Single Nucleotide Polymorphism Facilitated Down-Regulation of the Cohesin Stromal Antigen-1: Implications for Colorectal Cancer Racial Disparities

Abstract
The biological underpinnings for racial disparities in colorectal cancer (CRC) incidence remain to be elucidated. We have previously reported that the cohesin SA-1 down-regulation is an early event in colon carcinogenesis which is dramatically accentuated in African-Americans. In order to investigate the mechanism, we evaluated single nucleotide polymorphisms (SNPs) for association with SA-1-related outcomes followed by gene editing of candidate SNP. We observed that rs34149860 SNP was significantly associated with a lower colonic mucosal SA-1 expression and evaluation of public databases showed striking racial discordance. Given that the predicted SNP would alter miR-29b binding site, we used CRISPR knock-in in CRC cells and demonstrated that the SNP but not wild-type had profound alterations in SA-1 expression with miR-29b inhibitor. This is the first demonstration of high-order chromatin regulators as a modulator of racial differences, risk alteration with SNPs and finally specific modulation by microRNAs.

Introduction
Colorectal cancer (CRC) remains the second leading causes of cancer mortality in the United States and Western countries despite effective modalities for screening fecal tests and colonoscopy [1]. Barriers to implementation include compliance, resource utilization and complication/discomfort. This is juxtaposed by the relatively low rate (~7%) of advanced adenomas, the immediate precursor of CRC, meaning that the vast majority of colonoscopies do not have cancer preventive implications. This underscores the need of more effective risk prediction strategies to personalize CRC screening strategies [2]. Approximately 30% of CRC risk is considered genetic making inherited factors a logical target for risk stratification efforts [3].

Genetic predilection for CRC is epitomized by high penetrance but relatively infrequent syndromes such as familial adenomatous polyposis and Lynch syndrome. More commonly, risk alleles are due to single nucleotide polymorphisms (SNPs) in genes important in colon carcinogenesis. Efforts to harness this have been through genome-wide association studies (GWAS) with a plethora of report albeit most with low risk factors. Furthermore, variation by population has also been reported in GWAS.

Since many SNPs have a racial predilection, this provides a potential mechanism for racial disparities in CRC. Strikingly the relative incidence and mortality rates of CRC were ~25% and ~50% higher, respectively, in Blacks compare to Whites [1,4]. While socio-economic factors are undoubtedly contributory, there are also marked biological differences between CRCs in Blacks and Whites [5]. These might be genetically related with studies showing SNPs in p53 [6], TGFβ-related genes (BMP4, GREM1, CDH1, SMAD7 and RPHN2) [7] and mannose-binding lectin 2 (MBL2) gene [8].
Given the myriad of differentially expressed genes dysregulated, high order chromatin remodelers as a modality for these pleotropic genomic effects [9–12]. The cohesins, including SA-1 (also known as STAG1 and SCC3), are a recently discovered high order chromatin remodelers that are involved in chromatin looping (bringing enhancers/promoters in proximity to coding regions to foster transcription) and thus facilitate long-range nuclear interactions and gene expression [13]. While mutations in this family of proteins are rare in CRC, we have recently reported that SA-1 was significantly down-regulated in the endoscopically normal rectal mucosa of patients harboring adenomas elsewhere in their colon (field carcinogenesis) presumably through epigenetic mechanisms [14]. Moreover, SA-1 loss triggered critical pathways in carcinogenesis [15–17]. Importantly, we noted that rectal mucosal SA-1 was dramatically reduced in Blacks when compared to Whites providing a potential mechanism for racial differences in CRC rates. Since SNPs may segregate racially, we aimed to test the hypothesis that SNPs may relate to decrease in SA-1 expression selectively by race. We combined in silico work, human colon, and cell culture (CRISPR knock-in) work to demonstrate that an exonic polymorphism rs34149860 (selected based on p-value and racial segregation) had a racial predilection and was associated with decreased SA-1 expression. Mechanistically, we discovered that the rs34149860 impacted mRNA expression.

**Materials and Methods**

**Human Studies**

We recruited 81 patients undergoing screening/surveillance colonoscopy and obtained two biopsies from the endoscopically normal rectal mucosa. We also evaluated CRC samples (n = 14) from our repository.

**SNP Selection**

SA-1 SNPs (n = 27) were selected for genotyping based on: (1) eQTL potential for SA-1 expression in any tissue based on the GTEx database (http://www.gtexportal.org/home/), (2) missense mutations within SA-1 identified in the SnpEff databases (http://snpeff.sourceforge.net/), (3) SNPs capturing the haplotype diversity in SA-1.

**SNP Array**

SNP genotyping was performed using 7900HT Fast Real-Time PCR System (Life Technologies, Grand Island, NY) and analyzed using SDS software version 2.3.

---

![Figure 1](image_url)

**Figure 1.** Cohesin SA-1 (STAG1) gene mapping for the (A) CEU and (B) ASW populations formed by genotyped SNPs. Abbreviations: CEU – European American (White) population represented by Utah residents with Northern and Western European ancestry. ASW – African American (Black) population represented by the African ancestry in Southwest USA.
**CRISPR-KI Method**

CRISPR gene knock-in technique was used to change the SNP, rs34149860 in the SA-1 (STAG1) genomic DNA in HCT116 and RKO cells. Gene expression performed using Step-One-Plus RT Thermo cycler (Life Technologies).

**DNA Sequencing**

Genotyping of CRC cells for SNP rs34149860 and validation of CRISPR knock-in was done by PCR amplification and purification followed by Sanger DNA sequencing at Genewiz (Cambridge, MA) using the following primers: rs34149860.Fwd: CAGAGCCA CACTCTTCACAAGT; rs34149860.Rev: ACCTGGTGAGACAGCCGTACATT.

**Statistical Methods**

Logistic and linear regression models (PLINK) were used for the analysis to test for associations between SA-1 SNPs, SA-1 expression, and CRC risk. Interaction tests were used to determine whether the effect of these SNPs differed between Blacks and Whites. To assess the significance of the observed \( P \) values and correct for multiple testing, we created 1000 replicates of the dataset by randomly permuting genotypes and phenotypes.

**Results**

**In Silico Analysis Linkage Disequilibrium and Haplotype Diversity in Cohesion SA-1 Between Blacks and Whites**

Linkage disequilibrium (LD) patterns have been well-characterized in population-genetics studies as well as mapping disease-causing genes by comparing allele frequencies between case and control groups. The LD across the different regions of the world varies considerably due to biological and random factors and is partly attributed to age of population and the number of founders. Based on our recent report demonstrating significant loss of SA-1 gene expression (both baseline and in cancer) in Black population, we investigated the LD distribution of SA-1 gene between Blacks and Whites within the general population using public databases.

We downloaded genome-wide SNP genotyping datasets using HapMap 3 (International HapMap Project, Phase 3 samples) Genome Browser release #28 and used the Haplovie 4.2 software tool (Broad Institute/MIT) to compute linkage disequilibrium (LD)

**Table 1. Effect of SA1 Polymorphisms on SA-1 Expression in Total Patient Cohort**

<table>
<thead>
<tr>
<th>SNP</th>
<th>SA1 Deregulation</th>
<th>( P )-Value</th>
<th>% AFR</th>
<th>% EUR</th>
<th>Type of Variant</th>
<th>Selection Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs34149860</td>
<td>161% down</td>
<td>7.992e-05</td>
<td>31</td>
<td>0</td>
<td>Missense</td>
<td>Missense</td>
</tr>
<tr>
<td>rs1070232</td>
<td>125% down</td>
<td>0.03096</td>
<td>30</td>
<td>26</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs13071220</td>
<td>143% down</td>
<td>0.00073</td>
<td>7</td>
<td>20</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs1681818</td>
<td>197% down</td>
<td>0.00111</td>
<td>31</td>
<td>26</td>
<td>Non coding transcript exon variant</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs17364492</td>
<td>115% down</td>
<td>0.01076</td>
<td>14</td>
<td>20</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs2655008</td>
<td>159% down</td>
<td>0.00012</td>
<td>33</td>
<td>27</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs34992220</td>
<td>199% down</td>
<td>0.00018</td>
<td>44</td>
<td>67</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs6788433</td>
<td>197% down</td>
<td>0.0023</td>
<td>34</td>
<td>29</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs6773604</td>
<td>128% down</td>
<td>0.048</td>
<td>48</td>
<td>31</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs696517</td>
<td>141% down</td>
<td>0.0003</td>
<td>27</td>
<td>20</td>
<td>Non coding transcript exon variant</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs7646127</td>
<td>109% down</td>
<td>0.03778</td>
<td>32</td>
<td>0</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
<tr>
<td>rs9861150</td>
<td>133% down</td>
<td>0.0184</td>
<td>14</td>
<td>22</td>
<td>Intron</td>
<td>tag SNP</td>
</tr>
</tbody>
</table>

**Abbreviations:** AFR, African; EUR, European.
The racial prevalence estimate (% AFR and % EUR) are obtained from well-annotated databases (dbSNP and Ensembl).

and population haplotype patterns for Blacks represented by the African ancestry in Southwest USA (ASW; \( n = 90 \)) cohort and Whites represented by Utah residents with Northern and Western European ancestry from the CEPH collection (CEU; \( n = 180 \)). As shown in Figure 1A and B, we observed striking heterogeneity in the LD \( (r^2) \) between the Black and White populations within the genomic region of SA-1.

**Effect of SA-1 SNPs on SA-1 Gene Expression With Race-Specific Implication**

Based on the dramatic difference in the LD of SA-1 gene, we next investigated whether SA-1 polymorphisms are associated with the loss of SA-1 gene expression and risk of CRC. The SA-1 gene is a member of the SCC3 family and expressed in the nucleus. It is located on Chromosome 3 (3q22.3). The gene contains 34 exons (33 coding exons) with a transcript length of 6055bps and translational length of 1258 amino acid residues. We observed 23,484 SNPs in human SA-1. Out of these SNPs, 236 were missense mutations, 22,789 were introns, 146 were 3’UTR and 89 were 5’ UTR. To identify a candidate SA-1 variant that is associated with decreased gene expression and risk of CRC, we selected 27 SNPs based on eQTL potential, impact on protein function, and haplotype tagging. Using Taqman SNP genotyping array we have identified 13 SNPs which were found significantly associated with SA-1 down-regulation (without race adjustment) (Table 1). We focused on SNP rs34149860 which encoded a missense alteration in SA-1 (allele change – \( \text{CAC} \rightarrow \text{CAG} \); \( \text{CAG} \rightarrow \text{CAC} \); residual change – \( \text{Q} \rightarrow \text{H} \) because it was observed to have marked racial prevalence differences (31% of black and 0% of whites). Furthermore, when we evaluated our 14 CRC patients, we also observed that rs34149860 was associated with CRC risk (OR: 1.96; 95% CI: 0.70 ± 5.453; \( P = 0.2 \)). These data show for the first time that SA-1 SNPs have racial predilection providing a biological basis into the preferential loss of the SA-1 tumor suppressor gene in African Americans and hence higher incidence of CRC.

**In-Vitro Engineering of rs34149860 SNP Through CRISPR Knock-In**

We genotyped a panel of CRC cells for presence of SNP rs34149860. RKO cells which are heterozygous for the SNP
rs34149860 (CG, one alternate allele—C and one reference allele G) showed higher SA-1 expression compared to HCT116 cells which are homozygous for the SNP (GG) (Figure 2A). We successfully designed CRISPR knock-in to create the reciprocal allele alterations within RKO and HCT116 cells. In RKO cells, changing to a homozygous variant (CG to GG) alleles decreased SA-1 by 36% ($P = .002$) whereas in HCT116 cells changing from reference allele variant to alternate allele (GG to CC) increased expression by 91% ($P = .004$). These results suggest a significant contribution of rs34149860 SNP in the regulation of SA-1 expression.

**miRNA Mediated Down-Regulation of SA-1 Targeting SNPs**

Our discovery that rs34149860 SNP is associated with decreased SA-1 expression is intriguing. However, it is not intuitive to believe a missense alteration going from glutamine $\rightarrow$ histidine will result in decreased SA-1 mRNA expression as noted. Since it does not involve promoter or enhancer regions but rather exons, it is unlikely to be methylation. We posited that SNP rs34149860 might play a role in miRNA (miR) binding and mRNA destabilization. Presence of SNPs can disrupt normal miR binding and/or introduce new binding sites which might be associated with disease pathogenesis. This association emphasizes the importance of detecting miR targets and predicting the possible effects of SNPs on binding sites. Dysregulation of miRs are common in colon carcinogenesis [18]. Our group has noted that miRs are dysregulated in field carcinogenesis [19]. We used RegRNA (A Regulatory RNA Motifs and Elements Finder) to identify the predicted miRNA binding to the rs34149860 region (±25bp around SNP transcript) that would be disrupted by variant. Our analysis on the possible effects of SNPs on miR binding sites [20] showed binding of miR-29b-1* as the most promising candidate (Figure 2C). In order to get empiric data, we used RKO cells (baseline heterozygous) and engineered variant status (homozygous) with CRISPR knock-in. We noted that the induction of SA-1 expression with miR-29b-1* inhibitor showed a minimal (15%. $P = .23$) increase in the control cells whereas the same concentration of inhibitor showed a dramatic 226% increase ($P = .004$) in the variant (Figure 2D). To determine the clinical significance, we validated that miR-29b-1* was expressed in rectal mucosa of ~30 randomly selected patients (data not shown). These emerging findings are highly novel and provide novel mechanistic insights into CRC racial disparities.

**Discussion**

In this proof of concept study, we took a candidate approach to identify a SA-1 SNP that was associated with decreased SA-1 expression in the colonic mucosa. We report herein, that SNP rs34149860 enabled miR-mediated degradation of colonic mucosal cohesin SA-1 expression which we previously reported is the hallmark of CRC risk. Importantly, the racial predilection of this SNP provides a compelling potential mechanism for CRC disparities.

High order chromatin has been shown to be a fundamental event in carcinogenesis. We have demonstrated that high order chromatin is altered in the endoscopically-normal rectal mucosa patients harboring colon neoplasia using partial wave spectroscopic microscopy (PWS) [21]. Members of SWI/SNF family including Arid 1a has been shown to be one of the most commonly mutated genes involved in chromatin-remodeling and gene regulation [22,23]. A new family of high order chromatin remodelers is the cohesins. These exist as a family (including Smc1, Smc3, Rad21/Scc1 and SA-1 or SA2) that
can play a critical role in sister chromatid segregation (cohesion) and more relevant to CRC initiation is change in high order chromatin allowing genes to be transcribed via topologically-associated domains and chromatin loops [13,15,24,25]. A case in point has been a seminal report on glioblastoma that human isocitrate dehydrogenase 1 (IDH-1) mutant gliomas exhibit hypermethylation at cohesin and CCCTC-binding factor (CTCF)-binding sites, compromising binding of this methylation-sensitive insulin protein [16]. Reduced CTCF binding is associated with loss of insulation between topological domains and aberrant gene activation through allowing interaction of a constitutive enhancer element with PDGFRα. These cohesins are turned off with factors such as WAPL providing a mechanism to restrict chromatin looping functions through triggering cohesin release from chromatin [25,26]. The ability of cohesins to impacting promoter-enhancer function is potentially apropos given the emerging data on the role of enhancers in colon carcinogenesis [9]. In this regard, Cohen et al. showed that approximately half of the GWAS CRC risk loci co-localized with recurrently activated enhancers. Moreover, several of these enhancers are occupied by cohesin complex members [9]. Additionally, it has been noted that enhancer activation via Arid1a loss is sufficient to induce intestinal tumorigenesis thus emphasizing the potential biological implications [10–12].

While somatic mutations in cohesins are well established, germline mutations or SNPs as a risk factor has not been described. To our knowledge this is the first report of SNPs in chromatin remodelers associated with colon cancer risk and thus may provide important insights. Furthermore, the racial predilection may explain, at least partly, the striking differences in CRC rates for Blacks compared to Whites. The patient database was small and thus the odds ratio for CRC was not significant but the decreased colonic SA-1 expression is compelling given the unequivocal data that SA-1 down-regulation is an earliest event in colon carcinogenesis [27].

MicroRNAs (miRs) are short ~22–27 nucleotide long non-coding RNAs that bind specific sequences on mRNA and promote degradation through RNA-induced silencing complex (RISC). Dysregulation of miRs is an important epigenetic mediator during colon carcinogenesis and has been exploited as a biomarker (found in blood and stool). We have earlier reported that miRs are dysregulated in field carcinogenesis [19] thus providing a mechanism for SA-1 degradation. In this proof of concept study, we documented the novel interaction between SA-1 and miR which is based partly on the critical hypothesis that miRs may act selectively on SNPs or certain miRs may be selectively altered in the colon by race. In addition, miR-29b has been shown to be an important CRC biomarker supporting its relevance [28].

There are numerous strengths of this study. To our knowledge, this is the first to link SNPs in cohesins in particular and high order chromatin modulators in general to CRC risk. This is particularly important given the emerging knowledge regarding the importance of these. The comparison to actual mRNA in the colon mucosa is novel and powerful in a manner to go beyond the typical association studies. The elucidation of a mechanism through which SNP leads to decreased RNA expression provides biological plausibility and insights.

There are several limitations that need to be acknowledged. The SNP selection was based on candidate approach and not comprehensive with other SNPs may also be important. It bears reiteration that our goal was simply to demonstrate proof of principle that SNPs can impact SA-1 expression and CRC risk. Therefore, we deferred standard biomarker approaches (independent validation set, adjustment for multiple comparators etc.) for future studies. There might be other epigenetic modulations including methylation that may also contribute to the racial differences. The link between SA-1 and CRC is indirect based on the compelling association between mucosal SA-1 and CRC risk [27]. Our pilot CRC data (n = 14) was suggestive (~2-fold risk) but failed to achieve statistical significance. Furthermore, given that self-reported race [29] accuracy has been recently impugned and we were unable to genotype our patients, we resorted upon well-annotated databases for racial prevalence estimate since genotyping our patients was not feasible.

In conclusion, we provide the first report on cohesin SA-1 SNP and colonic mucosal gene expression as a surrogate for CRC risk. The mechanistic insights into the role of miRNA mediated SA-1 degradation provide novel insights. While corroboration in large scale is needed for rs34149860 and other SA-1 SNPs, this has potential for personalization of screening targeting racial differences.

Acknowledgements
We thank Ms. Irene Simkin, Project Manager, Molecular Genetics Core at Boston University School of Medicine for performing the SNP array and Ms. Beth Parker for excellent support in manuscript preparation.

Grant Support
NIH grants – R03CA195143, R01CA165309, R01CA156186, R01CA200064, R01CA183101, R33CA225323-01.

The Role of the Study Sponsor
The National Institutes of Health had no role in the study.

Conflict of Interest Statement
Drs. Roy and Backman are co-founders and shareholders in Nanocytomics LLC. All other authors do not have any competing interests.

References
with increased colon cancer risk in African Americans. Cancer Res 72(6), 1467–1477.


