonstration of PQS motif bias in a differentially-expressed gene set.

PQS abundance and correlations with transcription suggest a hypothetical transcriptional regulatory model

The data from the present study clearly correlate PQS motifs with increased or decreased transcription in a position- and strand-dependent manner. In the absence of biochemical evidence, however, caution is urged in attributing PQS transcriptional correlation to effects due to non-B DNA structures. Nonetheless, these data might be interpreted as evidence that non-B DNA G-quadruplex structures have position-dependent regulatory effects on gene expression, and these effects can be modified by RecQ helicases. We therefore propose a hypothetical transcriptional model that can account for the observed PQS transcriptional correlations via position- and strand-dependent effects of intrastrand G-quadruplex structures (Fig. 6).

In this model, PQS located between 140 and 270 bp downstream of the TSS on the sense strand have a transcription-activating effect mediated by sense strand G-quadruplex formation, resulting in greater accessibility of the antisense strand to RNAPII, facilitating transcriptional initiation. This model for PQS-mediated transcriptional activation has previously been described by Du et al. [18]. A slight difference, however, is that we propose that this mechanism only applies to the sense strand at 140–270 bp downstream, verses both DNA strand located 0–500 bp downstream. An intriguing addition to the model includes the RecQ helicases as inhibitors of intrastrand G-quadruplex formation, which may functionally inhibit this transcriptional activation mechanism. The rationale for this addition to the model is the known function of the RecQ helicases as G-quadruplex resolving enzymes [7, 8], and also to the fact that both BS and WS have many more up-regulated genes than down-regulated genes (BS: 1012 up-regulated, 141 down-regulated, WS: 1046 up-regulated, 540 down-regulated (Additional file 23: Table S1)). RecQ helicases as inhibitors of G-quadruplex formation at this location may account for the fact that many genes are transcribed at higher levels upon RecQ helicase loss.

The majority of PQS positions near the TSS, however, are correlated with decreased transcription. In our model, we account for this by proposing that downstream PQS in both strands spontaneously form intrastrand G-quadruplex structures that hinder the passage of RNAPII, effectively reducing the level of transcription. We propose that the RecQ helicases resolve intrastrand G-quadruplex structures, partially relieving the G-quadruplex mediated transcriptional repressive effect. This model is consistent with our analysis of BS and WS showing that down-regulated genes have an abundance of PQS in positions correlated with reduced expression. This model also has support from *in vitro* experiments performed by others studying individual genes with promoter-proximal PQS using stabilizing ligands [21, 22].

It is slightly more challenging to explain the correlation of upstream PQS and reduced transcription. Our data is the first to suggest a function role of upstream PQS, and no mechanism currently exists to explain our observations in this region. This is a topic that clearly requires further investigation.

Relevance to other RecQ helicases

It is intriguing to consider whether correlations of PQS abundance/location with transcriptional sensitivity to RecQ helicase loss observed in this study may similarly

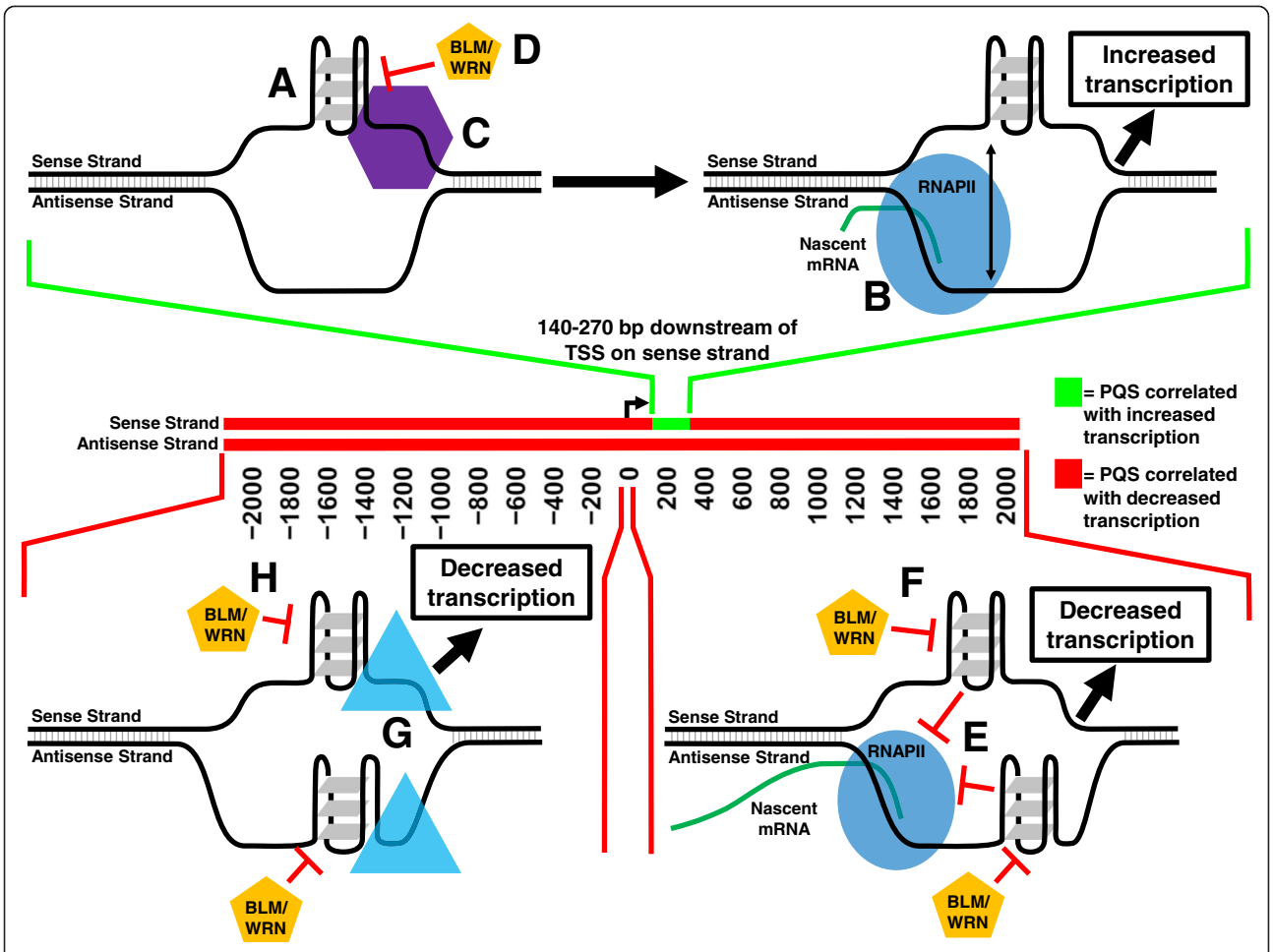

Fig. 6 A model to account for PQS transcriptional correlations. Model summarizes PQS position- and strand-dependent correlations with gene transcription determined from the present work and possible molecular mechanisms to explain these observations. An increase in transcription may be facilitated by the formation of intramolecular G-quadruplex structures between 140–270 bp downstream of the TSS on the sense strand (**a**), resulting in the release of the antisense strand from Watson-Crick base pairing and enabling easier access of the antisense strand to RNAPII, aiding transcription initiation (**b**). Formation of G-quadruplex structures in this region may also recruit G-quadruplex DNA-binding proteins (**c**) which may further stabilize G-quadruplex structures and enhance transcription. Resolution of G-quadruplex structures by BLM and WRN helicases (**d**) attenuates this transcriptional-activating effect. Transcriptional decrease mediated by PQS downstream of the TSS on both DNA strands may be facilitated by intramolecular G-quadruplex formation, resulting in a biophysical roadblock which prevents the passage of RNAPII and hinders transcript elongation (**e**). Resolution of G-quadruplex structures by BLM and WRN helicases in this region attenuates the transcriptional-repressing effect of the biophysical roadblock (**f**). Transcriptional decrease mediated by PQS upstream of the TSS may be recruitment of trans-acting factors that attenuate transcription (**g**). Resolution of G-quadruplex structures in this region attenuates this transcriptional repressing effect (**h**)

be found for other DNA helicases such as the RecQ4 helicase (deficient in Rothmund-Thomson photosensitivity- and cancer-associated syndrome), FANCJ helicase (deficient in Fanconi Anemia), and RecQ1 and RecQ5 helicases (members of the RecQ helicase family, but lacking evidence of a human phenotype). Of these helicases, BLM, WRN, RecQ4, and FANCJ reportedly have measurable affinity and helicase activity for G-quadruplex DNA *in vitro*, in addition to other diverse specificities and activities that are unique to each helicase. It is possible that similar patterns to those observed for BLM and WRN helicases in the present work may also be observed for RecQ4 and FANCJ helicases, although testing this experimentally is beyond the scope of the present work. There is less experimental evidence, however, that RecQ1 and RecQ5 have significant affinity and helicase activity for G-quadruplex DNA, although these helicases do have affinity and helicase activity for other DNA structures. It would therefore be less likely that significant correlation would be observed between PQS abundance/location and transcriptional change in RecQ1 and RecQ5-deficient cell lines.

Conclusions

We have used analysis of PQS in BS and WS in combination with genome-wide correlation of PQS motif

position with transcription using data from seven human cell lines to imply the functional roles of PQS motifs. Our studies reveal that PQS have position- and strand-dependent correlations with both increased and decreased transcription, and suggest that the RecQ helicases are functionally important in moderating PQS transcriptional effects. Particularly novel insights this these high-resolution analyses are that i) PQS abundance in BS and WS differentially-expressed genes varies strongly with both position and strand; ii) PQS motifs in BS and WS differentially expressed genes are significantly biased in composition; and iii) PQS in multiple human cell lines are associated with activated or repressed transcription in a strand- and position-dependent manner.

Availability of supporting data

Supporting data are included as additional files.

Additional files

Additional file 1: Genomic coordinates and motif features for all PQS. Reference coordinates from GRCh37 calculated using the Quadparser algorithm, and analysis of individual PQS sequence features including total length, loop lengths, number of G stacks, and base fractions of adenine, thymine, and cytosine. (ZIP 13239 kb)

Additional file 2: R code used for mapping genomic PQS to known TSS. Program written in R, used for mapping PQS motifs to known TSS. Takes GENCODE annotations in .csv format and Additional file 1 as input. (R 7 kb)

Additional file 3: Mapped coordinates of all PQS within 2 kbp of known TSS. Genomic coordinates and gene localization of all PQS within 2 kbp of known TSS. Each row contains data on a PQS motif and its host gene. PQS entries are repeated multiple times if a given motif is within 2 kbp of multiple TSS. Genes without any PQS motifs are represented as single rows with missing PQS motif data. (ZIP 11660 kb)

Additional file 4: Differentially expressed genes in Werner Syndrome. List of genes that are differentially expressed in human fibroblasts in Werner Syndrome. (CSV 116 kb)

Additional file 5: R code for PQS abundance analysis in BS and WS differentially-expressed genes. Program written in R for analyzing differences in PQS abundance between BS/WS differentially-expressed genes and the rest of the genes in the genome. Takes Additional files 3 and 5 as input. (R 50 kb)

Additional file 6: R code for GENCODE annotation processing protocol for epigenetic modeling of gene expression method. Program written in R for filtering low quality transcript annotations, assigning unique TSS numberings, and unique gene numberings. Takes GENCODE annotations in .gtf format as input. (R 3 kb)

Additional file 7: R code for CAGE data processing for epigenetic modeling of gene expression method. Program written in R that takes CAGE data in .gtf format and the output of Additional file 6 as input, merges rows with GENCODE annotation table (output of Additional file 6), calculates aggregate expression of all transcripts originating from a given TSS, and identifies TSS for each gene with the highest expression. (R 4 kb)

Additional file 8: R code for mapping PQS to TSS for epigenetic modeling of gene expression method. Program written in R that identifies all PQS within 2 kbp of TSS, calculates relative distance between PQS and TSS, identifies whether PQS are located on sense or antisense strand. Takes Additional file 1 and the output of Additional file 8 as input. (R 13 kb)

Additional file 9: R code for extracting DNAse 1 genomic track signal near TSS for epigenetic modeling of gene expression method. Program written in R that calculates DNAse track signal in 100 bp bins spanning 1 kbp upstream and downstream of the TSS. Takes .bigwig file of DNAse 1 hypersesitivity epigenomic track and output of Additional file 8 as input. (R 4 kb)

Additional file 10: R code for extracting histone modification track signal near TSS for epigenetic modeling of gene expression method. Program written in R that calculates histone modification track signal in 100 bp bins spanning 1 kbp upstream and downstream of the TSS. Takes multiple .bigwig files of histone modification epigenomic tracks and the output of Additional file 9 as input. (R 5 kb)

Additional file 11: R code for determining optimal pseudocount and bin for epigenetic modeling of gene expression method. Program written in R that identifies optimal pseudocount and bin for calculating correlations of epigenomic track signals with gene expression. Takes output of Additional file 10 as input. (R 5 kb)

Additional file 12: GM12878 epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 2493 kb)

Additional file 13: H1-hESC epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 6250 kb)

Additional file 14: HeLa-S3 epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 4263 kb)

Additional file 15: HepG2 epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 4911 kb)

Additional file 16: HUVEC epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 3089 kb)

Additional file 17: K562 epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 4722 kb)

Additional file 18: NHEK epigenetic modeling of gene expression dataset. Expression modeling dataset used for epigenetic modeling of gene expression analysis. Each row corresponds to a TSS and contains information on positions of PQS within 2 kbp of the TSS as well as "bestbin" epigenomic track signals for DNase hypersensitivity and histone

modifications. This file is an output of Additional file 11 and is taken as input by Additional file 19. (CSV 2268 kb)

Additional file 19: R code for implementing epigenetic modeling of gene expression and analyzing PQS transcriptional effects. Program written in R that takes Additional files 12, 13, 14, 15, 16, 17, 18 as inputs and trains a Bayesian linear regression model upon epigenomic data to predict gene expression levels. Models are then applied to gene sets not used for training, and the subsequent sorting of modeling prediction errors by PQS position is used to calculate PQS transcriptional effects. Also conducts statistical bootstrapping simulations to estimate the FDR for this analysis. (R 44 kb)

Additional file 20: R code for PQS feature analysis in BS and WS via ANOVA. Program written in R that analyses promoter-proximal PQS motif length, loop lengths, and base fraction in differentially expressed genes in BS and WS and compares the result to the genomic distribution. Takes Additional file 1 an Additional file 4 as input. (R 78 kb)

Additional file 21: R code for PQS motif analysis in BS and WS differentially-expressed genes via SOM method. Program written in R for conducting self-organizing map classification of PQS motifs and analyzing enrichment of certain types of motifs in genes differentially expressed in BS and WS. (R 11 kb)

Additional file 22: PQS motif classifications from SOM method. PQS motif classifications from self-organizing map analysis. Number of nearest multidimensional cluster is given as the last column in the dataset. (CSV 8064 kb)

Additional file 23: Supplemental Figures S1 to S3 and Tables S1 to S7 with corresponding legends. (PDF 677 kb)

Competing interests
JS and LJM have no conflicts of interest to declare regarding the content of this work.

Authors' contributions
JS conceived the study, conducted the analyses, and drafted the manuscript. LJM consulted with study design and edited the manuscript. Both authors read and approved the final manuscript.

Authors' information
Not applicable.

Acknowledgements
We wish to thank Dr. Justin Peters for his technical advice relating to the use of the Mayo Research Computing Facility Linux cluster.

Author details
[1]Mayo Clinic Medical Scientist Training Program, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55905, USA. [2]Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55905, USA.

Received: 23 January 2015 Accepted: 23 September 2015
Published online: 08 October 2015

References
1. Epstein CJ, Martin GM, Schultz AL, Motulsky AG. Werner's syndrome a review of its symptomatology, natural history, pathologic features, genetics and relationship to the natural aging process. Medicine (Baltimore). 1966;45:177–221.
2. Lauper JM, Krause A, Vaughan TL, Monnat Jr RJ. Spectrum and risk of neoplasia in Werner syndrome: a systematic review. PLoS One. 2013;8, e59709.
3. Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, et al. Positional cloning of the Werner's syndrome gene. Science. 1996;272:258–62.
4. Bloom D. Congenital telangiectatic erythema resembling lupus erythematosus in dwarfs; probably a syndrome entity. AMA Am J Dis Child. 1954;88:754–8.
5. German J. Bloom syndrome: a mendelian prototype of somatic mutational disease. Medicine (Baltimore). 1993;72:393–406.
6. German J, Bloom D, Passarge E. Bloom's syndrome. VII. Progress report for 1978. Clin Genet. 1979;15:361–7.
7. Sun H, Karow JK, Hickson ID, Maizels N. The Bloom's syndrome helicase unwinds G4 DNA. J Biol Chem. 1998;273:27587–92.
8. Fry M, Loeb LA. Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)n. J Biol Chem. 1999;274:12797–802.
9. Harmon FG, DiGate RJ, Kowalczykowski SC. RecQ helicase and topoisomerase III comprise a novel DNA strand passage function: a conserved mechanism for control of DNA recombination. Mol Cell. 1999;3:611–20.
10. Ahn B, Harrigan JA, Indig FE, Wilson 3rd DM, Bohr VA. Regulation of WRN helicase activity in human base excision repair. J Biol Chem. 2004;279:53465–74.
11. Opresko PL, Mason PA, Podell ER, Lei M, Hickson ID, Cech TR, et al. POT1 stimulates RecQ helicases WRN and BLM to unwind telomeric DNA substrates. J Biol Chem. 2005;280:32069–80.
12. Crabbe L, Jauch A, Naeger C, Holtgreve-Grez H, Karlseder J. Telomere dysfunction as a cause of genomic instability in Werner syndrome. Febs Journal. 2007;274:7–7.
13. Crabbe L, Verdun RE, Haggblom CI, Karlseder J. Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. Science. 2004;306:1951–3.
14. Johnson JE, Cao KJ, Ryvkin P, Wang LS, Johnson FB. Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. Nucleic Acids Res. 2010;38:1114–22.
15. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33:2908–16.
16. Nguyen GH, Tang W, Robles AI, Beyer RP, Gray LT, Welsh JA, et al. Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs. Proc Natl Acad Sci U S A. 2014;111:9905–10.
17. Cheung HH, Liu X, Canterel-Thouennon L, Li L, Edmonson C, Rennert OM. Telomerase protects werner syndrome lineage-specific stem cells from premature aging. Stem cell reports. 2014;2:534–46.
18. Du Z, Zhao YQ, Li N. Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription (vol 18, pg 233, 2008). Genome Res. 2008;18:516–6.
19. Dong XJ, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13.
20. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982;43:59–69.
21. Cogoi S, Xodo LE. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Res. 2006;34:2536–49.
22. Lemarteleur T, Gomez D, Paterski R, Mandine E, Mailliet P, Riou JF. Stabilization of the c-myc gene promoter quadruplex by specific ligands' inhibitors of telomerase. Biochem Biophys Res Commun. 2004;323:802–8.