

Noncoding cancer risk variants at 8q24: predicted regulatory effects on *MYC*

Christopher Zheng¹, Michael S. Cuoco²

¹ Ridge High School, Basking Ridge, NJ

² University of California, San Diego, La Jolla, CA

KEYWORDS: cancer, genomics, enhancers, *MYC*, variants, artificial intelligence, population

OVERVIEW: Noncoding cancer-associated variants at the 8q24 locus may influence *MYC*, but the mechanisms through which these variants do this remain unclear. This study utilized AlphaGenome to investigate whether selected 8q24 cancer-associated variants demonstrate evidence of regulatory activity near *MYC* and whether risk alleles alter predicted chromatin accessibility or *MYC* expression.

Abstract

Noncoding cancer-associated variants at the 8q24 locus may influence *MYC*, but the mechanisms through which these variants do this remain unclear. We hypothesized that selected 8q24 cancer-associated variants demonstrate evidence of regulatory activity near *MYC* and that risk alleles alter predicted chromatin accessibility. A panel of 12 variants from multiple GWAS of colon cancer, prostate cancer, and breast cancer was analyzed using ENCODE cCRE annotations, and then AlphaGenome was used to predict chromatin accessibility and contact maps, in silico mutagenesis, and gnomAD was used to plot ancestry differences. rs6983267 showed the strongest convergence of regulatory evidence, including being a proximal enhancer and meeting nomination requirements for AlphaGenome's chromatin analysis. There was no clear relationship between increased regulatory effect and ancestry-frequency disparity due to the small variant panel. These findings suggest that rs6983267 shows strong predicted evidence to modulate the activity of a positive *MYC* enhancer. rs6983267 is supported as a plausible functional candidate, prioritizing it for future testing in prostate cell lines, although linkage disequilibrium prevents causal identification from GWAS alone. However, AlphaGenome's results are predictions: the predictions generated in this study by AlphaGenome require experimental validation before they can be confirmed.

INTRODUCTION

Cis-regulatory elements (CREs) are noncoding DNA sequences that regulate the expression of nearby (proximal) or distant (distal) genes. This includes enhancers and silencers, which increase and reduce gene expression respectively (1). Enhancers are a type of CRE that can regulate transcription even when physically far away from their target gene. Enhancers often physically interact with a gene's promoter through DNA bending, allowing enhancers to regulate transcription across vast distances (2, 3). Active CREs are often associated with epigenetic features such as increased chromatin accessibility, certain histone marks, transcription factor binding, and physical contact with promoters (3).

Genome-Wide Association Studies (GWAS) scan the genome across a population, comparing allele counts with expressed traits to determine which loci may be associated with those traits (4). Interestingly, genetic-trait associations often fall within noncoding regulatory regions, suggesting such loci have a bona fide role in the biological manifestation of the trait (5). However, GWAS are complicated by linkage disequilibrium (LD). Since genetic variants are often inherited in blocks, certain single nucleotide polymorphisms (SNPs) are inherited together.

This creates tag-SNPs, which may be identified by GWAS as significantly associated with a trait, but could not be doing anything biologically. GWAS may identify a tag-SNP which is associated with a phenotype, but other inherited SNPs in the block could be causing it (4).

Gene deserts are large regions in the genome containing little or no protein-coding genes. 8q24 (chr8: 116,500,001 - 145,138,636) is a very unusual example of a gene desert. It contains almost no protein-coding genes, but some of the strongest cancer-associated genetic signals reported by GWAS (6). Indeed, 8q24 lies near *MYC* TSS (chr8:127,735,433), a gene commonly associated with cancer (7, 8). *MYC* is considered a likely regulatory target of elements in 8q24 because of observed long-range physical contact between certain 8q24 risk loci and *MYC* (9).

Past studies have already identified many cancer-associated variants in 8q24 through GWAS, have shown that certain 8q24 regions have enhancer-like activity, and demonstrated long-range physical interactions between 8q24 risk loci and *MYC* (6, 9). However, these studies do not prove which GWAS SNPs are functional. LD also plays a large role in this, making GWAS data limited for determining functional regions (4). Prior studies have also not found a consistent relationship between risk-allele status and *MYC* expression (6). We hypothesized that risk variants in 8q24 would demonstrate regulatory activity on *MYC*, with risk alleles altering chromatin accessibility. We evaluated variants for regulatory evidence and determined if variants with strong predicted regulatory effects demonstrated larger ancestry-related frequency differences. Among the 12 rsIDs, only rs6983267 met nomination criteria, displaying strong evidence of predicted regulatory effect and emerging as a candidate for further testing.

RESULTS

Four variants overlapped ENCODE cCREs

rsID	Cancer	Risk	Ref.	Alt.	<i>MYC</i> TSS distance (bp)	P-value
rs16901979	Prostate	A	C	A	622762	3×10^{-14}
rs6983561	Prostate	C	G	C	640798	7.9×10^{-19}
rs1016343	Prostate	C	C	T	654381	1×10^{-7}
rs10086908	Prostate	T	T	C	735741	7.9×10^{-8}
rs1447295	Prostate	A	A	C	262640	2×10^{-19}
rs4242382	Prostate	A	A	G	230105	3×10^{-19}
rs10090154	Prostate	T	T	C	215541	8.8×10^{-13}
rs7837688	Prostate	T	T	G	208318	1.2×10^{-25}
rs6983267	Prostate/ Colorectal	G	G	T	334373	9×10^{-13} (Prostate); 1×10^{-14} (Colorectal)

rsID	Cancer	Risk	Ref.	Alt.	MYC TSS distance (bp)	P-value
rs10505477	Prostate/Colorectal	A	A	G	340235	3×10^{-11} (Colorectal)
rs16902094	Prostate	G	A	G	427332	6×10^{-15}
rs13281615	Breast	G	A	G	392061	5×10^{-12}

Table 1. Cancer-associated variants included in the analysis. The table lists rsID, cancer association, risk, reference, and alternate alleles, distance from the *MYC* TSS, and reported P-value. P-values were obtained from Grisanzio and Freedman (2010) where available; additional values were taken from Haiman et al. (2007), Al Olama et al. (2009), and Takata et al. (2010). Risk, risk allele; Ref., reference allele; Alt., alternate allele.

To understand if any of the twelve variants (rsIDs, Table 1) fall within regions previously predicted to be in cCREs, we inspected the ENCODE cCREs at each of the 12 loci in the UCSC genome browser. ENCODE defines predicted cCRE status in a given biosample through a combined scoring method of DNase, ATAC, H3K4me3, H3K27ac, and CTCF signals (see details in Materials and Methods) (10). We identified four of twelve rsIDs that fell within an ENCODE cCRE. rs1016343 and rs6983267 were both within ENCODE-predicted proximal enhancers, while rs10086908 and rs10505477 fell within ENCODE-predicted distal enhancers. All four of these rsIDs were analyzed with a prostate cell-line context. However, prostate and tissue-specific track predictions were often not the same as the main ENCODE classification (Table 2). These findings serve as preliminary evidence that four of the twelve GWAS loci we collected fall within enhancer regions of *MYC*, independent of AlphaGenome predictions.

rsID	cancer	in cCRE?	Closest cCRE ID	cCRE Class	PC-3 Class	Predicted Effect
rs16901979	Prostate	FALSE	EH38E3858151	—	—	FALSE
rs6983561	Prostate	FALSE	EH38E2665673	—	—	FALSE
rs1016343	Prostate	TRUE	EH38E3858121	Proximal Enhancer	Proximal Enhancer	TRUE
rs10086908	Prostate	TRUE	EH38E3858034	Distal Enhancer	Low Chromatin Accessibility	TRUE
rs1447295	Prostate	FALSE	EH38E2665962	—	—	FALSE
rs4242382	Prostate	FALSE	EH38E2665989	—	—	FALSE
rs10090154	Prostate	FALSE	EH38E4482920	—	—	FALSE

rsID	cancer	in cCRE?	Closest cCRE ID	cCRE Class	PC-3 Class	Predicted Effect
rs7837688	Prostate	FALSE	EH38E3858546	—	—	FALSE
rs6983267	Prostate/Colorectal	TRUE	EH38E3858447	Proximal Enhancer	Proximal Enhancer	TRUE
rs10505477	Prostate/Colorectal	TRUE	EH38E2665913	Distal Enhancer	Low Chromatin Accessibility	TRUE
rs16902094	Prostate	FALSE	EH38E3858377	—	—	FALSE
rs13281615	Breast	FALSE	EH38E3858405	—	—	FALSE

Table 2. Manual ENCODE cCRE annotations for the 12 variants. cCRE, candidate cis-regulatory element; PC-3, prostate cancer cell line. Dashes indicate that the variant did not overlap a cCRE. The predicted effect indicates whether the rsID met previous requirements (see Materials and Methods) for cCRE classification, and does not indicate final predicted regulatory effect.

rs6983267 chromatin accessibility stood out

To compare the chromatin accessibility across the *MYC*-8q24 locus in PC-3, we analyzed the ATAC signal across the window, focusing specifically on the rs6983267 candidate enhancer. Predicted PC-3 chromatin accessibility (ATAC signal) was high near the *MYC* promoter and rs6983267 relative to the genomic background. Notably, the *MYC* promoter demonstrated 4-5 fold higher predicted ATAC signal than the rs6983267 region, suggesting the *MYC* promoter is more accessible to transcriptional machinery in PC-3 than the rs6983267 region (Figure 1a, b). The mean ATAC score within 500 bp of the *MYC* TSS was predicted to be 1.81, while it was predicted at 0.425 for the rs6983267 region. While the predicted ATAC signal around the *MYC* TSS was much larger than at rs6983267, the ATAC signal from rs6983267 still stood out above the genomic background, demonstrating that the area around rs6983267 was predicted to have relatively higher local chromatin accessibility in PC-3. Notably, this is an established phenomenon—promoters tend to be more accessible than enhancers (3). The increased ATAC signal around rs6983267 could be evidence of predicted regulatory effect. Further tests would strengthen this conclusion.

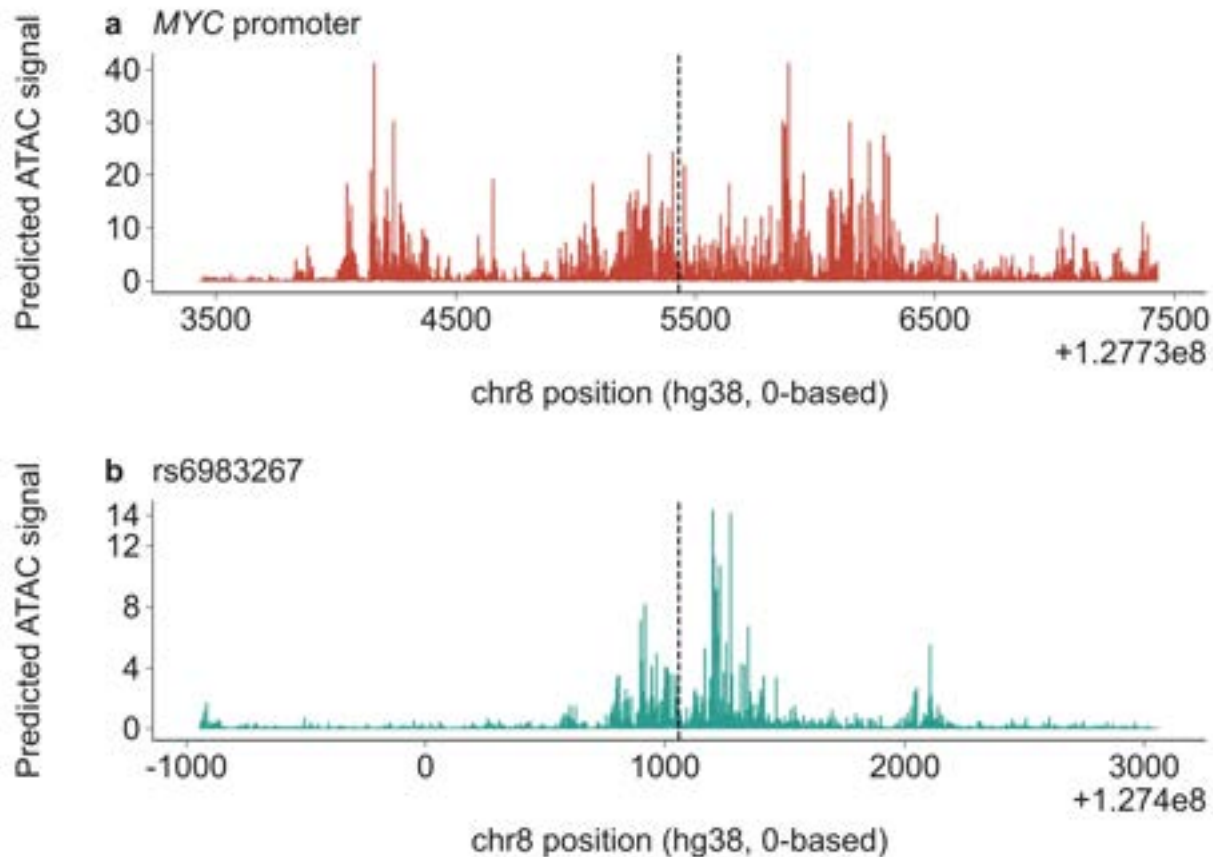


Figure 1. PC-3 predicted ATAC signal at the *MYC* TSS and rs6983267 candidate enhancer.

AlphaGenome-predicted ATAC signal at 1-bp resolution in PC-3 prostate cancer cell line, signaling chromatin accessibility at each location. A dashed line indicates the location of the *MYC* TSS (a) and rs6983267 (b). (a) *MYC* promoter (TSS $\pm$ 2 kb). (b) candidate enhancer around rs6983267 ($\pm$ 2 kb). This figure directly compares the predicted ATAC near the *MYC* TSS and the region around rs6983267.

Specific base-pair substitutions changed local ATAC near rs6983267

To determine which nucleotide positions are most sensitive to sequence change near rs6983267, we performed an in-silico mutagenesis around the region. ISM revealed that base identities at certain positions, for example at position 1070 with base T, have a much larger effect on the predicted ATAC than others in the region. Base A around position 1160 had a much larger local predicted ATAC effect than other base pairs at the same position (Figure 2a). The predicted *MYC* RNA effect had few base pairs that stood out aside from important areas like base pair T at around 940 (Figure 2b). The predicted *MYC* RNA effect was more affected by base pair changes from around position 1000 to position 1200. Around rs6983267, only very specific base pair changes were predicted to strongly affect the local predicted ATAC signal. In contrast, predicted effects on *MYC* RNA expression were smaller in magnitude and distributed across the ISM window. This evidence points to the region around rs6983267 being functionally sensitive, containing specific nucleotide positions that are especially important for local

chromatin accessibility, even if RNA effect is smaller in magnitude, supporting the idea that the rs6983267 region has regulatory potential.

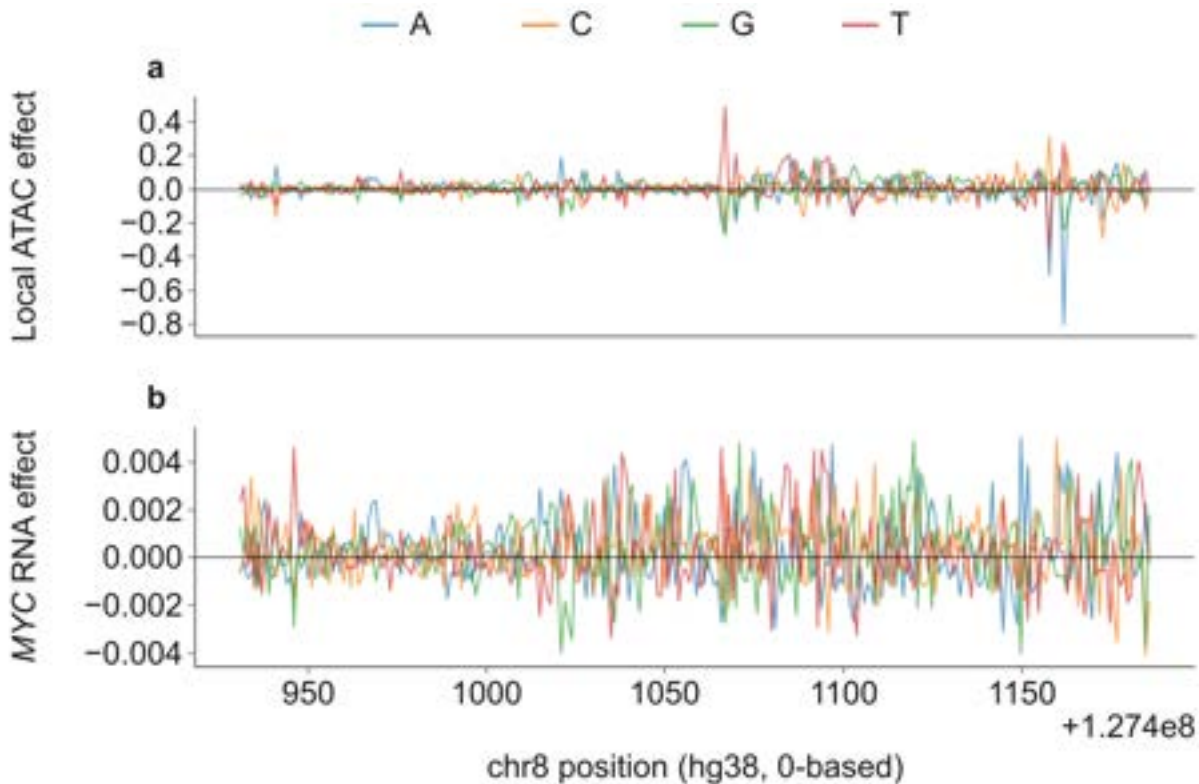


Figure 2. Delta ATAC and delta RNA across the rs6983267 ISM window. Predicted local ATAC effects (a) are shown above predicted *MYC* RNA log-fold change effects (b) across the 256 bp window centered on the candidate enhancer. Different colors indicate a different base, with large spikes demonstrating larger relative ATAC or RNA effect. This window analyzes the local ATAC and local *MYC* RNA effect near rs6983267.

rs6983267 was the only nominated likely enhancer

To prioritize SNP loci as likely *MYC* enhancers, we analyzed predicted contact maps and reference ATAC signal at each SNP to nominate enhancers which met the criteria focusing around rs6983267 (Figure 3). Nominated enhancers required a positive contact score with the *MYC* promoter, determined at 2,048 bp resolution and expressed as log-fold over the distance-expected contact, along with a reference chromatin accessibility, converted to a percentile against tiled background windows of the same width, of 90% or above. Only one SNP matched the criteria for being nominated, which was rs6983267 with an ATAC in the 98th percentile, and a *MYC* contact score of 0.014, passing both criteria. The other SNPs, while close to the *MYC* promoter, often lacked the contact requirement. Many SNPs were positioned near spots with high ATAC and demonstrated meaningful accessibility regulation. Positive predicted contact with *MYC* was rare, with only rs6983267 and rs10086908 having positive contact scores. While rs6983267 was the only SNP that passed nomination, rs1016343 and rs16902094 also had ATAC percentiles of 96% and 95% respectively. However, they both

received negative contact scores. Only rs6983267 passed the nomination test, so it was the most likely to be a predicted *MYC* enhancer based on the criteria.

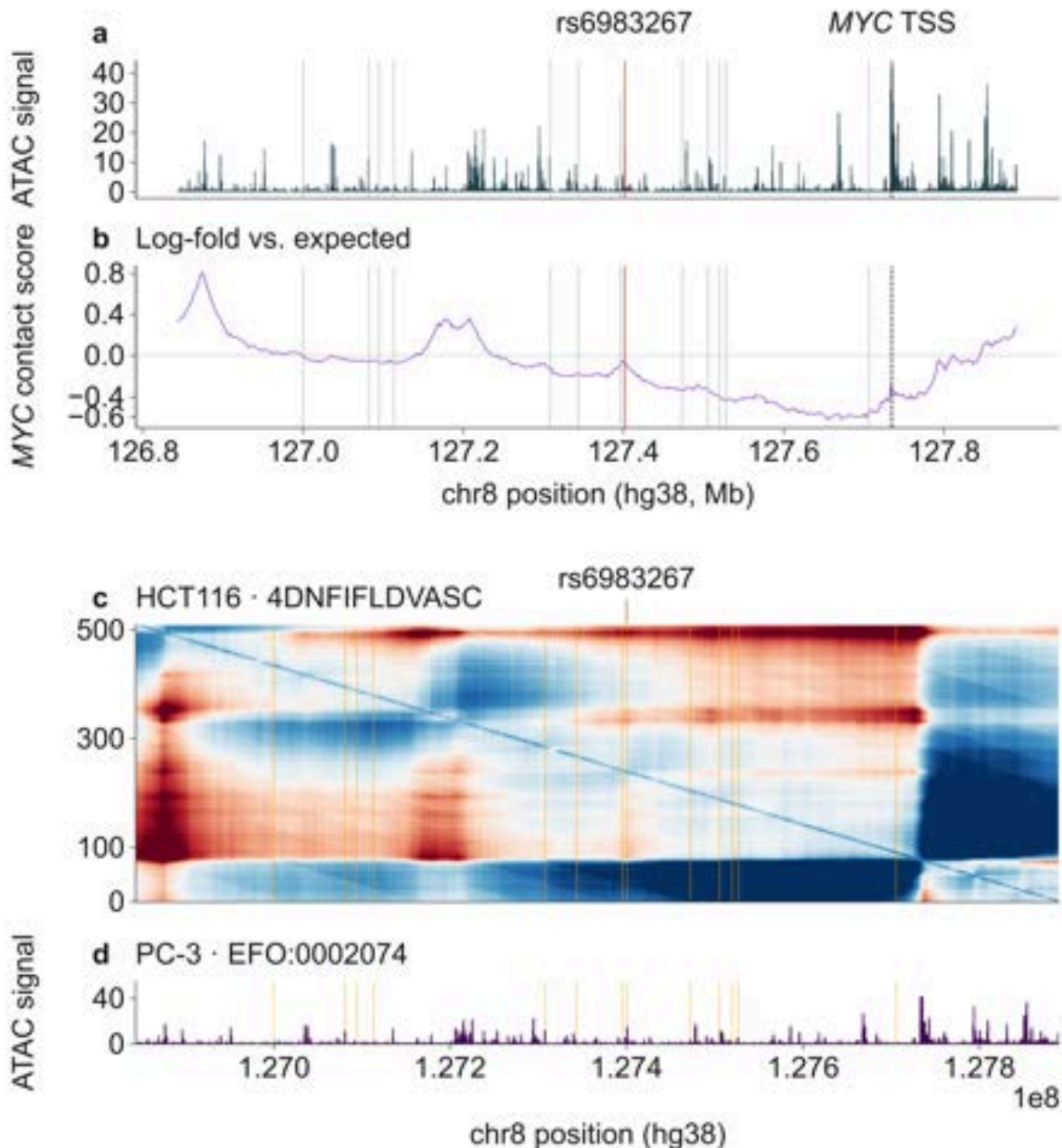


Figure 3. Predicted accessibility and *MYC* contact across 8q24. (a) Reference PC-3 ATAC signal and (b) predicted contact with the *MYC* promoter across the locus. (c) HCT116 contact map across the 8q24-*MYC* window. (d) Reference PC-3 ATAC signal beneath the contact map. The different peaks demonstrate elevated contact with *MYC* promoter region (b), and darker regions indicate greater contact levels (c). Each separate panel has rs6983267 marked in the same spot for quick comparison.

Risk and non-risk accessibility differed most for rs1016343 and rs6983267

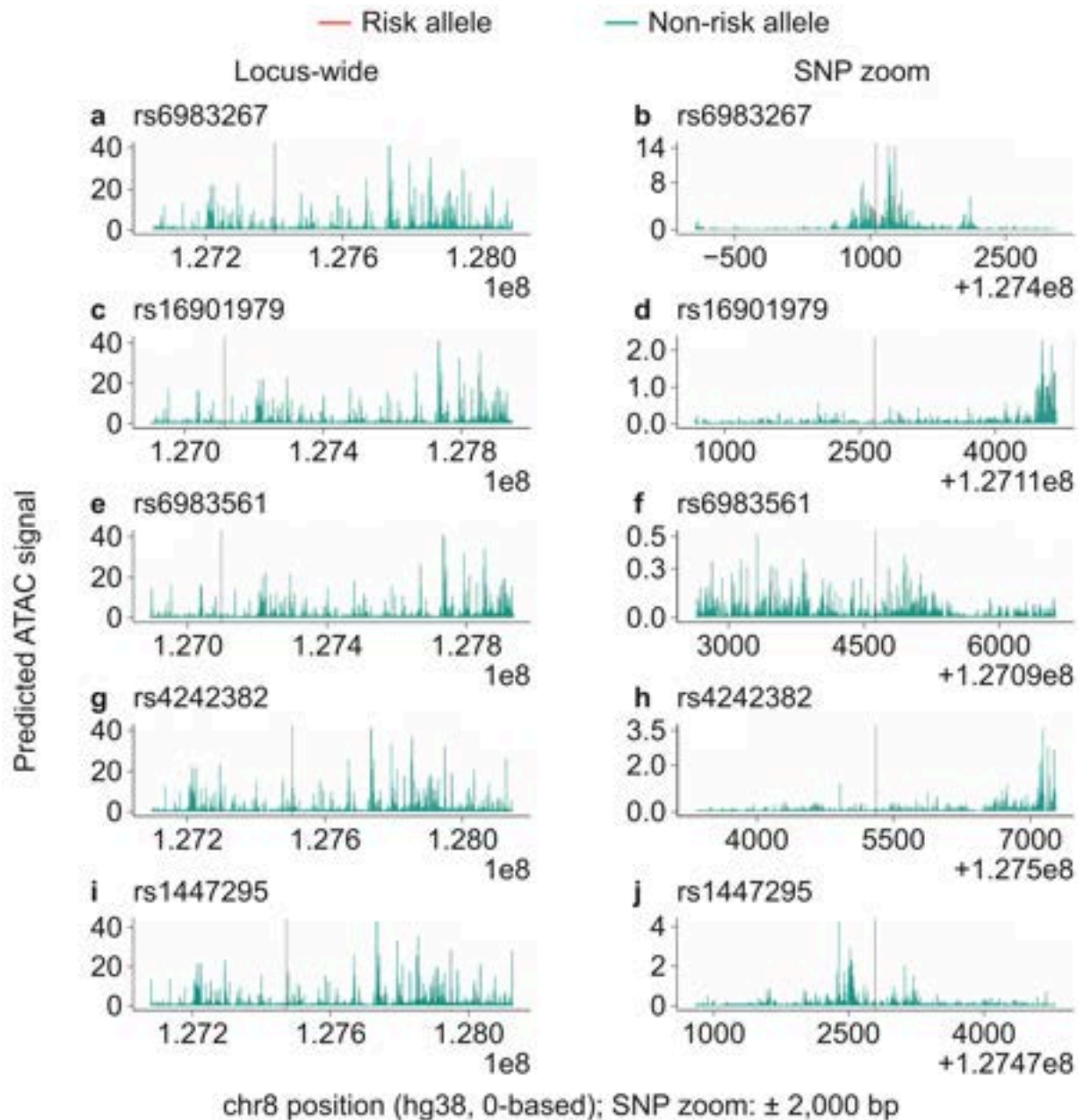


Figure 4. Risk-allele versus non-risk-allele predicted ATAC tracks. Locus-wide and SNP-zoom views are shown for rs6983267 and the four SNPs with the largest ancestry-frequency disparities. Panels a and b: rs6983267; c and d: rs16901979; e and f: rs6983561; g and h: rs4242382; i and j: rs1447295. Panels a, c, e, g, and i show locus-wide views; b, d, f, h, and j show SNP-zoom views (± 2 kb). These panels make for easy comparison of ATAC level differences between risk and non-risk allele, also helping to determine the scale of the change with both local and locus-wide windows.

To predict risk allele regulatory effects, we compared the risk and non-risk chromatin accessibility scores to determine differences (see methods for details). Predicted accessibility

scores were defined as the mean AlphaGenome-predicted PC-3 ATAC signal across a 1001-bp window centered on each SNP. SNP delta mean was then calculated by subtracting the non-risk accessibility score from the risk accessibility score. rs6983267 demonstrated a large ATAC difference between risk-allele and non-risk allele in the visualization (Figure 4a, b). rs6983561 had a large predicted ATAC difference at its position between risk and non-risk alleles, even though it had the smallest SNP delta mean effect of -1.7×10^{-5} (Figure 4e, f). While rs6983267 was the only nominated allele, rs1016343 had the largest SNP delta mean of +0.0215. rs6983267 placed below it with an SNP delta mean of +0.0160. The nominated SNP, rs6983267, had the second largest SNP delta mean, furthering its likelihood as a predicted enhancer. Other regions like rs1016343 and rs6983561 were notable as well. rs1016343 was notable due to its high SNP delta mean, and rs6983561 was notable since it showed a visible local ATAC difference despite having a very low delta mean. These results further support the idea that the rs6983267 region has regulatory potential, since the higher ranking on the SNP delta mean list represented greater overall chromatin accessibility change than most other rsIDs.

No clear relationship emerged between population frequency and predicted regulatory effect

To determine if there is correlation between predicted risk allele regulatory effect and population frequency, we analyzed population frequency through gnomAD (11). When comparing ancestries, rs6983267, rs6983561, rs16901979, rs4242382, and rs1447295 had large differences in frequencies. These SNPs had a much larger AFR frequency than other populations, especially NFE (Figure 5a). Several prostate cancer-associated variants had higher risk-allele frequencies in AFR than NFE, with rs6983561 and rs16901979 showing the largest AFR to NFE ratios of 14.14 and 13.49 respectively. The leading rsID in Table 3, rs1016343, did not have a large population frequency difference. rs6983267, however, did have a population frequency difference and was very high on Table 3 as well, with a large predicted delta mean score. rs6983267 was also the only nominated variant based on previous criteria. Although rs6983267 placed high in Table 3 and had a much higher AFR risk allele frequency than other ancestries, rs1016343 placed high in Table 3 but did not have a clear uneven frequency distribution. These results showed that certain rsIDs were higher in frequency among certain ancestries than others, but there were not enough data points to determine whether risk allele frequency across ancestry groups is correlated with regulatory effect.

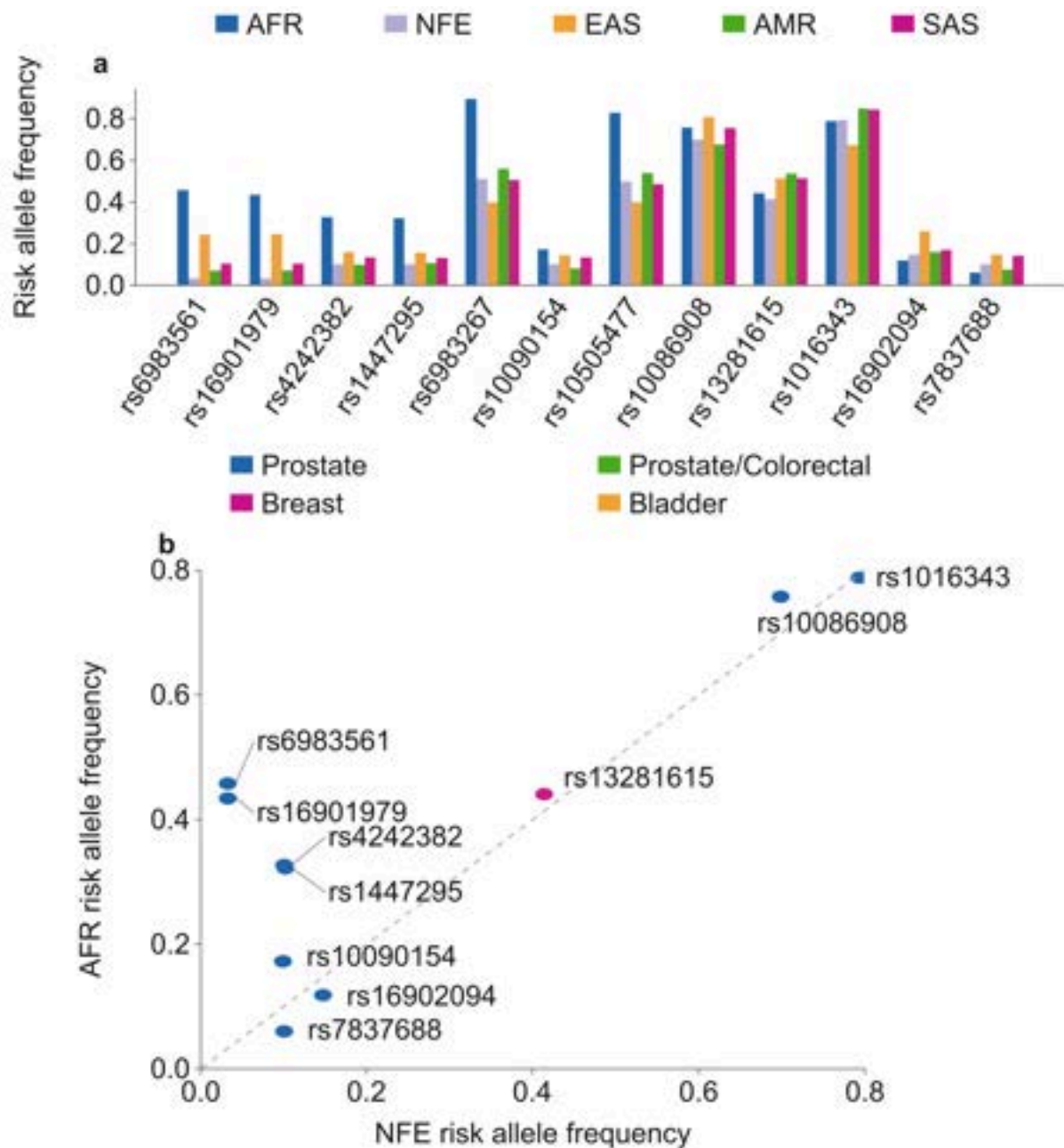


Figure 5. Risk-allele frequency across ancestry groups. (a) Risk-allele frequency for the variant panel across AFR, NFE, EAS, AMR, and SAS populations. Different colors represent different populations, so that differences in frequency are easy to spot. (b) AFR versus NFE risk-allele frequency for the variant panel. Different colors represent different cancers, and rsIDs farther from the mid-line, which signifies a ratio of 1, demonstrate a larger difference between AFR and NFE population frequencies. rs6983267 and rs10505477 are absent due to their AFR frequency exceeding 0.8, outside of the chart's risk allele frequency window.

rsID	snp_delta_mean	ref_atac_pctile	hic_contact_myc
rs1016343	0.021468	96.032505	-0.009033
rs6983267	0.015958	98.32696	0.014221
rs10505477	-0.002498	66.013384	-0.018555
rs1447295	0.001329	81.070746	-0.267578
rs7837688	-0.000797	55.497132	-0.330078
rs16901979	0.00039	16.347992	-0.005066
rs10086908	-0.000311	37.523901	0.046875
rs10090154	0.000117	20.172084	-0.310547
rs13281615	-0.000114	42.112811	-0.128906
rs4242382	-0.000086	50.43021	-0.245117
rs16902094	0.000027	95.028681	-0.077637
rs6983561	-0.000017	26.481836	-0.007263

Table 3. Full-panel AlphaGenome summary. The table lists SNP delta mean, reference accessibility percentile, and predicted contact with *MYC* for each rsID.

DISCUSSION

We analyzed a panel of 12 variants from previous GWAS and nominated rs6983267 to have the highest evidence of enhancer regulatory effects on *MYC* expression. We manually viewed SNPs in the UCSC browser, predicting rs1016343, rs6983267, rs10086908, and rs10505477 to have regulatory effect. We compared chromatin accessibility across the window with rs6983267 as a candidate enhancer, determining that there is an increased ATAC near it. We performed a targeted ISM near rs6983267, demonstrating that only specific base pair changes had a large effect on ATAC, but a broad area of changes had an effect on *MYC* RNA. rs6983267 was then nominated due to meeting the criteria previously set, and we compared the ATAC differences between risk and non-risk alleles for all rsIDs, with rs1016343 having the largest mean difference. Ancestry data was compared, but there was not enough data to determine whether population frequency data and regulatory effect was correlated. However, there was enough data to determine that rs6983267 had a large population frequency difference between AFR and other ancestries. This difference means that rs6983267 effects will disproportionately affect AFR populations compared with NFE, EAS, AMR, and SAS populations.

Our findings suggest that GWAS signal from rs6983267 is partly explained by modulated enhancer activity because of its nomination status and predicted chromatin accessibility differences. This partially supports our hypothesis: only rs6983267 was predicted to demonstrate evidence of regulatory activity near *MYC*, altering chromatin accessibility.

rs6983267 was initially one of the four rsIDs predicted to have regulatory effect due to its ENCODE classification as proximal enhancer. rs6983267 was the only nominated SNP due to meeting both the ATAC requirement and the contact requirement. During rs6983267's ISM analysis, certain base-pair substitutions had a much larger predicted effect on local ATAC signal than others, suggesting that these substitutions may be predicted to contribute to regulatory activity. However, the effect on *MYC* RNA between base pairs was smaller in magnitude. The region surrounding rs6983267 also showed a relatively high predicted ATAC signal, demonstrating potential regulatory activity. rs6983267 also scored high in predicted SNP delta mean, showing that the difference between risk and non-risk allele for chromatin accessibility was larger than most other SNPs. In the population frequency analysis, while it may not be correlated with regulatory effect, rs6983267 demonstrated a much larger frequency in AFR populations. These findings support rs6983267 as a predicted regulatory element that may influence enhancer activity near *MYC*. The findings of this study are consistent with previous evidence that the region containing rs6983267 exhibits enhancer activity. rs6983267 has been previously considered to demonstrate regulatory effect, along with physically forming long range interactions with the *MYC* promoter (6, 9).

Although rs6983267 showed the strongest evidence of regulatory activity in the study, it is still unclear how the GWAS signal is associated with cancer risk. It's possible that GWAS association is due to linkage disequilibrium, with rs6983267 acting as a tag-SNP with nearby variants, possibly suggesting that rs6983267 may not be the functional variant (4), with another region making the biological impact. HCT116 colorectal contact predictions were used alongside PC-3 prostate accessibility predictions due to PC-3 contact predictions being unavailable, possibly resulting in a mismatch between the two predictions. This entire study was done digitally, so most results are predictive and data in this study generated by AlphaGenome is not experimentally validated (12). Therefore, future studies are required to validate if these variants have bona fide regulatory activity of *MYC* in the proper cell context, potentially through physical tests determining contact between risk regions and the *MYC* TSS. We hope that this study can provide valuable resources for future experimental groups to nominate variants for testing.

MATERIALS AND METHODS

Variant panel Assembly

A panel of variants was assembled from published GWAS labeled by rsIDs on hg38 coordinates by Grisanzio and Freedman (2010) (6, 13). Variants were selected from the summary of association results, removing those that did not fit into the 1 Mb AlphaGenome window, centered between the *MYC* TSS and the farthest rsID possible, along with removing variants outside the study's focus on prostate, colorectal, and breast cancer. This list of variants included rs6983267, which has been previously established with prostate/colorectal cancer association as a control (9). Additional P-values for variants not summarized by Grisanzio and Freedman were taken from primary studies (14-16).

Manual regulatory annotation via UCSC Genome Browser

Each variant locus was analyzed for regulatory annotations using the UCSC Genome Browser, by inspecting the ENCODE candidate cis-regulatory elements (cCREs) at the hg38 location of the locus (10, 17). The UCSC Genome Browser session used is available at: https://genome.ucsc.edu/s/chris_419/Annotations%20v4. If variants fell within a promoter-like or

enhancer-like cCRE, the predicted effect would be 'yes'. If it overlaps a CTCF-only or weakly-supported element, the predicted effect would be 'uncertain'. If it did not overlap any cCRE, the predicted effect would be 'no'. Only the ENCODE cCRE track was used for these predictions (10). Each variant was entered into the browser and recorded the closest cCRE ID, along with whether the variant fell within the cCRE. This allowed predictions to be made based on the cCRE class alone.

Resolving reference and alternate alleles

Certain rsIDs are multiallelic, meaning that more than two different bases have been observed at that position and leaving more than just a reference and alternate allele. Therefore, instead of taking values from the source table directly, the Ensembl database was referenced for each rsID to determine its current hg38 position and allele information (18). This allowed for high-resolution inputs and multi-allelic rsIDs being fixed.

AlphaGenome and window design

AlphaGenome is a state of the art sequence-to-function model developed by Google DeepMind which, given up to 1 Mb of DNA sequence, predicts chromatin accessibility, gene expression, and DNA contact maps across multiple cellular contexts. AlphaGenome is capable of interpreting long stretches of data without a resolution tradeoff, making it ideal for an analysis between 8q24 candidate enhancers and *MYC*, due to their distance (12). DNA sequence inputs can also be manipulated by the user, allowing for predictions for regulatory effects of mutated input sequences. This makes AlphaGenome a great tool for predicting the regulatory effects of the aforementioned risk alleles. Since AlphaGenome analyzes 1 Mb of DNA sequence, it is able to stretch 500 kb on each side (12). However, since some SNPs are farther than 500 kb from the *MYC* TSS, AlphaGenome's window was centered on the midpoint between the farthest SNP and the *MYC* TSS to include both in the analysis, leading to a window of chr8:126,843,275–127,891,851 (hg38, 0-based). This constant window was used for all analyses.

Chromatin accessibility screen

Chromatin accessibility was analyzed for *MYC* near rs6983267 and the *MYC* TSS, with rs6983267 as the candidate enhancer, in the PC-3 cell line context. PC-3 was chosen because it is a prostate cell line that is relevant to rs6983267 and multiple other SNPs. AlphaGenome was utilized to make predictions across the window, and the accessibility for *MYC* based on the predicted ATAC signal was visualized. ATAC-seq measures which parts of the genome are open and which are closed, providing a measure for chromatin accessibility (19). After this broad comparison, the ATAC signal near the *MYC* TSS and rs6983267 were separately visualized, spanning 2 kb on each side. This compared the mean and max signals near rs6983267 and the *MYC* TSS to demonstrate which positions had a higher relative ATAC signal. The mean predicted ATAC signal was computed and tabulated for each separate visualization to determine which region had a larger overall ATAC.

Targeted in silico mutagenesis (ISM)

AlphaGenome was utilized to perform in-silico mutagenesis (ISM). Each base pair across the rs6983267 enhancer window was mutated, with 3 alternate bases for each of 256 positions (12). The enhancer window was defined by centering the 256 base pair window on the candidate

enhancer. This was then scored with GeneMaskLFCScorer for gene-level RNA effects (12). GeneMaskLFCScorer is AlphaGenome's gene-level expression-change scorer, converting AlphaGenome's predictions into an effect score. ATAC was scored as usual, with ATAC-seq measuring which parts of the genome are open. In this ISM, GeneMaskLFCScorer scores how much AlphaGenome predicts *MYC* RNA abundance will change. The resulting ATAC and *MYC* RNA effect scores were visualized according to mutation position and base pair. These results were then ranked by the maximum predicted *MYC* RNA effect at each position.

Enhancer nomination

The predicted chromatin accessibility was analyzed along with contact maps to nominate SNPs in regions with features consistent with potential *MYC* enhancers (12). This was used to focus on certain SNPs before chromatin accessibility was compared between all SNPs. Because AlphaGenome did not provide a prostate-specific contact-map track, HCT116 was used as a colorectal substitute (12). This choice was due to rs6983267 also raising colorectal cancer risk (9). The criteria for nomination was reference chromatin accessibility, converted to a percentile against tiled background windows of the same width, with a cutoff in the top 10%, and a predicted 3D contact with the *MYC* promoter at 2,048 bp resolution and expressed as log-fold over the distance-expected contact, with a cutoff of greater than 0 (12). Percentile scores were calculated with the largest *MYC*-SNP window.

Risk vs. non-risk allele comparison

The ATAC track was predicted for both the non-risk and the risk allele, and compared them with each other (12). Each risk and non-risk track had a mean created inside their 1 Mb window by calculating predicted ATAC signal over equal 1001 bp windows centered on each SNP, and the non-risk mean was subtracted from the risk mean for a delta mean for each rsID. rs6983267 was run first as a validation check against the chromatin accessibility screen, and plots were generated in full for rs6983267 and the four SNPs flagged for largest ancestry disparities. Each SNP is still scored in the summary table even if plots were not generated, resulting in a comprehensive comparison between risk and non-risk contact and chromatin accessibility table.

Population frequency comparison

Population frequency data was compared in gnomAD v.4.1.0 (11). gnomAD population frequency data was downloaded for each of the 12 variants and their frequencies were analyzed in African/African American (AFR), Non-Finnish European (NFE), East Asian (EAS), Admixed American (AMR), and South Asian (SAS) ancestries. With these frequencies, a visualization was created for each rsID, directly comparing each ancestry's frequency data. An AFR/NFE ratio was also created and visualized for each variant. This data was then compared with AlphaGenome predictions to determine correlation between predicted regulatory effect and population frequency.

ACKNOWLEDGMENTS

We thank Polygence for providing this opportunity. This work was made possible through the use of AlphaGenome, gnomAD, and the UCSC browser.

References

1. Wittkopp, Patricia J., and Gizem Kalay. "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence." *Nature Reviews Genetics*, vol. 13, 2012, pp. 59–69. <https://doi.org/10.1038/nrg3095>
2. Seufert, Isabelle, et al. "Deregulated enhancer-promoter communication in cancer through altered nuclear architecture." *International Journal of Cancer*, vol. 158, no. 2, 2026, pp. 409–422. <https://doi.org/10.1002/ijc.35424>
3. Hu, Zhenhua, and Wee-Wei Tee. "Enhancers and chromatin structures: regulatory hubs in gene expression and diseases." *Bioscience Reports*, vol. 37, no. 2, 2017, BSR20160183. <https://doi.org/10.1042/BSR20160183>
4. Peterson, Roseann E., et al. "Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations." *Cell*, vol. 179, no. 3, 2019, pp. 589–603. [https://www.cell.com/cell/fulltext/S0092-8674\(19\)31002-5](https://www.cell.com/cell/fulltext/S0092-8674(19)31002-5)
5. Maurano, Matthew T., et al. "Systematic localization of common disease-associated variation in regulatory DNA." *Science*, vol. 337, no. 6099, 2012, pp. 1190–1195. <https://doi.org/10.1126/science.1222794>
6. Grisanzio, Chiara, and Matthew L. Freedman. "Chromosome 8q24-associated cancers and MYC." *Genes & Cancer*, vol. 1, no. 6, 2010, pp. 555–559. <https://doi.org/10.1177/1947601910381380>
7. Mudge, Jonathan M., et al. "GENCODE 2025: reference gene annotation for human and mouse." *Nucleic Acids Research*, vol. 53, no. D1, 2025, pp. D966–D975. <https://doi.org/10.1093/nar/gkae1078>; GENCODE, release 46, <https://www.encodegenes.org/>. Accessed 26 July 2026.
8. Dang, Chi V. "MYC on the path to cancer." *Cell*, vol. 149, no. 1, 2012, pp. 22–35. <https://doi.org/10.1016/j.cell.2012.03.003>
9. Ahmadiyeh, Nasim, et al. "8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 21, 2010, pp. 9742–9746. <https://doi.org/10.1073/pnas.0910668107>
10. ENCODE Project Consortium. "Candidate cis-Regulatory Elements (cCREs) track." *ENCODE*. <https://www.encodeproject.org/>. Accessed 26 July 2026.
11. Chen, Siwei, et al. "A genomic mutational constraint map using variation in 76,156 human genomes." *Nature*, vol. 625, no. 7993, 2024, pp. 92–100. <https://doi.org/10.1038/s41586-023-06045-0>; gnomAD, version 4.1.0, <https://gnomad.broadinstitute.org/>. Accessed 18 May 2026.
12. Avsec, Žiga, et al. "Advancing regulatory variant effect prediction with AlphaGenome." *Nature*, vol. 649, 2026, pp. 1206–1218. <https://doi.org/10.1038/s41586-025-10014-0>
13. "GWAS Catalog." *EMBL-EBI*. <https://www.ebi.ac.uk/gwas/>. Accessed 10 Aug. 2026.



14. Haiman, Christopher A., et al. "Multiple regions within 8q24 independently affect risk for prostate cancer." *Nature Genetics*, vol. 39, no. 5, 2007, pp. 638–644. <https://doi.org/10.1038/ng2015>
15. Al Olama, Ali Amin, et al. "Multiple loci on 8q24 associated with prostate cancer susceptibility." *Nature Genetics*, vol. 41, 2009, pp. 1058–1060. <https://doi.org/10.1038/ng.452>
16. Takata, Ryo, et al. "Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population." *Nature Genetics*, vol. 42, no. 9, 2010, pp. 751–754. <https://doi.org/10.1038/ng.635>
17. "UCSC Genome Browser." *University of California, Santa Cruz*. <https://genome.ucsc.edu/>. Accessed 26 July 2026.
18. "Ensembl REST API." *EMBL-EBI*. <https://rest.ensembl.org/>. Accessed 10 Aug. 2026.
19. Buenrostro, Jason D., et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods*, vol. 10, 2013, pp. 1213–1218. <https://doi.org/10.1038/nmeth.2688>