Bioinformatics and Biology Insights



ORIGINAL RESEARCH

# Candidate Single Nucleotide Polymorphism Markers for Arsenic Responsiveness of Protein Targets

Raphael D. Isokpehi[1,2], Hari H.P. Cohly[1,2], Matthew N. Anyanwu[2,3], Rajendram V. Rajnarayanan[4], Paul B. Tchounwou[1], Udensi K. Udensi[1,2] and Barbara E. Graham-Evans[1]

[1]RCMI-Center for Environmental Health, College of Science, Engineering and Technology, Jackson State University, Jackson, Mississippi 39217, USA. [2]Center for Bioinformatics & Computational Biology, Department of Biology, Jackson State University, PO Box 18540, Jackson, Mississippi 39217, USA. [3]Department of Computer Science, University of Memphis, Memphis Tennessee, USA. [4]Department of Pharmacology and Toxicology, State University of New York at Buffalo, Buffalo, New York, USA. Corresponding author email: raphael.isokpehi@jsums.edu

**Abstract:** Arsenic is a toxic metalloid that causes skin cancer and binds to cysteine residues—a property that could be used to infer arsenic responsiveness of a target protein. Non-synonymous Single Nucleotide Polymorphisms (nsSNPs) result in amino acid substitutions and may alter arsenic binding with cysteine residues. Thus, the objective of this investigation was to identify and analyze nsSNPs that lead to substitutions to or from cysteine residues as an indication of increased or decreased arsenic responsiveness. We hypothesize that integration of data on molecular impacts of nsSNPs and arsenic-gene relationships will identify nsSNPs that could serve as arsenic responsiveness markers. We have analyzed functional and structural impacts data for 5,811 nsSNPs linked to 1,224 arsenic-annotated genes. In addition to the identified candidate nsSNPs for increased or reduced arsenic responsiveness, we observed i) a nsSNP that results in the breakage of a disulfide bond, as candidate marker for reduced arsenic responsiveness of KLK7, a secreted serine protease participate in normal shedding of the skin; and ii) 6 pairs of vicinal cysteines in KLK7 protein that could be binding sites for arsenic. In summary, our analysis identified non-synonymous SNPs that could be used to evaluate responsiveness of a protein target to arsenic. In particular, an epidermal expressed serine protease with crucial function in normal skin physiology was prioritized on the basis of abundance of vicinal cysteines for further research on arsenic-induced keratinocyte carcinogenesis.

**Keywords:** arsenic, keratinocytes, non-synonymous single nucleotide polymorphisms, toxicogenomics, skin cancer, vicinal cysteines

*Bioinformatics and Biology Insights* 2010:4 99–111

doi: 10.4137/BBI.S5498

This article is available from http://www.la-press.com.

## Introduction

Arsenic (As) is recognized as an environmental toxicant of concern for global public health and a leading cause of toxicity and carcinogenicity.[1,2] Arsenic targets the human skin and long-term exposure to arsenic, principally through drinking water, has been correlated with increased risk of skin cancer.[3–5] The cellular toxicity of arsenic has been well documented from case studies of poisoning incidents and medicinal use.[2,6] However, due to increased epidemiological reports of arsenic related cancers in places such as Southeastern Michigan (USA), Taiwan, China, India and Bangladesh, public health concerns about long-term exposure have arisen.[2,6–8] Inorganic arsenic is classified by the United States Environmental Protection Agency (U.S. EPA) as a Group A carcinogen based on sufficient evidence of carcinogenicity in humans.[9] Chronic oral exposure to inorganic arsenic can have adverse effects on tissues in the human body systems.[10] However, the human skin is the critical organ of arsenic toxicity because arsenic has a strong affinity for the keratin proteins which are rich in the sulphur containing cysteine residues.[11–13] Chronic exposure to arsenic induces sequential changes in the skin epithelium, proceeding from hypopigmentation to hyperkeratosis which may eventually lead to skin cancer.[11] Arsenic-induced skin lesions are early warning markers for development of cancers in internal organs.[14,15] A well-known beneficial use of arsenic is that arsenic trioxide is used for treatment of relapsed or refractory acute promyelocytic leukemia.[16–20] However, side effects of treatment with arsenic trioxide include significant adverse cardiac effects.[21]

The avalanche of genome sequences combined with genome-enabled datasets from high-throughput gene expression, genotyping, haplotyping and protein assays is making it possible to gain biological insights into previously unknown gene-toxicant interactions. Arsenicogenomics, an aspect of toxicogenomics, therefore, provides a means to i) understand how various genes respond to arsenic and ii) how arsenic modifies the function and expression of specific genes in the genome. Early physical manifestations of arsenic toxicity in endemic areas are skin lesions including melanosis and keratosis. However, not everyone exposed to arsenic in an endemic region would develop skin lesions.[22] Therefore, future research on arsenic-induced cancers and, in particular, skin lesions should consider the impact of genetic variation in individual susceptibilities to arsenic toxicity.

Single nucleotide alterations in the DNA sequence represent a major source of genetic heterogeneity[23] and the most common type of genetic variation in the human genome. The diversity of single nucleotide polymorphisms (SNPs) derived from arsenic responsive genes in different populations could provide biomarkers for an individual's susceptibility to arsenic-induced diseases. Genomic and bioinformatics techniques now exist to identify and analyze the presence of SNPs in populations.[24] Furthermore, the dense distribution of SNPs across the genome makes them ideal markers for large-scale genome-wide association studies to discover genes in common complex diseases, such as cancer. A SNP-induced amino acid substitution in the coding region can be broadly divided into synonymous (no change in amino acid) or non-synonymous (change in amino acid).[25] Furthermore, the functional impact of the SNP on protein function has been described as deleterious (disruptive) or non-deleterious (benign/neutral).[26] Non-synonymous substitutions could lead to missense or nonsense mutations in the encoded polypeptide. In particular, nonsense mutations that generate premature termination codons (PTCs) are responsible for approximately one-third of human genetic diseases.[27] In addition, substitutions in one or two amino acids of a protein sequence can alter the quantity of encoded protein during expression in mammalian cells.[28] Considering that arsenic trioxide is also used for treatment of acute promyelocytic leukemia,[16,18,19,29] genetic variation may also affect response to therapy.

Arsenic binds to sulfhydryl (SH) groups of cysteine (Cys) residues to form arsenic-thiol linkages, a property that could be used to infer arsenic responsiveness of a protein target as well as contribute to oxidative and protein folding stresses.[30–32] In the human arsenic (+3 oxidation state) methyltransferase (hAS3MT) sequence, Cys residues at positions 156, 206 and 250 play important roles in the enzymatic function and structure.[33] Mutation of the arsenic-sensing Cys151 in Kelch-like ECH-associated protein 1 (Keap1) abolished arsenic activation of nuclear factor erythroid 2-related factor 2,

a transcription factor responsible for induction of antioxidative cytoprotective genes.[34]

Non-synonymous Single Nucleotide Polymorphisms (nsSNPs) result in amino acid substitutions and may alter the number of cysteine residues available to arsenic for binding to a protein cellular target. Therefore, the objective of this investigation was to identify and analyze nsSNPs that lead to substitutions to or from cysteine residues as an indication of increased or decreased arsenic responsiveness of a protein. We hypothesize that integration of data on molecular impacts of non-synonymous single nucleotide polymorphisms and arsenic-gene relationships will help identify nsSNPs that are candidate arsenic responsiveness markers.

An integrative approach combining results from selected web-based toxicogenomics and genomics databases as well as bioinformatics tools was used to prioritize candidate nsSNPs markers for arsenic responsiveness of protein targets. In the first step, a list of genes annotated to interact with arsenicals was retrieved from the Comparative Toxicogenomics Database.[35] Subsequently, the nsSNPs linked to these arsenic-annotated genes were extracted from SNPs3D[36] and analyzed for functional and structural impacts data on protein isoforms. Significant amino acid substitutions to or from cysteine residues were then prioritized based on structural effect resulting in breakage of a disulphide bond as well as function in skin cells. Furthermore, structural homology modeling was used to identify vicinal (neigboring) cysteines in prioritized protein targets of arsenic. In order to facilitate additional investigations on prioritized SNPs, protein targets and molecular mechanisms of arsenic action, we have constructed a collection of over 100,000 sentences from over 16,000 PubMed[37] abstracts on arsenic. Finally, a web resource Arsenic Sentence Database was developed to enable web-based search of the sentences by keywords and PubMed identifiers.

## Methods

### Functional and structural impacts of single nucleotide polymorphisms on arsenic annotated genes

The molecular functional effects of non-synonymous SNPs based on sequence and structure analysis were retrieved from the SNPs3D web resource and database[36] for genes curated in the Comparative Toxicogenomics Database (CTD)[35] to have a relationship with arsenic. In the CTD, the term gene also includes mRNA and proteins. We described relationship in terms of arsenic modifying the function and/or expression of genes. Furthermore, we referred to the genes as arsenic-responsive genes or proteins. In SNPs3D, the classification into *in vivo* functional impact categories of the SNP was based on two Support Vector Machine (SVM) models: protein sequence conservation profiling and protein structure stability. In both machine learning models, an SVM is trained using 5 sequence profiles and 15 protein stability features. Additional details on the methods are available at the SNPs3D website http://www.snps3d.org/help/method.html. The nsSNPs were ranked according to SVM score. For both sequence and structure SVM scores, a nsSNP with negative score was classified as deleterious while a nsSNP with a positive score was classifed as non-deleterious. We observed that some nsSNPs in SNPs3D were assigned a SVM score of −0.00. However, they were not tagged as deleterious. Thus, these nsSNP were classified in this investigation as non-deleterious. Furthermore, High Confidence (HC) SVM scores were greater than 0.50 or less than −0.50.[38]

The computational workflow consisting of a suite of customized Perl and Unix scripts was developed to process results obtained from CTD and SNPs3D. The Entrez Gene[37] Identifiers and Gene Symbols were extracted from XML formatted results of arsenic-gene interactions from CTD. Furthermore, the Entrez Gene identifiers were then used to remotely download the SNP Analysis page in SNPs3D for each CTD arsenic-annotated gene. For example, the SNP Analysis page for a known arsenic-interacting gene Glutathione S-transferase Omega 1 (Gene Symbol: GSTO1 and Entrez Gene Identifier: 9446) in SNPs3D is http://www.snps3d.org/modules.php?name=SnpAnalysis&locus_ac=9446.

The collection of html files was processed to mine for relevant data to construct a dataset. The fields of the dataset were the dbSNP identifier,[37] RefSeq protein isoform identifier,[37] mutation, SVM sequence profile score, SVM structure score and description of impact of nsSNP on protein stability. There were instances that no data were predicted for the SVM structure score and the protein impact. However,

scores were predicted for all the SVM sequence profile. The gene symbols were also extracted from the CTD and combined with the dataset from SNPs3D. The final integrated dataset consisted of the Entrez Gene Identifier, Gene Symbol, the SNP Identifier, the Amino Acid Substitution, the SVM score computed for the sequence and structure profile models and the structural consequence of the SNP. In SNPs3D, we preferred to remotely download the SNP Analysis pages so as to extract data and links to additional information such as i) sequence alignment evidence of tolerance of the amino acid position to mutation and ii) values associated with the 15 protein stability factors. These additional datasets were not available in the files available for download.

### Arsenic responsiveness based on substitution to or from cysteine residues

We conjectured that substitution to or from a cysteine residue could lead to increased or decreased responsiveness of a protein isoform to arsenic. In order to identify these cysteine substitutions, a suite of customized Perl and Unix scripts was developed to extract records that met the criteria from the integrated CTD and SNPs3D dataset. For example, in SNPs3D SNP Analysis page for GSTO1, nsSNP rs45529437 is linked to substitution from cysteine to tyrosine in position 32 (C32Y). Furthermore, rs11509436 is linked to substitution to cysteine from serine in position 86 (S86C).

### Structural homology modeling

Structural models of protein mutants link to candidate SNPs that alter arsenic responsiveness were generated using MODELLER 7v7[39] with appropriate homologous high resolution X-ray crystal structure templates from the Protein Data Bank (PDB).[40] SYBYL (Tripos Inc) was used to identify vicinal cysteines following a quick minimization routine using AMBER force field.

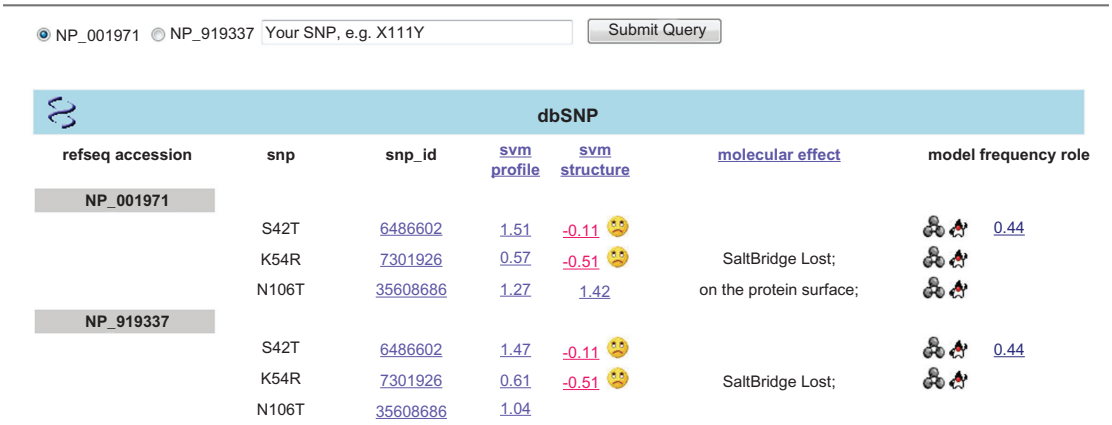
### Construction of sentences collection from PubMed abstracts on arsenicals

In order to identify descriptors of interest in sentences and cluster sentences with identical descriptors, we implemented a sentence splitting algorithm on a collection of PubMed[37] abstracts annotated with at least one of the Medical Subject Heading (MeSH) terms: arsenic or arsenicals. The sentence splitting algorithm implemented uses Perl regular expressions to enhance the Comprehensive Perl Archive Network (CPAN) Text: Sentence splitter module (http://search.cpan.org/) to achieve a high accuracy in sentence disambiguation. A web interface for searching the catalog of sentences was also developed.

## Results

### Functional and structural impacts of single nucleotide polymorphisms on arsenic annotated genes

The set of genes for predicting potential responsiveness to arsenic was obtained from the Comparative Toxicogenomics Database (CTD).[35] A total of 1,604 genes consisting of 1,492 human genes and 112 non-human genes documented to have a relationship with arsenicals (Medical Subject Heading Identifier [MeSH ID]: D001152) were retrieved on May 10, 2010 (Supplementary Data). We describe the gene set as a list of arsenic-annotated genes and their protein isoforms as arsenic-annotated proteins. The functional impacts of SNPs on protein function, as predicted by support vector machine (SVM), were retrieved from SNPs3D.[36] Support Vector Machines (SVM) are supervised machine learning techniques that have been applied to numerous classification tasks to predict the class of an example based on training examples. The SNPs3D database provides an SVM profile score for the functional impact (deleterious or non deleterious) of an nsSNP as well as 3-dimensional protein structure of the impact of the nsSNP on protein stability. According to Yue et al[36] the SVM used in SNPs3D is trained on monogenic disease data, thus deleterious is defined as "sufficiently damaging to protein function *in vivo* as to be consistent with a monogenic disease outcome". A screenshot of a section of SNPs3D page for a gene is presented in Figure 1. The pages of the arsenic-annotated genes available in the SNPs3D database were the data source for extracting relevant data. The protein stability impacts in SNPs3D were i) Four classes of electrostatic interaction: reduction of charge–charge, charge–polar or polar–polar energy, or introduction of electrostatic repulsion; ii) three solvation effects: burying of charge or polar groups, and reduction in non-polar area buried on folding; iii) and two terms representing steric strain: backbone


NP_001971 NP_919337 Your SNP, e.g. X111Y Submit Query

dbSNP

| refseq accession | snp | snp_id | svm profile | svm structure | molecular effect | model frequency role | |
|---|---|---|---|---|---|---|---|
| NP_001971 | | | | | | | |
| | S42T | 6486602 | 1.51 | -0.11 | | | 0.44 |
| | K54R | 7301926 | 0.57 | -0.51 | SaltBridge Lost; | | |
| | N106T | 35608686 | 1.27 | 1.42 | on the protein surface; | | |
| NP_919337 | | | | | | | |
| | S42T | 6486602 | 1.47 | -0.11 | | | 0.44 |
| | K54R | 7301926 | 0.61 | -0.51 | SaltBridge Lost; | | |
| | N106T | 35608686 | 1.04 | | | | |

**Figure 1.** Screenshot of a SNPs3D page for a gene. The functional and structural impacts, molecular effect and frequency of non-synonymous SNPs associated with the protein isoforms (RefSeq accession) is documented on the page. The negative SVM score (value in red) indicates a deleterious substitution.

strain and over-packing; cavity formation (affecting van der Waals energy); iv) and loss of a disulfide bridge.

From SNPs3D, we extracted the Entrez Gene Identifier, the SNP Identifier, the Amino Acid Substitution and the SVM score computed for the sequence and structure profile models. The data-set constructed consisted of 5,811 nsSNPs linked to 1,224 arsenic-annotated genes. Furthermore, a total of 8,992 nsSNP-induced substitutions (3,700 deleterious, 5,292 non-deleterious) were linked to 1,872 protein isoforms in the National Center for Biotechnology Information (NCBI) Reference Sequence Database.[37] The SVM scores observed for the substitutions ranged from –4.73 to 6.59 with 743 unique scores (Fig. 2). Of the 3,700 nsSNP-predicted substitutions, there were 2,739 high confidence deleterious substitutions (SVM sequence profile score $<-0.5$) linked to 745 genes, 1,785 nsSNPs and 1,094 protein isoforms. Furthermore, there were 4,191 high confidence non-deleterious substitutions (SVM sequence profile score $>0.5$) linked to 964 genes, 2,829 and 1,459 nsSNPs. A summary of

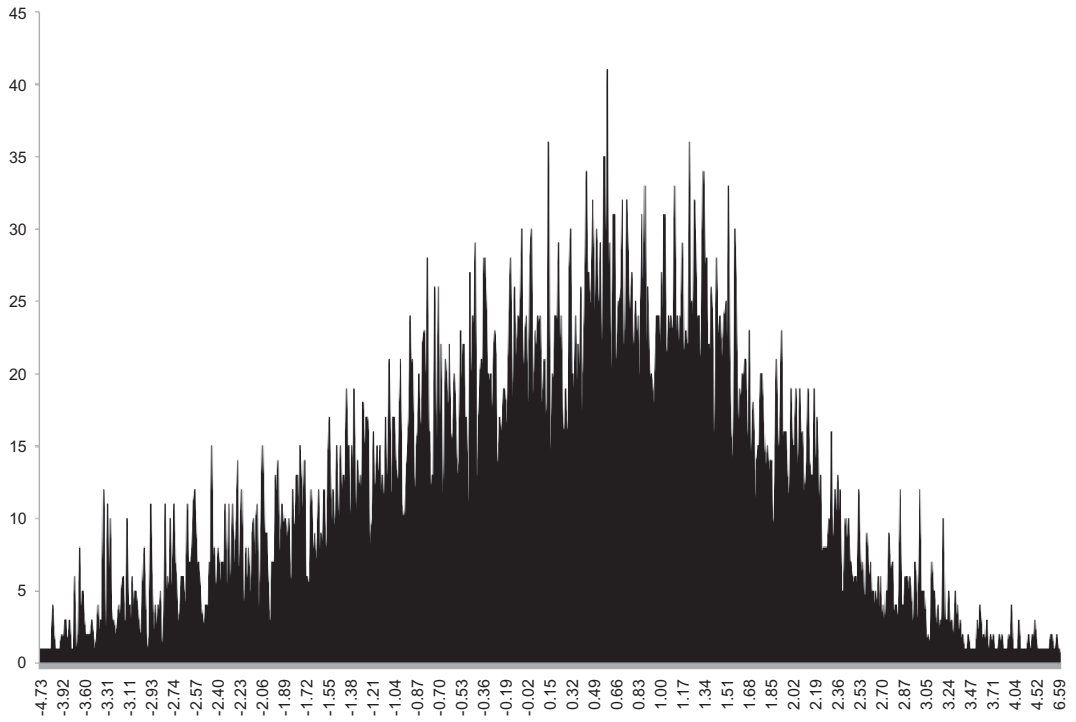



**Figure 2.** Plot of frequencies of unique Support Vector Machine (SVM) scores for dataset of non-synonymous SNPs linked to arsenic-annotated genes. The SVM score predicted for each nsSNP substitution was extracted from the SNPs3D page.

**Table 1.** Summary of datasets.

| Dataset | Count |
|---|---|
| Arsenic-annotated genes Retrieved from CTD | 1,604 |
| Arsenic-annotated genes with nsSNPs in SNPs3D | 1,224 |
| Non-synonymous SNPs | 5,811 |
| Arsenic-annotated protein isoforms | 1,872 |
| Amino acid substitutions | 8,992 |
| Deleterious substitutions | 3,700 |
| Non-deleterious substitutions | 5,292 |
| High confidence deleterious substitutions* | 2,739 |
| High confidence non-deleterious substitutions** | 4,191 |

SVM sequence profile score: *SVM <–0.5; **SVM >0.5

the dataset is presented in Table 1. In order to facilitate selection of nsSNPs according to confidence of SVM score, we classed substitutions into categories with a 0.5 interval (Table 2).

## Arsenic responsiveness based on substitution to or from cysteine residues

A total of 196 nsSNPs linked to 144 genes and 225 protein isoforms were observed to cause substitutions to cysteine residues. In the case of substitutions from cysteine residues, 92 nsSNPs linked to 79 genes and 122 protein isoforms were observed. In the protein isoforms analyzed, the substitutions to or from cysteine residues were restricted to the following six amino acid residues: Phenylalanine (F), Glycine (G), Arginine (R), Serine (S), Tryptophan (W) and Tyrosine (Y).

Four classes of nsSNPs were identified on the basis of significant deleterious and non-deleterious effects on protein function as well as SNP-associated residue changes to or from cysteine (Table 3). The 111 nsSNPs that resulted in non-deleterious substitutions to or from cysteine are candidates for evaluating increased or decreased responsiveness to arsenic, respectively (Supplementary Data). A set of nsSNPs that mutates the residue to or from cysteine with significant impact on protein function and structure (with agreement of both SVM scores for sequence and structure profiles) are presented in Table 4 and Table 5, respectively. All the identifiers for SNP are from the dbSNP and begin with "rs".

## Structural homology modeling

In SNPs3D, breakage of a disulfide bond is assigned to any mutation that replaces a cysteine residue in an S–S bond with a non-cysteine residue. Eight nsSNPs from 6 genes (9 protein isoforms) were identified to result in breakage of a disulfide bond (Table 6). The impact of these nsSNPs on protein stability has pointed us to potential regions of arsenic binding to vicinal (neigboring) cysteines. Since, we are interested in arsenic-induced skin cancer, we further analyzed the

**Table 2.** Distribution of SVM scores for nsSNP substitutions.

| SVM score class* | Non-deleterious | | Deleterious | |
|---|---|---|---|---|
| | Frequency | Percent frequency | Frequency | Percent frequency |
| 0.0–0.5 | 1083 | 20.46% | 977 | 26.41% |
| 0.5–1.0 | 1186 | 22.41% | 831 | 22.46% |
| 1.0–1.5 | 1170 | 22.11% | 627 | 16.95% |
| 1.5–2.0 | 826 | 15.61% | 448 | 12.11% |
| 2.0–2.5 | 560 | 10.58% | 348 | 9.41% |
| 2.5–3.0 | 246 | 4.65% | 255 | 6.89% |
| 3.0–3.5 | 126 | 2.38% | 132 | 3.57% |
| 3.5–4.0 | 49 | 0.93% | 63 | 1.70% |
| 4.0–4.5 | 25 | 0.47% | 18 | 0.49% |
| 4.5–5.0 | 12 | 0.23% | 1 | 0.03% |
| 5.0–5.5 | 2 | 0.04% | | |
| 5.5–6.0 | 3 | 0.06% | | |
| 6.0–6.5 | 3 | 0.06% | | |
| 6.5–7.0 | 1 | 0.02% | | |

**Note:** *Absolute values.

**Table 3.** Categories of nsSNPs observed in analysis of arsenic-annotated genes.

| Category | Gene count | SNP count | Protein isoforms |
|---|---|---|---|
| From Cys, deleterious | 45 | 53 | 67 |
| From Cys, non-deleterious | 36 | 39 | 58 |
| To Cys, deleterious | 101 | 130 | 145 |
| To Cys, non-deleterious | 63 | 72 | 108 |

stratum corneum chymotryptic serine protease KLK7 (kallikrein-related peptidase 7) for annotated structural consequences of SNP marker for reduced responsiveness, the effects of arsenic on expression as well as distribution of cysteine residues. A screenshot of SNPs3D page on structural impact of nsSNP rs17855561 is presented in Figure 3. The nsSNP rs17855561 in both KLK7 protein isoforms is predicted to result in breakage of a disulfide bond and potentially reducing responsiveness to arsenic by changing the Cys in position 226 to Tryptophan (W) (Table 6). Furthermore, according to data extracted from Bae et al[41] by CTD curators, sodium arsenite results in decreased expression of KLK7 mRNA in the virally immortalized human keratinocyte cell line RHEK-1. Structures of KLK7 from protein sequences NP_005037 and NP_644806 were generated using high resolution X-ray crystal structure of human kallikrein (PDB ID: 2QXI). Structural homology models of wild type KLK7 structure revealed six cysteine pairs 36–165; 55–71; 137–239; 144–211; 176–190 and 201–226 (Fig. 4).

## Construction of sentence collection from PubMed abstracts on arsenicals

In order to facilitate further studies on these identified genes, single nucleotide polymorphisms and other aspects of arsenic, we have segmented 16,057 PubMed abstracts into a collection of 108,235 sentences. The abstracts were selected based on annotation of the abstract in PubMed with at least one of the Medical Subject Heading (MeSH) terms: arsenic or arsenicals. An Arsenic Sentence Database that facilitates query of the sentences with keywords as well as retrieval of sentences for specific PubMed abstracts is available at http://compbio.jsums.edu/arsenic_pubmed.

The utility of the database was demonstrated by a search for "GSTO1" the symbol for gene encoding the enzyme glutathione S-transferase omega 1 which catalyzes the monomethyl arsenate reduction, the rate-limiting step for inorganic arsenic biotransformation in humans.[42,43] A cluster of 80 sentences containing the symbol were retrieved from the database. Furthermore, a subset of the GSTO1 sentences and containing the

**Table 4.** Candidate nsSNPs that increase protein's arsenic responsiveness through amino acid change to cysteine.

| Entrez gene | Gene symbol | nsSNP | Protein isoform | Mutation | SVM sequence score | SVM structure score | Impac on protein stability** |
|---|---|---|---|---|---|---|---|
| 834 | CASP1 | 3203613 | NP_150635 | S33C | 2.76 | 1.33 | PS |
| 834 | CASP1 | 3203613 | NP_150634 | S126C | 2.36 | 1.33 | PS |
| 11200 | CHEK2 | 28909981 | NP_665861 | S471C | 2.20 | 1.13 | PS; HBL |
| 834 | CASP1 | 3203613 | NP_150636 | S33C | 1.99 | 1.33 | PS |
| 8644 | AKR1C3 | 35575889 | NP_003730 | R170C | 1.82 | 1.59 | |
| 55713 | ZNF334 | 41283032 | NP_955473 | R237C | 1.71 | 1.08 | PS |
| 55713 | ZNF334 | 41283032 | NP_060572 | R275C | 1.68 | 1.08 | PS |
| 6389 | SDHA | 1041948 | NP_004159 | S346C | 1.62 | 0.92 | HBL |
| 2877 | GPX2 | 17880492 | NP_002074 | R146C | 1.61 | 0.84 | SBL |
| 983 | CDC2 | 8755 | NP_203698 | R59C | 1.53 | 1.22 | PS |
| 10935 | PRDX3 | 11554910 | NP_054817 | Y53C | 1.19 | 0.99 | PS |
| 10935 | PRDX3 | 11554910 | NP_006784 | Y71C | 1.14 | 0.99 | PS |
| 5265 | SERPINA1 | 1802962 | NP_000286 | S325C | 0.88 | 1.37 | PS |
| 5265 | SERPINA1 | 1802962 | NP_001002235 | S325C | 0.88 | 1.37 | PS |
| 5265 | SERPINA1 | 1802962 | NP_001002236 | S325C | 0.88 | 1.37 | PS |
| 983 | CDC2 | 8755 | NP_001777 | R59C | 0.81 | 1.22 | PS |
| 4233 | MET | 34589476 | NP_000236 | R970C | 0.77 | 0.99 | PS |

***Abbreviations:** PS, On the protein surface; HBL, hydrogen bond lost; SBL, saltbridge lost.

**Table 5.** Candidate nsSNPs that reduce protein's arsenic responsiveness through amino acid change from cysteine.

| Entrez gene | Gene symbol | nsSNP | Protein isoform | Mutation | SVM sequence score | SVM structure score | Impact on protein stability* |
|---|---|---|---|---|---|---|---|
| 23660 | ZKSCAN5 | 28411998 | NP_055384 | C579W | –3.63 | –0.84 | BP; OP |
| 23660 | ZKSCAN5 | 28411998 | NP_659570 | C579W | –3.63 | –0.84 | BP; OP |
| 7465 | WEE1 | 17854721 | NP_003381 | C379R | –3.62 | –1.18 | |
| 9753 | ZSCAN12 | 2232432 | NP_001034732 | C332R | –3.59 | –1.11 | ESR |
| 9446 | GSTO1 | 45529437 | NP_004823 | C32Y | –3.53 | –1.20 | OP |
| 5650 | KLK7 | 17855561 | NP_644806 | C226W | –2.44 | –1.20 | OP; BDB |
| 5650 | KLK7 | 17855561 | NP_005037 | C226W | –2.43 | –1.20 | OP; BDB |
| 5879 | RAC1 | 7673785 | NP_008839 | C157R | –2.06 | –1.24 | BC; OP |
| 3945 | LDHB | 3575 | NP_002291 | C294Y | –1.83 | –1.05 | BP; OP |
| 5879 | RAC1 | 7673785 | NP_061485 | C176R | –1.77 | –1.52 | OP |
| 5265 | SERPINA1 | 8350 | NP_000286 | C256W | –1.76 | –0.51 | BP |
| 5265 | SERPINA1 | 8350 | NP_001002235 | C256W | –1.76 | –0.51 | BP |
| 5265 | SERPINA1 | 8350 | NP_001002236 | C256W | –1.76 | –0.51 | BP |
| 2868 | GRK4 | 35824641 | NP_892027 | C215R | –1.66 | –0.79 | ESR; OP |
| 2868 | GRK4 | 35824641 | NP_001004057 | C215R | –1.55 | –0.79 | ESR; OP |
| 835 | CASP2 | 11551881 | NP_116764 | C370F | –1.54 | –0.99 | OP |
| 2868 | GRK4 | 35824641 | NP_001004056 | C183R | –1.51 | –0.80 | ESR; OP |
| 10465 | PPIH | 11550298 | NP_006338 | C131R | –1.28 | –1.41 | BC; ESR; OP |
| 3107 | HLA-C | 41563216 | NP_002108 | C125R | –1.15 | –1.28 | BC; OP; BDB |
| 3107 | HLA-C | 41543517 | NP_002108 | C188W | –0.81 | –0.86 | BP; OP; BDB |

***Abbreviations:** BP, buriedpolar; BC, buriedcharged; BDB, breakage of a disulfide bond; ESR, electrostaticrepulsion; OP, overpacking; SBL, saltbridge lost.

word "polymorphism" allowed us to identify 13 sentences from 7 PubMed abstracts (Table 7). These abstracts were on genetic variation observed in GSTO1 and other genes involved in arsenic metabolism.

## Discussion

Since amino acid substitutions to or from cysteine residues attributed to SNPs might be a determinant of responsiveness of target proteins to arsenic and possibly arsenic-induced skin cancer, we undertook to prioritize genes and SNPs to understand keratinocyte carcinogenesis resulting from arsenic exposure. Our analysis of the functional and structural impacts of 5,811 nsSNPs associated with 1,224 putative arsenic responsive genes identified i) 196 candidate nsSNPs for increased arsenic responsiveness by substitutions to cysteine residues for 144 genes; ii) 92 candidate nsSNP for decreased arsenic responsiveness by substitutions from cysteine residues for 79 genes; iii) nsSNP rs17855561 that results in breakage of a disulfide bond, as candidate marker for reduced arsenic responsiveness in KLK7, a secreted serine protease that has been demonstrated to participate in normal shedding of the skin[44] and iv) 6 pairs of vicinal cysteines in KLK7 protein.

**Table 6.** Non-synonymous SNPs that predict potential region of arsenic-binding to vicinal cysteines.

| Gene symbol | nsSNP | Protein isoform | Mutation |
|---|---|---|---|
| FST | 1127760 | NP_006341 | C239S |
| FST | 1127760 | NP_037541 | C239S |
| HLA-C | 41563216 | NP_002108 | C125R |
| HLA-C | 41543517 | NP_002108 | C188W |
| HLA-C | 41562916 | NP_002108 | C188F |
| IL4 | 4986964 | NP_000580 | C27R |
| IL4 | 4986964 | NP_758858 | C27R |
| KLK7 | 17855561 | NP_005037 | C226W |
| KLK7 | 17855561 | NP_644806 | C226W |
| TFRC | 9852079 | NP_003225 | C363W |
| TLR4 | 2770145 | NP_612564 | C306W |

The bioinformatics analysis pipeline identified genes with evidence for potential SNP-induced increased or decreased responsiveness to arsenic. To the best of our knowledge this report is the first large-scale analysis of SNP-induced substitutions evaluating the abundance of cysteines in putative protein targets of arsenic. A recent large-scale analysis of the curation of chemical-gene relationships from

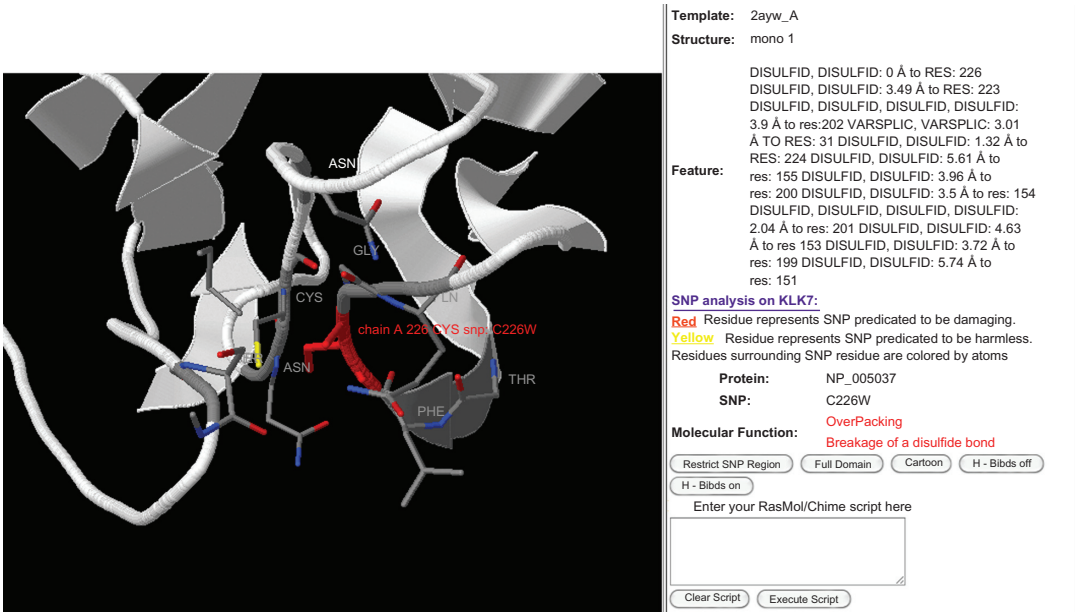



**Figure 3.** Screenshot of predicted structural impact of nsSNP rs17855561 on human kallikrein-7 preproprotein (NP_005037). The nsSNP results in steric strain (over-packing) and breakage of a disulfide bond by changing Cysteine (C) residue in position 226 to a Tryptophan (W).

biomedical literature has provided over 1,400 genes whose activity were perturbed by arsenic in a variety of conditions and/or cell types.[35] Our analysis extends the curation efforts by CTD by integrating SNP data that could help understand the molecular mechanisms of arsenic action in diverse cell types including keratinocytes.

We have used the structural impact annotation "breakage of a disulfide bond" as an evidence of the presence of potential arsenic-binding vicinal cysteines in a protein sequence. The function and structure of proteins are often determined by Cys residues since many proteins folding are dependent on disulfide bonds.[33] Arsenic binding to target protein depends on the number, accessibility and relative positioning of Cys residues.[45] Furthermore, the ability for trivalent arsenicals to bind to thiol groups of biomolecules is an accepted mechanism for being more toxic than pentavalent arsenicals. The mRNA from the tissue serine protease KLK7 that was identified by our pipeline was down-regulated by arsenic in human keratinocyte cell line RHEK-1.[41] The proposed functions of KLK7 in the normal skin physiology include i) activating interleukin 1 beta (IL-1b) and ii) basal permeability barrier function of stratum corneum by degrading two major lipid processing enzymes beta-glucocerebrosidase and acidic sphingomyelinase.[44,46,47] The crucial function of KLK7 in normal skin function and potential perturbation by arsenic justifies a need to determine the potential energy of each Cys residues in KLK7 combined with their proximity to enzyme active sites or other functional regions of the protein. We hypothesize that arsenic binds to at least one of the 6 pairs of vicinal cysteines resulting in conformational changes that down regulate KLK7 function. Our hypothesis for KLK7 can be tested using similar experiments conducted to determine the role of the 8 Cys residues in arsenic binding for human beta-tubulin.[32] In this investigation, we have verified the sequence-based prediction of vicinal cysteine with structural homology modeling.

The functional implications of SNP modified polypeptides of KLK7 warrant further investigation. Our text mining approach using the gene symbol GSTO1 as a search term in a collection of over 100,000 sentences from over 16,000 PubMed abstracts retrieved publications that could guide further research on the impact of arsenic as well as SNP-induced polymorphism on KLK7 function. In the case of the two enzymes Glutathione S-transferase omega 1 and omega 2 (GSTO1 and GSTO2) that catalyze monomethyl arsenate reduction, variant allozymes have been shown to degrade more rapidly than their respective wild type allozymes.[48] Similar protein degradation experiments for KLK7, could

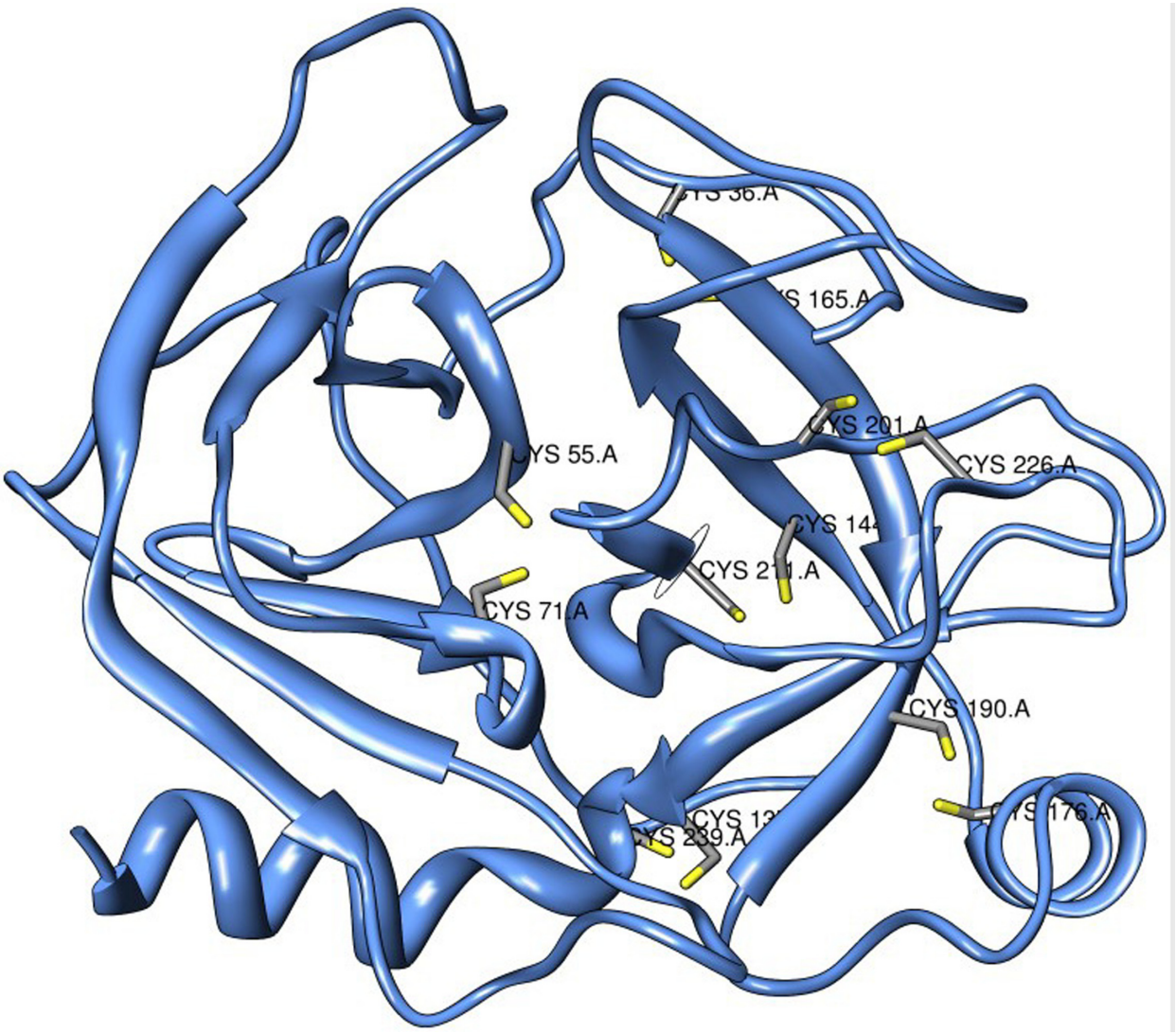



**Figure 4.** Predicted structure of human kallikrein-7 preproprotein (NP_005037). The vicinal cysteine pairs are 36–165; 55–71; 137–239; 144–211; 176–190; 201–226. The homology structure shows that the Cys201 pairs with Cys226. Arsenic is known to reduce the expression of KLK7 in an epidermal cell line.

unravel the impact of molecular differences resulting in susceptibility of arsenic-induced skin cancer in arsenicosis-endemic populations.

## Conclusions

Single nucleotide polymorphisms (SNPs) can alter the physico-chemical properties of proteins. The susceptibility to arsenic-induced diseases as well as response to arsenic-based drugs has been linked to single nucleotide polymorphisms. Our analysis identified non-synonymous SNPs that could be used to evaluate responsiveness of a protein target to arsenic. Furthermore, an epidermal expressed serine protease KLK7 with crucial function in normal skin physiology was prioritized on the basis of abundance of vicinal cysteines to understand arsenic-induced keratinocyte carcinogenesis.

## Acknowledgements

Research Centers in Minority Institutions (RCMI)—Center for Environmental Health at Jackson State University (NIH-NCRR 2G12RR013459); Mississippi NSF-EPSCoR Grant Awards (EPS-0556308, EPS-0903787); Mississippi Computational Biology Consortium Seed Grant Program; Pittsburgh Supercomputing Center's National Resource for Biomedical Supercomputing (T36 GM008789); U.S. Department of Homeland Security Science and Technology Directorate (2007-ST-104-000007; 2009-ST-062-

**Table 7.** Cluster of sentences from PubMed abstracts on GSTO1 polymorphism.

| Sentence identifier* | Sentence text |
|---|---|
| 18216717_10 | The percent of inorganic arsenic in the urine of 205 Chilean participants showed a bimodal distribution that was not associated with the Ala140Asp, Glu155del or Ala236Val polymorphisms in GSTO1-1. |
| 19686770_5 | Genotyping of CYP2E1, GSTO1 and GSTO2 was determined by polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP). |
| 16638819_4 | We identified 31 and 66 polymorphisms in GSTO1 and GSTO2, respectively, with four nonsynonymous-coding single nucleotide polymorphisms (cSNPs) in each gene. |
| 12928150_5 | We screened two genes responsible for arsenic metabolism, human purine nucleoside phosphorylase (hNP), which functions as an arsenate reductase converting arsenate to arsenite, and human glutathione S-transferase omega 1-1 (hGSTO1-1), which functions as a monomethylarsonic acid (MMA) reductase, converting MMA(V) to MMA(III), to develop a comprehensive catalog of commonly occurring genetic polymorphisms in these genes. |
| 12928150_9 | In hGSTO1-1, 33 polymorphisms were observed. |
| 12928150_11 | In contrast to hNP, in which the IA group was more polymorphic than the EA group, in hGSTO1-1 the EA group was more polymorphic than the IA group, which had only 1 polymorphism with a frequency $>$10%. |
| 19635583_4 | This study was conducted to investigate the relationship between polymorphisms in the GSTO1 and GSTO2 genes and arsenic metabolism and oxidative stress status in Chinese populations chronically exposed to different levels of arsenic in drinking water. |
| 19635583_5 | Two polymorphisms (GSTO1*A140D and GSTO2*N142D) with relatively higher mutation frequencies in the Chinese population were determined by polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP). |
| 19635583_8 | Multivariate analysis revealed that there was no association between the urinary profile or oxidative stress status and the polymorphism of GSTO1*A140D or GSTO2*N142D. |
| 19635583_9 | Collectively, polymorphisms in GSTO1 or GSTO2 do not appear to contribute to the large individual variability in arsenic metabolism or susceptibility to arsenicosis. |
| 14680363_3 | To understand this variability, we studied the relationship between polymorphisms in the gene for human monomethylarsonic acid (MMA(V)) reductase/hGSTO1 and the urinary arsenic profiles of individuals chronically exposed to arsenic in their drinking water. |
| 14680363_9 | These two subjects had the same unique polymorphisms in hGSTO1 in that they were heterozygous for E155del and Glu208 Lys. |
| 18414634_9 | RESULTS: Among four candidate genes, PNP, As3MT, GSTO1, and GSTO2, we found that distribution of three exonic polymorphisms, His20His, Gly51Ser, and Pro57Pro of PNP, was associated with arsenicism. |

**Note:** *Sentence identifier consists of PubMed identifier (PMID) and the location of the sentence in the abstract with abstract title as the first sentence.

000014; 2009-ST-104-000021) and NIH RIMI Grant 1P20MD002725-01 to Tougaloo College. We thank Dr. Robert Rice and Dr. Susan Bridges for their suggestions. Disclaimer: The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

## Disclosures

This manuscript has been read and approved by all authors. This paper is unique and not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

## References

1. Hall M, Chen Y, Ahsan H, et al. Blood arsenic as a biomarker of arsenic exposure: results from a prospective study. *Toxicology*. 2006;225:225–33.
2. Tchounwou PB, Patlolla AK, Centeno JA. Carcinogenic and systemic health effects associated with arsenic exposure—a critical review. *Toxicologic Pathology*. 2003;31:575–88.
3. Ding W, Hudson LG, Liu KJ. Inorganic arsenic compounds cause oxidative damage to DNA and protein by inducing ROS and RNS generation in human keratinocytes. *Molecular and Cellular Biochemistry*. 2005;279:105–12.
4. Pi J, He Y, Bortner C, et al. Low level, long-term inorganic arsenite exposure causes generalized resistance to apoptosis in cultured human keratinocytes: potential role in skin co-carcinogenesis. *International Journal of Cancer*. 2005;116:20–6.
5. Tchounwou PB, Centeno JA, Patlolla AK. Arsenic toxicity, mutagenesis, and carcinogenesis—a health risk assessment and management approach. *Molecular and Cellular Biochemistry*. 2004;255:47–55.

6. Mead MN. Arsenic: in search of an antidote to a global poison. *Environmental Health Perspectives*. 2005;113:A378–86.
7. McDonald C, Hoque R, Huda N, Cherry N. Prevalence of arsenic-related skin lesions in 53 widely-scattered villages of Bangladesh: an ecological survey. *Journal of Health and Population Nutrition*. 2006;24:228–35.
8. Meliker JR, Wahl RL, Cameron LL, Nriagu JO. Arsenic in drinking water and cerebrovascular disease, diabetes mellitus, and kidney disease in Michigan: a standardized mortality ratio analysis. *Environmental Health*. 2007;6:4.
9. U.S. EPA (U.S. Environmental Protection Agency). National primary drinking water regulations: arsenic and clarifications to compliance and new source contaminants monitoring. Final Rule. *Federal Register*. 2001;66: 6976–7066.
10. Hughes MF. Biomarkers of exposure: a case study with inorganic arsenic. *Environmental Health Perspectives*. 2006;114:1790–6.
11. Centeno JA, Mullick FG, Martinez L, et al. Pathology related to chronic arsenic exposure. *Environmental Health Perspectives*. 2002;110 Suppl 5: 883–6.
12. Lansdown AB. Physiological and toxicological changes in the skin resulting from the action and interaction of metal ions. *Critical Reviews in Toxicology*. 1995;25:397–462.
13. Ralph SJ. Arsenic-based antineoplastic drugs and their mechanisms of action. *Metal Based Drugs*. 2008;2008:260146.
14. Bates MN, Smith AH, Hopenhayn-Rich C. Arsenic ingestion and internal cancers: a review. *American Journal of Epidemiology*. 1992;135:462–76.
15. Yu HS, Liao WT, Chai CY. Arsenic carcinogenesis in the skin. *Journal of Biomedical Science*. 2006;13:657–66.
16. Soignet SL, Maslak P, Wang ZG, et al. Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide. *New England Journal of Medicine*. 1998;339:1341–8.
17. Yedjou C, Thuisseu L, Tchounwou C, et al. Ascorbic acid potentiation of arsenic trioxide anticancer activity against acute promyelocytic leukemia. *Archives of Drug Information*. 2009;2:59–65.
18. Yedjou CG, Moore P, Tchounwou PB. Dose- and time-dependent response of human leukemia (HL-60) cells to arsenic trioxide treatment. *International Journal of Environmental Research and Public Health*. 2006;3:136–40.
19. Yedjou CG, Tchounwou PB. In-vitro cytotoxic and genotoxic effects of arsenic trioxide on human leukemia (HL-60) cells using the MTT and alkaline single cell gel electrophoresis (Comet) assays. *Molecular and Cellular Biochemistry*. 2007;301:123–30.
20. Yedjou CG, Rogers C, Brown E, Tchounwou PB. Differential effect of ascorbic acid and n-acetyl-L-cysteine on arsenic trioxide-mediated oxidative stress in human leukemia (HL-60) cells. *Journal of Biochemical and Molecular Toxicology*. 2008;22:85–92.
21. Raghu KG, Yadav GK, Singh R, et al. Evaluation of adverse cardiac effects induced by arsenic trioxide, a potent anti-APL drug. *Journal of Environmental Pathology, Toxicology and Oncology*. 2009;28:241–52.
22. Breton CV, Zhou W, Kile ML, et al. Susceptibility to arsenic-induced skin lesions from polymorphisms in base excision repair genes. *Carcinogenesis*. 2007;28:1520–5.
23. Yamada R. Primer: SNP-associated studies and what they can teach us. *Nature Clinical Practice Rheumatology*. 2008;4:210–7.
24. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. 2001;29:308–11.
25. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*. 2006; 7:61–80.
26. Clifford RJ, Edmonson MN, Nguyen C, et al. Bioinformatics tools for single nucleotide polymorphism discovery and analysis. *Annals of the New York Academy of Sciences*. 2004;1020:101–9.
27. Kang JQ, Macdonald RL. Making sense of nonsense GABA(A) receptor mutations associated with genetic epilepsies. *Trends in Molecular Medicine*. 2009;15:430–8.
28. Mukherjee B, Salavaggione OE, Pelleymounter LL, et al. Glutathione S-transferase omega 1 and omega 2 pharmacogenomics. *Drug Metabolism and Disposition: The Biological Fate of Chemical*. 2006;34:1237–46.
29. Shen ZX, Chen GQ, Ni JH, et al. Use of arsenic trioxide (As2O3) in the treatment of acute promyelocytic leukemia (APL): II. Clinical efficacy and pharmacokinetics in relapsed patients. *Blood*. 1997;89:3354–60.
30. Mizumura A, Watanabe T, Kobayashi Y, Hirano S. Identification of arsenite- and arsenic diglutathione-binding proteins in human hepatocarcinoma cells. *Toxicology and Applied Pharmacology*. 2010;242:119–25.
31. Ramadan D, Rancy PC, Nagarkar RP, Schneider JP, Thorpe C. Arsenic (III) species inhibit oxidative protein folding in vitro. *Biochemistry*. 2009; 48:424–32.
32. Zhang X, Yang F, Shim JY, et al. Identification of arsenic-binding proteins in human breast cancer cells. *Cancer Letters*. 2007;255:95–106.
33. Song X, Geng Z, Zhu J, et al. Structure-function roles of four cysteine residues in the human arsenic (+3 oxidation state) methyltransferase (hAS3MT) by site-directed mutagenesis. *Chemico-Biological Interactions*. 2009;179:321–8.
34. He X, Ma Q. Critical cysteine residues of Kelch-like ECH-associated protein 1 in arsenic sensing and suppression of nuclear factor erythroid 2-related factor 2. *Journal of Pharmacology and Experimental Therapeutics*. 2010;332:66–75.
35. Davis AP, Murphy CG, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study. *BMC Medical Genomics*. 2008;1:48.
36. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*. 2006;7:166.
37. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 2009;37:D5–15.
38. Yue P, Moult J. Identification and analysis of deleterious human SNPs. *Journal of Molecular Biology*. 2006;356:1263–74.
39. Marti-Renom MA, Stuart AC, Fiser A, et al. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*. 2000;29:291–325.
40. Kouranov A, Xie L, de la Cruz J, et al. The RCSB PDB information portal for structural genomics. *Nucleic Acids Research*. 2006;34:D302–5.
41. Bae DS, Hanneman WH, Yang RS, Campain JA. Characterization of gene expression changes associated with MNNG, arsenic, or metal mixture treatment in human keratinocytes: application of cDNA microarray technology. *Environmental Health Perspectives*. 2002;110 Suppl 6:931–41.
42. Mukherjee B, Salavaggione OE, Pelleymounter LL, et al. Glutathione S-transferase omega 1 and omega 2 pharmacogenomics. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*. 2006;34:1237–46.
43. Ferrario D, Croera C, Brustio R, et al. Toxicity of inorganic arsenic and its metabolites on haematopoietic progenitors “in vitro”: comparison between species and sexes. *Toxicology*. 2008;249:102–8.
44. Ekholm IE, Brattsand M, Egelrud T. Stratum corneum tryptic enzyme in normal epidermis: a missing link in the desquamation process? *Journal of Investigative Dermatology*. 2000;114:56–63.
45. Kitchin KT, Wallace K. Arsenite binding to synthetic peptides based on the Zn finger region and the estrogen binding region of the human estrogen receptor-alpha. *Toxicology and Applied Pharmacology*. 2005;206:66–72.
46. Pampalakis G, Sotiropoulou G. Tissue kallikrein proteolytic cascade pathways in normal physiology and cancer. *Biochimica et Biophysica Acta: Protein Structure and Molecular Enzymology*. 2007;1776:22–31.
47. Hachem JP, Man MQ, Crumrine D, et al. Sustained serine proteases activity by prolonged increase in pH leads to degradation of lipid processing enzymes and profound alterations of barrier function and stratum corneum integrity. *Journal of Investigative Dermatology*. 2005;125:510–20.
48. Mukherjee B, Salavaggione OE, Pelleymounter LL, et al. Glutathione S-transferase omega 1 and omega 2 pharmacogenomics. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*. 2006;34:1237–46.

## Supplementary Data

BBI-4-Isokpehi-Supplementary.xls