Niron Sukumar[a], Erika Scott[a], Apostolos Dimitromanolakis[a], Alexandra Misiak, Ioannis Prassas, Eleftherios P. Diamandis and Ana Konvalinka*

Mining for single nucleotide variants (SNVs) at the kallikrein locus with predicted functional consequences

Abstract: Kallikreins (KLKs) are a group of 15 serine proteases encoded by the KLK locus on chromosome 19. Certain single nucleotide variants (SNVs) within the KLK locus have been linked to human disease. Next-generation sequencing of large human cohorts enables reexamination of genomic variation at the KLK locus. We aimed to identify all KLK-related SNVs and examine their impact on gene regulation and function. To this end, we mined KLK SNVs across Ensembl and Exome Variant Server, with exome-sequencing data from 6503 individuals. PolyPhen-2-based prediction of damaging SNVs and population frequencies of these SNVs were examined. Damaging SNVs were plotted on protein sequence and structure. We identified 4866 SNVs, the largest number of KLK-related SNVs reported. Fourteen percent of noncoding SNVs overlapped with transcription factor binding sites. We identified 602 missense coding SNVs, among which 148 were predicted to be damaging. Nine missense SNVs were common (>1% frequency) and displayed significantly different frequencies between European-American and African-American populations. SNVs predicted to be damaging appeared to alter tertiary structure of KLK1 and KLK6. Similarly, these missense SNVs may affect KLK function, resulting in disease phenotypes. Our study represents a mine of information for those studying KLK-related SNVs and their associations with diseases.

[a]These authors contributed equally to this manuscript.
***Corresponding author: Ana Konvalinka,** Division of Nephrology, Department of Medicine, Toronto General Hospital, University Health Network, 8N – 859, 200 Elizabeth Street, Toronto, Ontario M5G 2C4, Canada, e-mail: ana.konvalinka@mail.utoronto.ca
Niron Sukumar, Erika Scott, Apostolos Dimitromanolakis, Alexandra Misiak and Ioannis Prassas: Department of Laboratory Medicine and Pathobiology, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada
Eleftherios P. Diamandis: Department of Laboratory Medicine and Pathobiology, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada;and Department of Clinical Biochemistry, Toronto General Hospital, University Health Network, Toronto, Ontario, Canada

Keywords: data mining; kallikrein; missense; nonsense; single nucleotide polymorphism; single nucleotide variant.

DOI 10.1515/hsz-2014-0136
Received February 14, 2014; accepted July 3, 2014

Introduction

The kallikrein (KLK) proteins are encoded by a locus of 15 genes located on chromosome 19q13.3–13.4 (Riegman et al., 1989; Borgono and Diamandis, 2004). The majority of these enzymes (KLKs 1, 2, 4, 5, 6, 8, 10, 11, 12, 13, and 14) display trypsin-like serine protease activity and contain acidic residues such as aspartic and glutamic acid in their active sites, whereas KLKs 3, 7, and 9 are chymotrypsin-like serine proteases with hydrophobic residues such as tyrosine and phenylalanine in their active sites. The acidic and basic residues are located within the S1 specificity pocket of the proteases. All KLKs contain the well-conserved serine protease catalytic triad, consisting of histidine, aspartic acid, and serine. KLK1 contains additional exosites, allowing it to bind kininogen. It is now established that KLKs mediate a plethora of crucial physiological functions in several tissues, including in skin, breast, prostate, pancreas, and brain tissue. As part of the general extracellular web of proteases, KLK activity is tightly regulated by endogenous inhibitors, activating proteolytic cascades and regulatory feedback that maintain balanced KLK activity (Borgono and Diamandis, 2004). Disruption of this balance and concomitant dysregulation of KLK function has been linked to several diseases, including skin diseases (Eissa and Diamandis, 2008), neurodegenerative diseases, chronic inflammation, diabetes mellitus (Paliouras and Diamandis, 2006; Batra et al., 2012), and cancer (Bharaj et al., 2002; Ahn et al., 2008; Antoniou et al., 2010; Batra et al., 2011). The recent advances in sequencing technologies have brought

to light numerous variations, such as single nucleotide variants (SNVs), of several KLK genes, as major contributors to KLK dysregulation *in vivo*. For example, SNVs rs17632542, rs62113212, and rs62113214 in KLK3 have all been associated with prostate cancer (Kote-Jarai et al., 2011; Parikh et al., 2011). Furthermore, a missense SNV rs5515 discovered in *KLK1* gene is known to result in a change in amino acid residue from arginine (R) to histidine (H), which, in turn, leads to a marked decrease in KLK1 activity in tissue and urine (Slim et al., 2002; Waeckel et al., 2013). Individuals harboring this particular SNV display endothelial dysfunction and a defect in flow-dependent vasodilation (Bergaya et al., 2004a), as well as a propensity to develop hypertension (Margolius et al., 1976; Bergaya et al., 2004a). A crystallographic KLK1 structure demonstrated that the presence of histidine alters the substrate-binding properties of KLK1 to kininogen, and this explains the diminished activity of the enzyme (Katz et al., 1998). Further proof of concept came from a mouse knockout of *KLK1*, which displayed similar phenotypic features, verifying the pathophysiological significance of this particular SNV (Bergaya et al., 2004b). Based on these findings, we decided to expand current knowledge of potentially damaging SNVs within the KLKs (as predicted by the functional SNV class and its impact on KLK structure) and investigate their frequencies in certain populations and their potential to result in clinical phenotypes.

Single nucleotide polymorphisms (SNPs) are variations at a single nucleotide position and are considered to be the most frequent type of variation within the human genome, comprising about 0.1% of the human genome (Collins et al., 1998; International HapMap et al., 2007). Rare SNPs are present at a frequency above 1% in the general population, whereas common SNPs are present at a frequency above 5–10% (Kruglyak and Nickerson, 2001; Ladiges et al., 2004). In this study, we use the term single nucleotide variation (SNV) to denote variations at a single nucleotide position regardless of their frequencies. Previous studies have examined SNVs in the KLK locus (Goard et al., 2007; Batra et al., 2012). These studies mined databases such as dbSNP and 1000 Genomes to discover SNVs at the KLK locus and contributed valuable information. However, to our knowledge, none of the studies to date have linked functional SNV classes at the KLK locus with potential structural defects, nor have examined the frequencies of potentially damaging KLK SNVs in various human populations. The availability of the exome sequencing project data, which includes 6503 individuals, enables a new look at the genomic variation at the KLK locus.

Results

SNVs at the human kallikrein locus

Our first goal was to mine several open access public databases (the Exome Variant Server, the database of Single Nucleotide Polymorphisms (dbSNP), the International HapMap project, the 1000 Genomes Project, Ensembl Genome Browser, and the Catalogue of Somatic Mutations in Cancer (COSMIC) database) and compile all data in a unique catalog of KLK-related SNVs. The Exome Variant Server (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA http://evs.gs.washington.edu/EVS/ [June 2013]) and the Ensembl Genomic Database were chosen as the most extensive nonredundant sources. We thus based our mining on these two databases. We found a total of 4866 SNVs in the 15 kallikreins. Among these, there were 98 SNVs with no rs-identifier. The number of SNVs in each KLK did not display marked differences (Figure 1A and Table 1). More SNVs were noted in KLKs 2, 3, 5, 7, and 13, while fewer were found in KLK9. Compared to prior studies (Goard et al., 2007; Batra et al., 2012), we expanded the number of identified SNVs, as a result of the availability of sequencing information from larger cohorts of individuals. The total number of our identified SNVs is similar to that recently described by Batra et al., who identified 4331 KLK SNVs. Older studies, such as the one by Goard et al., reported on 1856 KLK SNVs in 2007. Goard and colleagues found a lower number of SNVs per KLK, but much like in our current study, the number of SNVs across all KLKs appeared uniform (Figure 1A). There were 1094 SNVs shared between our study and Goard et al. (Figure 1B). The remaining 762 SNVs discovered by Goard, but not found in our study, mainly represent SNVs in the inter-genic region, which are underrepresented in the Ensembl and the Exome Variant Server. We were unable to plot the results of Batra et al. in Figure 1, given that their SNV data were not publically available. Our complete SNV data are submitted as Supplemental data.

Functional classes of SNVs at the kallikrein locus

Our next objective was to examine the functional classification of these newly discovered SNVs. SNVs were categorized into gene locus region SNVs, splice junction SNVs, intronic SNVs, synonymous SNVs, missense SNVs, and untranslated region SNVs. The gene locus region was defined as the region 2000 base pairs upstream and 500

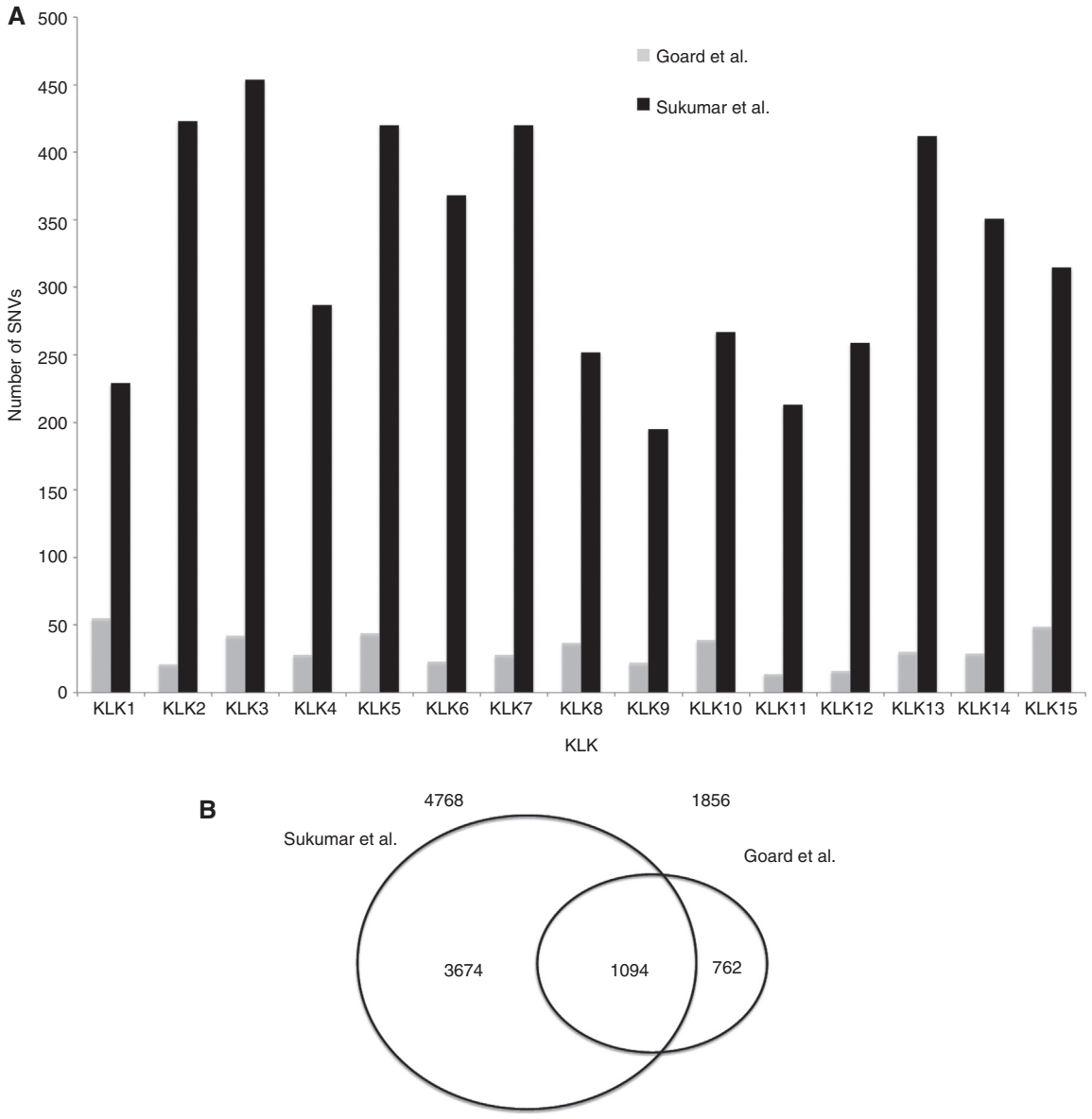


Figure 1 (A) Total SNVs found in each KLK. Comparison of SNV frequencies for each KLK between our study and Goard et al. (B) Total SNVs found as compared to prior studies. Venn diagram portraying SNVs common and unique to our study and Goard et al. The total number of SNVs in our study is smaller than reported, as we also discovered 98 SNVs with no rsIDs, which could not be adequately compared to Goard's studies.

base pairs downstream of a gene, similar to Goard et al. The splice junction region was defined by the Ensembl database as the 2 base pairs at the 5′ and 3′ ends of intronic regions. Synonymous SNVs were defined as the SNVs that did not lead to an amino acid substitution in the resulting polypeptide chain, and missense SNVs were SNVs that led to an amino acid substitution. Intronic SNVs refer to SNVs found in intronic (spliced out) regions of the primary transcript that are not included in the mRNA, and untranslated region SNVs refer to SNVs found in transcribed regions of DNA that are not ultimately translated to protein. We also separately examined whether there were any nonsense SNVs in the 15 kallikreins. Nonsense SNVs were defined as nucleotide substitutions that would lead to the conversion of an amino acid codon to a stop codon. Additionally, we mined a number of SNVs that did not fall into any of these categories.

The bar graph in Figure 2 demonstrates the number of SNVs in each functional class. Most of the SNVs were in the intronic and gene locus regions, a finding consistent across studies. Of all KLK SNVs, we found 38% in gene locus region, 6.4% untranslated, 36.5% intronic, 1.6%

Table 1 Summary of SNVs for each KLK. Numbers of total SNVs, probably damaging SNVs and nonsense SNVs for each KLK.

Kallikrein	Total SNVs	Probably damaging SNVs	Nonsense SNVs
KLK1	230	11	1
KLK2	423	9	0
KLK3	454	15	0
KLK4	287	13	3
KLK5	420	13	0
KLK6	368	6	6
KLK7	420	8	6
KLK8	252	1	5
KLK9	195	7	2
KLK10	267	7	6
KLK11	213	11	4
KLK12	259	10	0
KLK13	412	13	8
KLK14	351	18	13
KLK15	315	6	0
Total:	4866	148	54

splice junction, 5.1% synonymous, and 12.3% missense SNVs. In comparison, Goard and colleagues reported 26.9% SNVs in gene locus region, 9.5% untranslated, 55% intronic, 0.2% splice junction, 3.5% synonymous, and 4.8% missense SNVs. Batra et al. reported 10% of SNVs in the gene locus region, 4% untranslated, 27% intronic, 0% splice junction, 1.5% synonymous, and 2.5% missense SNVs. This study also reported 55% of SNVs as unclassified. The frequencies were similar in our and Goard's studies for all SNV classes, with the exception of missense SNVs, which were more than twice as frequent in our study compared to Goard et al. In comparison to Batra et al., we found a higher proportion of SNVs in the gene locus region and a higher proportion of missense SNVs.

The greatest number of SNVs fell in the noncoding regions. We thus examined common noncoding SNVs with frequency >1%, and there were 1145 of these. Given that noncoding regions may contain regulatory elements, we investigated whether these noncoding SNVs overlapped with transcription factor-binding sites (TFBSs). The noncoding SNVs were divided into intronic, gene locus region, untranslated, and splice junction. UCSC Genome Browser (Kent et al., 2002) was used to identify TFBSs populated by each noncoding SNV, by employing experimentally derived annotation tracks by ChIP-seq (Hudson and Snyder, 2006; Euskirchen et al., 2007). The data in the Browser were generated and analyzed as part of the ENCODE project, a genome-wide consortium with the aim of cataloging all functional elements in the human genome (Rosenbloom et al., 2013). We identified 163 SNVs that were located in TFBSs (Supplemental data). The number of TFs binding to a SNV site ranged from 1 to 47. Most of the SNVs overlapping with TFBSs were in the gene locus region (120). There were also 36 intronic SNVs in TFBSs, six untranslated region SNVs, and one splice junction SNV. As an example, we have displayed all common *KLK3* SNVs together with the associated TFs (Table 2). We chose KLK3 because this is arguably the most studied kallikrein. It is of interest that five TFs with binding sites that overlap with *KLK3* SNVs had been experimentally demonstrated to regulate *KLK3* gene expression in prior studies. These TFs include YY1 (Deng et al., 2009), MYC (Bernard et al., 2003), EGR1 (Yang and

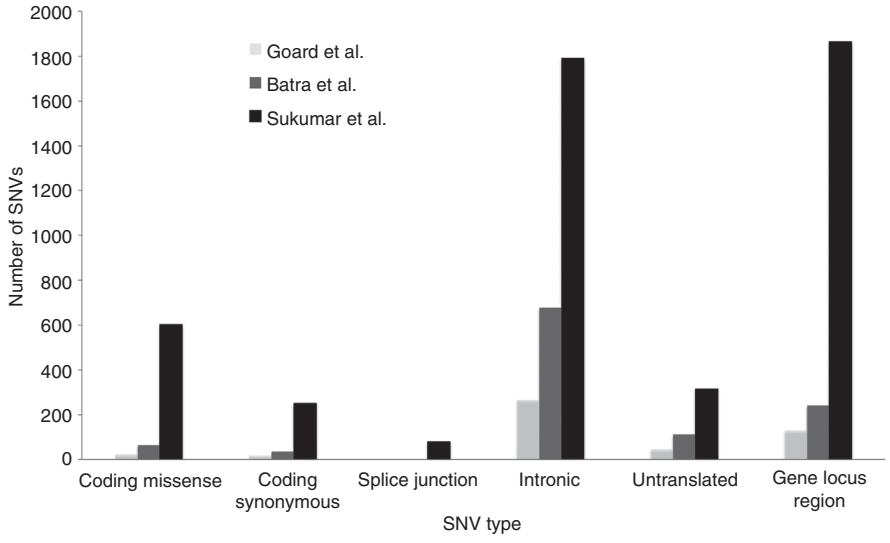


Figure 2 Functional SNV types in all KLKs. Comparison of SNV number for each SNV type between our study, Goard et al., and Batra et al.

Table 2 Common noncoding *KLK3* SNVs in sites of transcription factor binding. Downstream variant gene is in the gene locus region.

Kallikrein gene	rsID	SNV type	Transcription factor
KLK3	rs56297463	Downstream gene variant	EZH2
KLK3	rs11549920	Downstream gene variant	YY1[a]
KLK3	rs4846	Downstream gene variant	NR2F2, JUND, MYC[a], ZBTB7A, MAX, MAZ[a], TEAD4, BHLHE40, EP300, HDAC, GATA2[a], STAT1, ATF1, STAT2, EGR1[a], TAL1
KLK3	rs76975855	Downstream gene variant	CTCF
KLK3	rs8103659	Downstream gene variant	CTCF
KLK3	rs198964	Downstream gene variant	CTCF
KLK3	rs198963	Downstream gene variant	CTCF, RAD21
KLK3	rs198959	Downstream gene variant	JUND
KLK3	rs28419746	Downstream gene variant	JUND

[a]Transcription factors demonstrated to regulate transcription of the *KLK3* gene.

Abdulkadir, 2003), GATA2 (Deng et al., 2009), and MAZ (Tsui et al., 2011). Whether other TFs binding in KLK SNV sites also regulate specific KLK expression, remains for the most part unknown.

Given the potential functional importance of nonsense SNVs, we were particularly interested in the numbers and frequencies of these SNV types among the different KLKs. In theory, heterozygous nonsense SNVs would result in a 50% reduction of the functional protein, while homozygous nonsense SNVs would result in complete absence of the functional protein, which may in turn manifest in a clinical phenotype. We have discovered a total of 54 nonsense SNVs in all 15 KLKs with the highest number present in the *KLK14* gene (Table 1).

Frequencies of missense kallikrein SNVs in European-American and African-American populations

We went on to examine the population frequencies of missense and nonsense SNVs at the KLK locus. We were particularly interested in these types of SNVs because they are more likely to impact protein function. We divided the SNVs into common (>1% population frequency) and rare (<1%), based on their frequency in a population of European Americans (EA) and African Americans (AA). SNV information accessible through the databases was available from the study population of 6503 individuals (4300 EA and 2203 AA). Common SNVs with 1% frequency would thus imply around 65 individuals from this study group. We first examined these common SNVs. By mining the Exome and Ensembl databases, we found 24 common missense SNVs. These common SNVs included three *KLK1* SNVs, two *KLK2* SNVs, four *KLK3* SNVs, two *KLK6* SNVs, one *KLK7* SNV, one *KLK8* SNV, two *KLK9* SNVs, two *KLK10* SNVs, two *KLK12* SNVs, one *KLK13* SNV, and four *KLK14* SNVs. Out of these 24 SNVs, the information about frequencies in EA and AA populations was available for nine SNVs (Table 3 and Supplemental data). All nine of these common SNVs were found at significantly distinct frequencies in EA and AA population (Table 3). Variants rs5517 (*KLK1*), rs17632542 (*KLK3*), rs2075690 (*KLK10*), and rs3745535 (*KLK10*) were significantly more frequent in EA compared to AA population ($p<2.2\times10^{-16}$ for all). In contrast, variants rs5516 (*KLK1*), rs5515 (*KLK1*), rs77760094 (*KLK6*), rs61469141 (*KLK6*), and rs61742847 (*KLK12*) were significantly more frequent in the AA population ($p<2.2\times10^{-16}$ for all except rs5516, which had $p=7.16\times10^{-6}$). Not surprisingly, the frequency of nonsense SNVs was close to 0 in both EA and AA populations.

In addition to common missense SNVs, we identified a much larger number of rare SNVs at the KLK locus. There were 24 rare missense SNVs in *KLK1*, 9 in *KLK2*, 26 in *KLK3*, 4 in *KLK4*, 44 in *KLK5*, 21 in *KLK6*, 24 in *KLK7*, 13 in *KLK8*, 14 in *KLK14*, and 2 in *KLK15* (Supplemental data). Although these SNVs are potentially interesting, their frequencies in the population were extremely low, thus, not allowing us to make any statistical predictions based on the differences in the EA and AA populations.

SNVs predicted to be damaging

After identifying 'common' SNVs with significant population differences between EA and AA ancestry, we wondered about potential functional consequences related to these missense SNVs. Functional consequences of certain SNVs could be predicted by sophisticated algorithms. We examined the PolyPhen-2, SIFT, and Grantham score algorithms to determine the potential for deleterious effects of SNVs on protein function *in silico*. The PolyPhen-2 score

Table 3 Common missense KLK SNVs defined by allele frequency >1%.

Gene	Identifier (rs-ID)	EA allele frequency	AA allele frequency	PolyPhen-2 prediction	*p*-Value
KLK1	rs5517	0.268605	0.196323	Benign	$<2.2e^{-16}$
KLK1	rs5516	0.331163	0.370858	Benign	$7.16e^{-6}$
KLK1	rs5515	0.036047	0.09714	Possibly damaging	$<2.2e^{-16}$
KLK3	rs17632542	0.075349	0.013618	Benign	$<2.2e^{-16}$
KLK6	rs77760094	0.0041860	0.035406	Possibly damaging	$<2.2e^{-16}$
KLK6	rs61469141	0.0004651	0.111665	Possibly damaging	$<2.2e^{-16}$
KLK10	rs2075690	0.450548	0.287969	Benign	$<2.2e^{-16}$
KLK10	rs3745535	0.358625	0.078025	Benign	$<2.2e^{-16}$
KLK12	rs61742847	0.000465	0.056591	Probably damaging	$<2.2e^{-16}$

The frequencies of minor allele in European-American (EA) and African-American (AA) populations are given. *p*-Value is calculated by Pearson χ^2-test for the comparison between EA and AA alleles. PolyPhen-2 based predictions on the functional effects of SNVs are listed.

was chosen because of its comprehensive algorithm, which takes into account the three-dimensional context of the SNV. The PolyPhen-2 score rates SNVs as benign, possibly damaging or probably damaging. We identified 148 total SNVs predicted to be probably damaging in all 15 KLKs (Table 1). The highest number of probably damaging SNVs was 18 in *KLK14*. PolyPhen-2 also predicted 223 SNVs to be possibly damaging (Supplemental data), with the highest number in *KLK15*.

We next examined whether any of the SNVs with significantly different frequencies in EA and AA populations were predicted to be damaging. Indeed, three of the nine common SNVs (rs5515 in *KLK1*, rs77760094 in *KLK6*, rs61469141 in *KLK6*) were possibly damaging, and one common SNV (rs61742847) in *KLK12* was predicted to be probably damaging (Table 3). The frequencies of homozygotes with these damaging SNVs were reported to be 0.57% for rs5515, 0.03% for rs77760094, 0.51% for rs61469141, and 0.12% for rs61742847. Individuals homozygous for the SNVs predicted to be damaging would be more likely to have impaired or absent protein function. All four of these potentially deleterious SNVs had significantly higher frequencies in the AA compared to EA population. The *KLK1* SNV, rs5515, had been linked to endothelial dysfunction and predilection for hypertension in prior studies (Margolius et al., 1976; Slim et al., 2002; Bergaya et al., 2004a).

Structural representations of SNVs

Predicted deleterious missense SNVs may result in adverse residue interactions in key functional areas in primary, secondary, or tertiary protein structure. We thus examined whether the probably damaging missense SNVs predicted by the PolyPhen-2 algorithm would congregate in a particular region of a KLK amino acid sequence. A representative example of this is shown in Figure 3. We show *KLK14* SNVs, given that it had the highest number of probably damaging SNVs. As demonstrated, there was no discernable pattern between the location of probably damaging SNVs and the KLK amino acid sequence, and this was true for all 15 KLKs. In addition, deleterious SNVs did not appear to congregate in any specific areas of the protein domains.

Despite no obvious relationship between the probably damaging missense SNVs and KLK amino acid sequences, we postulated that these particular SNVs may congregate near the active sites of KLKs, impair substrate binding, or interfere with neighboring residues in tertiary protein structure. To evaluate this, we utilized UCSF Chimera software version 1.9 (http://www.cgl.ucsf.edu/chimera), which allowed us to display both native and 'mutated' forms of proteins and to examine steric effects of altered amino acids in the vicinity of active sites. We were particularly interested in examining common missense SNVs (Table 3). PDB files 1SPJ (KLK1), 2ZCH (KLK3), and 3VFE (KLK6) were used to study the structures of interest. All three structures had resolution ≤2.8A and almost complete sequence information (Laxmikanthan et al., 2005; Menez et al., 2008; Liang et al., 2012). There was no available structure for KLK10 or KLK12. Once the structure was displayed in the Chimera viewer, residues of interest were selected, together with all the residues in the surrounding 0.5 A. Interactions were examined. The catalytic triad residues were also highlighted. The residues of interest were then mutated by using the Rotamer function. Finally, the most energetically favorable and the most probable rotameric state of the residue was selected.

We were also interested in potential interactions or clashes between the mutated residues and any other residues in the vicinity that would render the mutation

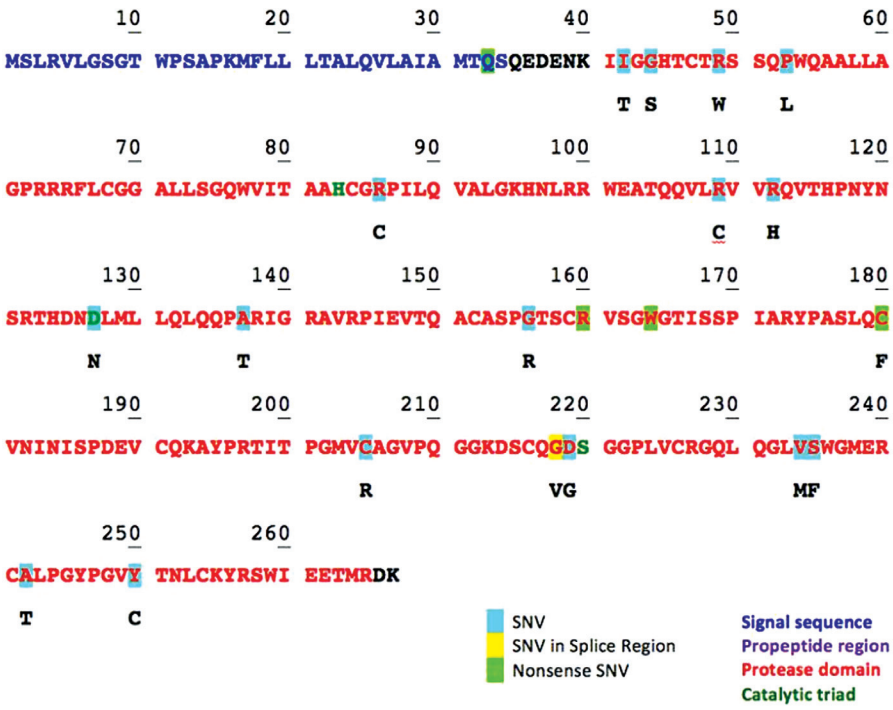


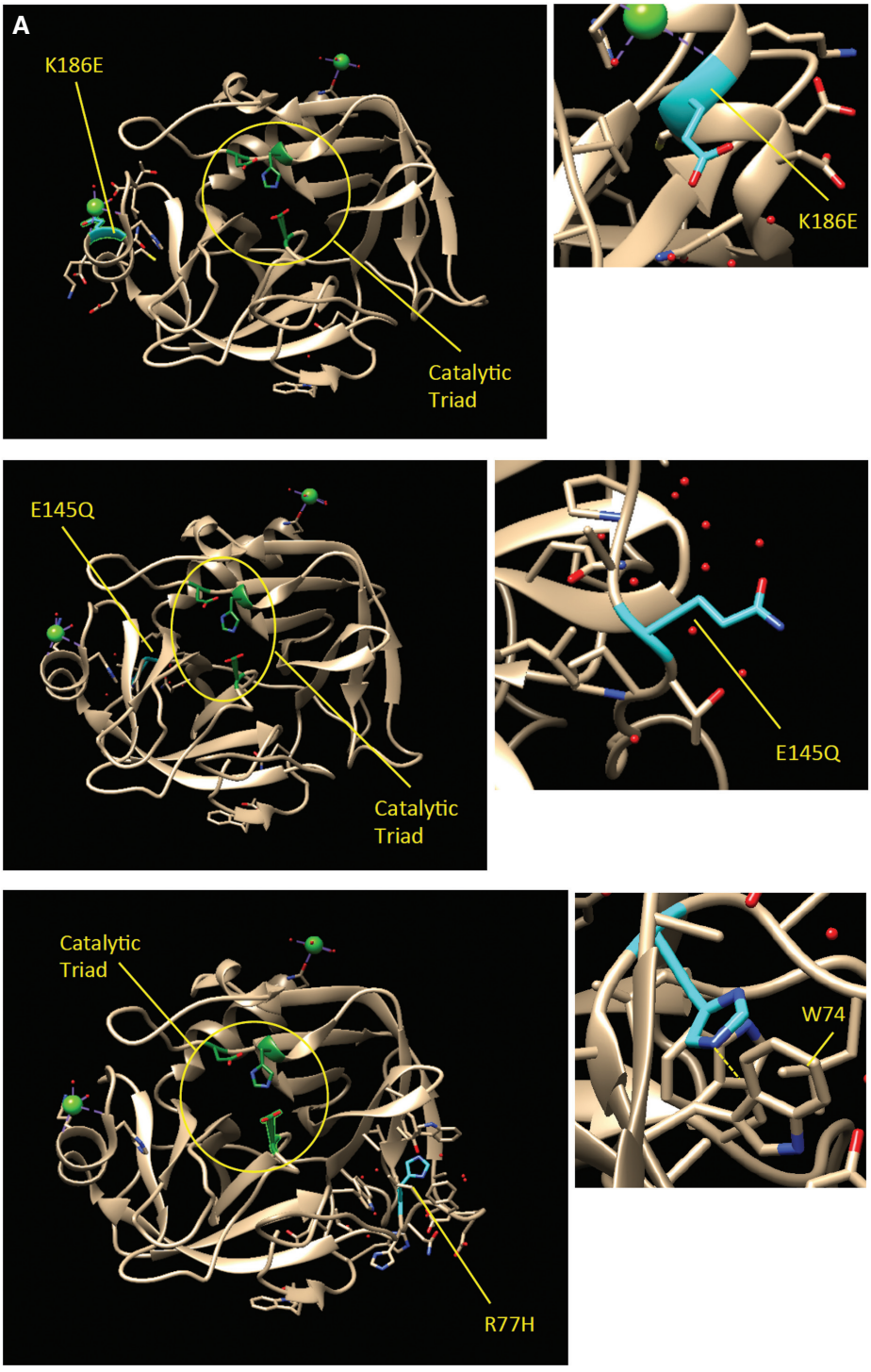
Figure 3 Visual representation of damaging missense SNVs on KLK14 amino acid sequence. All 'probably damaging' SNVs in KLK14 predicted by PolyPhen-2 are plotted. Each letter represents an amino acid residue, and highlighted letters indicate a type of SNV (missense, splice junction, nonsense). Font colors represent the catalytic triad and important domains of the primary structure (signal sequence, propeptide sequence, protease domain). It is interesting to note that one SNV replaces the aspartic acid residue in the catalytic triad.

energetically unfavorable. We thus performed structure minimization in Chimera. The structure with the residue mutated *in silico* was then examined to find any clashes and contacts with all other atoms within 0.6 A. Clashes and contacts represent potential steric hindrance of the selected residue with neighboring atoms and residues in the structure. Any clashing residues were selected to be minimized to the most energetically favorable state, and the potential energy change was calculated in kJ/mol. Proteins fold in a way that minimizes potential energy, and so even local clashes may disrupt the native protein conformation.

Using this method, we first examined common missense KLK1 SNVs (Figure 4A). KLK1 SNVs rs5517 and rs5516 resulted in residue substitutions far from the active site, and close to the surface of the protein, with no interference with other residues, and no requirement for minimization. These SNVs were predicted to be benign and were common in both EA and AA populations. In contrast, KLK1 SNV rs5515 involving arginine mutation to histidine (R77H) resulted in significant clashes between H77 and W74 residues that could not be minimized (Figure 4A). This particular SNV is predicted to be probably damaging and was previously associated with decreased function and disease in experimental models as well as in humans (Katz et al., 1998; Slim et al., 2002).

We next examined KLK3 SNV rs17632542, with I179T substitution (Figure 4B). Small threonine residue did not result in interference with any neighboring residues. This substitution was also predicted to be benign based on PolyPhen-2. In contrast, KLK6 SNVs rs77760094 and rs61469141 were predicted to be possibly damaging, and structural analysis lent support to this prediction. SNV rs77760094 with T234M substitution results in an energetically unfavorable clash between M234 and R129 (Figure 4C). The initial energy of the structure was 148487 kJ/mol, and the potential energy change after minimization was -4133 kJ/mol. This SNV was predicted to be possibly damaging and demonstrated significantly higher frequency in AA population. Similarly, arginine to tryptophan (R78W) substitution (rs61469141) was predicted to be possibly damaging and demonstrated significantly higher frequency in the AA population. As shown in Figure 4C, the substituted tryptophan residue was outside the active site, but demonstrated steric hindrance with a glutamic acid residue (D153), and resulted in exposure of the hydrophobic residue on the protein surface. The initial energy of this structure was unfavorable at 53,834 kJ/mol, and the

potential energy change from minimization was -3997 kJ/mol. This residue may, thus, result in an aberrant protein conformation. Furthermore, W78 is located in a flexible loop, which changes its conformation in active state (Pathak et al., 2013). Structural *in silico* analysis of missense SNVs was supported by PolyPhen-2 prediction and limited experimental data, in the case of KLK1.



(Figure 4 Continued)

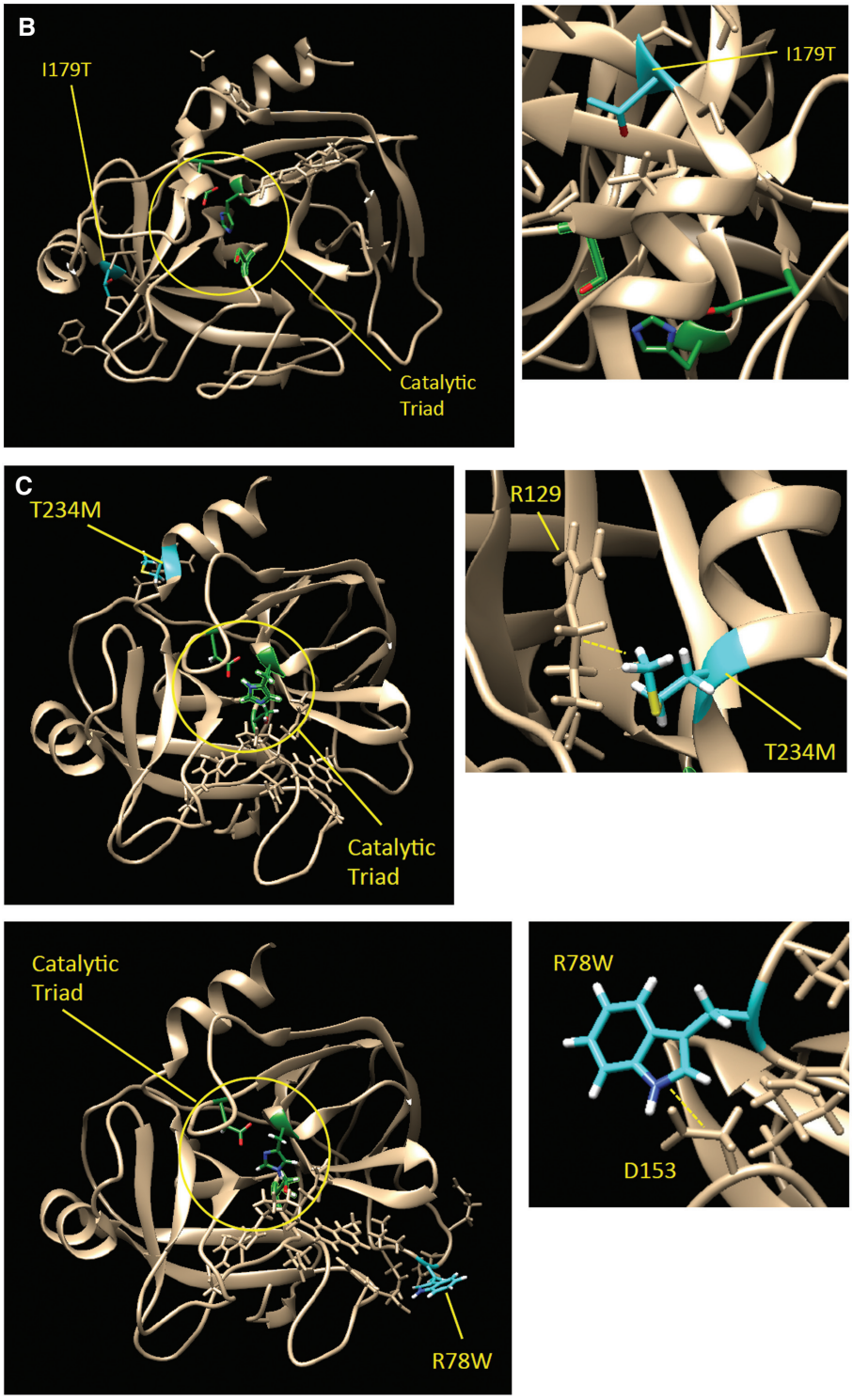


Figure 4 (A) UCSF Chimera representation of KLK1 (PDB file 1SPJ) SNV rs5517 (K186E substitution), rs5516 (E145Q substitution), and rs5515 (R77H substitution). Only rs5515 SNV results in steric hindrance between H77 and W74 residues depicted by a broken line. The other two SNVs have no predicted consequences on tertiary KLK1 structure. (B) KLK3 (PDB file 2ZCH) SNV rs17632542 with I179T substitution is demonstrated. The substitution leads to energetically favorable state, without interference with neighboring residues. (C) KLK6 (PDB file 3VFE) SNV rs77760094 with T234M substitution and SNV rs61469141 with R78W substitution are displayed. M234 residue interferes with R129 in its vicinity as seen in the zoomed view. The relatively large potential energy change from minimization, -4132 kJ/mol, supports the potentially detrimental nature of this substitution and the plausible alteration in the original protein conformation. W78 residue interferes with D153 in its vicinity. The relatively large potential energy change from minimization, -3997 kJ/mol, suggests a potentially detrimental nature of the substitution with plausible effect on protein conformation. All residues within 0.5 A are displayed. The catalytic triad is circled and highlighted in green.

Discussion

KLKs are important serine proteases that have been implicated in various diseases. The main goal of this paper was to perform a comprehensive database mining and *in silico* analysis of the SNVs in the KLK locus. To achieve this goal, we had three objectives: 1) to mine nonredundant databases for the most complete collection of SNVs at the KLK locus; 2) to examine the functional annotations and frequencies of KLK SNVs, particularly focusing on nonsynonymous (missense and nonsense) SNVs; and 3) to perform *in silico* prediction of deleterious SNVs and link the effect of deleterious SNVs to protein structure.

Our first objective was to uncover all SNVs in the KLK locus. Exome Variant Server and Ensembl databases are the largest sequencing databases currently available, and much of the information they contain was derived from next-generation sequencing and was not accessible even 2–3 years ago, when the study by Batra et al. was likely conducted. These databases contain identical SNV accessions and significant amounts of complementary data, notwithstanding the resulting redundancy one must be cautious of. The number of SNVs recovered by mining of these two databases is comparable to Batra et al., who found approximately 500 fewer SNVs in 2012. In contrast, Goard et al. reported approximately one third of our SNVs in the KLK locus in 2007. For the most part, the difference in numbers is accounted for by increasing data completeness. Goard et al. and Batra et al. used dbSNP search engine to identify SNVs (although this information is currently included in Ensembl database). Our study identified a large number of low-frequency SNVs. While rare SNVs are likely to be evolutionarily recent, and to reflect population-specific variation, recent study suggested that these rare SNVs are likely to make an important contribution to disease susceptibility (Tennessen et al., 2012). Batra and colleagues reported only validated SNVs, and this may explain some of the difference in numbers between our studies. Additionally, 55% of their SNVs were unclassified, and these potentially represent inter-genic regions, if we are to extrapolate from Goard et al. study.

Our second objective was to examine the functional annotations and frequencies of the KLK SNVs. The frequencies of almost all functional classes of SNVs have remained unchanged between our study and Goard et al. Intronic and gene locus region SNVs predominated in both studies. Novel insights from the ENCODE consortium have shed light on the fact that even intronic SNVs, previously believed to be inconsequential, may contain important regulatory elements (Consortium et al., 2012; Gerstein et al., 2012; Schaub et al., 2012). Indeed, 14% of the common SNVs we identified in the noncoding regions overlap with TFBSs. There is further experimental evidence that at least some of TFs that bind to these sites regulate KLK expression, as was shown in the case of KLK3. Batra et al. used computational imputation to predict SNVs in TFBSs (Batra et al., 2012). They reported on 15 SNVs that overlapped with TFBSs. However, none of these SNVs overlapped with TFBSs based on the ChIP-seq data from the Genome Browser. This discrepancy is not alarming, given the potential lack of completeness of ENCODE ChIP-seq data or incorrect computational prediction. Future studies could address how individual SNVs affect the binding of TFs and their regulation of KLK genes.

Our study identified a higher fraction of missense SNVs compared to the two older studies, mostly due to a much larger number of sequenced individuals, and to perhaps sequencing artifacts, at least in some cases. Nonetheless, we decided to take advantage of this observation, given that missense SNVs are potentially detrimental to protein function. Along the same lines, we found 54 nonsense SNVs. These types of SNVs represent a potential to detect human knockout phenotypes and, subsequently, better understand the relationship between genotype and phenotype. As expected, the frequencies of these deleterious SNVs were extremely low. It remains unclear why most of the nonsense SNVs occurred in *KLK14*, but it is possible that decreased function of this particular KLK does not have a significant impact on clinical phenotype.

We next examined the population frequencies of the nonsynonymous SNVs. We focused on SNVs likely to be damaging and those with a high enough frequency to be detected in human populations with relative ease. Numerous whole-genome sequencing studies have identified rare SNVs that were subsequently linked to human disease (Helgason et al., 2013; Cruchaga et al., 2014). Although we did not focus on these rare variants, our collected data will remain available for future studies in this area.

In some cases, the SNVs implicated in disease shed light on the propensity of a certain population to develop a disease of interest. A characteristic example of this is a demonstration that two SNVs in the apolipoprotein-L1 (*APOL1*) gene on chromosome 22 that are only present in African-Americans, lead to markedly increased risk of developing kidney disease and progressing to end stage renal disease (Genovese et al., 2010; Parsa et al., 2013). Groundbreaking studies involving *APOL1* provide a compelling reason to closely examine those SNVs that demonstrate dramatically different frequencies in different populations. Notwithstanding this, large-scale studies of nonsynonymous SNVs have uncovered significantly

higher frequencies of these SNVs in European compared to African populations (International HapMap et al., 2007; Lohmueller et al., 2008). Even those SNVs predicted to be probably damaging tended to be more common in the European population (Lohmueller et al., 2008). We found nine common missense SNVs with significantly different frequencies in EA and AA populations. The four SNVs predicted to be damaging were between 3- and 100-fold more common in the AA population. The impact of the frequency differences we report cannot be assessed, although we note that there are cases where a difference in frequency in a SNV accounts for difference in disease susceptibility between African and European individuals (APOL1). Among these nine SNVs, *KLK1* SNV rs5515 leads to a substitution in an arginine residue for a histidine residue and was previously demonstrated to impair kininogen cleavage into bradykinin. This particular SNV was linked to human disease, and the high frequency of this SNV in the AA population was noted before (Waeckel et al., 2013). With AA population frequency of almost 10%, compared to 3.6% in EA population, this SNV could plausibly contribute to an increased risk of hypertension and renal disease in African-Americans.

Another SNV found at a significantly different frequency between the two populations was previously implicated in disease. This was SNV rs17632542 in *KLK3*, which has been linked to non-aggressive prostate cancer (Parikh et al., 2011). This particular SNV was significantly higher in EA population, with a frequency of 7.5%, compared to 1.4% in the AA population ($p<2.2\times10^{-16}$). Prostate cancer development and aggressiveness may vary between AAs and EAs, but studies that examined these relationships reported mixed conclusions (Sucheston et al., 2012; Bensen et al., 2013). We postulate that this SNV, which was previously linked to prostate cancer, may account for higher likelihood of developing non-aggressive prostate cancer in EAs. On the other hand, this difference in frequency may simply reflect the phenomenon of higher frequency of missense SNVs in the EA population. This remains to be studied in the future.

We found two SNVs (rs77760094 and rs61469141) in *KLK6* with significantly different frequencies in the EA and AA populations. Moreover, these two SNVs were predicted to be damaging by PolyPhen-2. *KLK6* is most highly expressed in the brain, with lower expression in the genitourinary tract. Earlier studies have implicated KLK6 in cancer and neurodegenerative diseases (Yousef et al., 2004; Bayani and Diamandis, 2012). Given that the two SNVs we identified display a dramatically higher frequency in AAs, and that they are predicted to be deleterious both by the algorithm and by our structural *in silico* analysis, we propose that these two SNVs are worth studying further to establish whether there is a link between one or both of these SNVs and a predilection for a certain phenotype in the AA population.

Additionally, two of the SNVs in *KLK10*, predicted to be benign, demonstrated markedly increased frequency in the EA population. The frequencies of the two SNVs were 58% and 70%, respectively, in the population at large. We did not focus on these SNVs, given the lack of predicted deleterious potential and their high frequency, supporting a benign nature of these variants.

Lastly, we noted one probably damaging SNV in *KLK12* (rs61742847), with a higher frequency in the AA population. Not much is known about KLK12, but its expression was found to be dysregulated in certain cancers, particularly of the lung. No crystallographic structure is available for this KLK, so we were not able to examine the effect of this SNV on protein structure. Given that this SNV was predicted to be damaging, together with its significant prevalence in the AA population, a search for the potential phenotype in this population may be warranted.

Our final objective was to link the predicted deleterious SNVs to KLK sequence and structure *in silico*. The PolyPhen-2 algorithm was selected for prediction of the functional effect of KLKs. This algorithm has also been formally validated (Wang et al., 2012). There was a substantial number of potentially damaging SNVs across all KLKs, albeit at a low frequency per SNV. Having compiled the list of probably damaging SNVs, we examined the relationship between SNV position in the sequence and secondary protein domains. There was no obvious relationship between the two, as deleterious SNVs scattered throughout the sequence. Examination of the tertiary protein structure allowed a more accurate assessment of SNVs with relation to three-dimensional active sites and other residues in the vicinity. All three KLK SNVs (one KLK1 and two KLK6 SNVs) predicted to be damaging by PolyPhen-2 were also suggested to be damaging based on interference with neighboring residues in tertiary protein structure, exerting a potential effect on the final protein conformation, and thus protein function. Future studies should address the relevance of these SNVs to protein function and human disease.

Our study is not devoid of limitations. First, all functional predictions were made *in silico*, and thus, the assertions we make are hypothesis generating. Furthermore, we narrowed our focus on specific aspects of our findings, such as nonsynonymous and potentially detrimental SNVs. Despite these drawbacks, our study describes exhaustive database mining with the most complete

discovery of SNVs at the KLK locus to date and the most complete description of nonsynonymous KLK SNVs. We included novel important aspects, such as examination of the frequencies of deleterious SNVs, which can inform the prevalence and potential significance of those particular SNVs. We also extended our knowledge gained from data mining to the effects of SNV on protein sequence and structure. Finally, we created a resource for future studies. As such, we separated common and rare SNVs and filtered out nondamaging SNVs in order to facilitate future searches of our data. Future studies should address the functional relevance of identified SNVs.

Materials and methods

Database mining

The SNV databases explored include The Exome Variant Server (http://evs.gs.washington.edu/EVS/), Database of Single Nucleotide Polymorphisms (dbSNP)(http://www.ncbi.nlm.nih.gov/SNP/), The International HapMap project(http://hapmap.ncbi.nlm.nih.gov), The 1000 Genomes Project (http://www.1000genomes.org), Ensembl Genome Browser (http://www.ensembl.org/index.html), and Catalogue of Somatic Mutations in Cancer (COSMIC) database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/). SNVs were classified differently in each of these databases, so we chose to adopt the SNV types used by the Goard et al. study (2007) and classified the SNVs from Ensembl and Exome Variant Server accordingly. Database mining was performed by two individuals (N.S. and E.S.) separately, who pooled the results of their analyses together. Databases were mined for population-based allelic frequencies, base pair and amino acid substitutions, and algorithmic scores predicting SNV functionality (PolyPhen-2, SIFT, and Grantham were the main algorithms used) for kallikreins 1 through 15.

We selected PolyPhen-2 as the optimal prediction algorithm because it takes into account sequence homology, Pfam annotations, and structural information from PDB database (Adzhubei et al., 2010). Furthermore, PolyPhen-2 algorithm has been validated demonstrating specificity for pathogenic variants of 95%, positive predictive value (PPV) of 80%, and a false positive rate of 5% (Wang et al., 2012). Contrary to PolyPhen-2, Grantham scores only reflect chemical dissimilarity between the substituted residues and may not adequately represent the effect of SNV in the context of tertiary protein structure. Similarly, SIFT algorithm only considers sequence homology and physicochemical similarity between the substituted residues (Ng and Henikoff, 2001).

SNVs in the context of KLK amino acid sequence and protein structure

Amino acid sequences were extracted from Uniprot FASTA files for each of the KLKs. Information concerning the secondary domains was also obtained from Uniprot.

UCSF Chimera software (Pettersen et al., 2004) was used to visually display protein structures. KLK PDB file IDs were used to fetch the structures of interest. Only high-quality files with resolution ≤2.0 A and almost complete sequence information were used. Once the structure was displayed in the Chimera viewer, residues of interest were selected, together with all the residues in the surrounding 0.5 A. Interactions were examined. The residues of interest were then mutated by using the Rotamer function. Finally, the most energetically favorable and the most likely position of the residue based on probability was selected. We then performed energy minimization by moving only the residue of interest and the clashing residue(s), while keeping other residues fixed. The initial energy and potential energy change are calculated in kJ/mol. Energy minimization simply moves the structure to local minimum and does not search for global minimum.

Statistics

Comparisons of frequencies between the European and African-American populations were done based on a Pearson χ^2-test on the 2×2 table of the allele counts. Software package R was utilized. p-Values were adjusted for multiple hypothesis testing using Hochberg and Benjamini false detection rate method. p-Values <0.01 were considered significant.

Acknowledgments: The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926), and the Heart GO Sequencing Project (HL-103010).

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Ahn, J., Berndt, S.I., Wacholder, S., Kraft, P., Kibel, A.S., Yeager, M., Albanes, D., Giovannucci, E., Stampfer, M.J., Virtamo, J., et al. (2008). Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. Nat. Genet. *40*, 1032–1034; author reply 1035–1036.

Antoniou, A.C., Beesley, J., McGuffog, L., Sinilnikova, O.M., Healey, S., Neuhausen, S.L., Ding, Y.C., Rebbeck, T.R., Weitzel, J.N., Lynch, H.T., et al. (2010). Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction. Cancer Res. *70*, 9742–9754.

Batra, J., Nagle, C.M., O'Mara, T., Higgins, M., Dong, Y., Tan, O.L., Lose, F., Skeie, L.M., Srinivasan, S., Bolton, K.L., et al. (2011). A Kallikrein 15 (KLK15) single nucleotide polymorphism located

close to a novel exon shows evidence of association with poor ovarian cancer survival. BMC Cancer *11*, 119.

Batra, J., O'Mara, T., Patnala, R., Lose, F., and Clements, J.A. (2012). Genetic polymorphisms in the human tissue kallikrein (KLK) locus and their implication in various malignant and non-malignant diseases. Biol. Chem. *393*, 1365–1390.

Bayani, J. and Diamandis, E.P. (2012). The physiology and pathobiology of human kallikrein-related peptidase 6 (KLK6). Clin. Chem. Lab. Med. *50*, 211–233.

Bensen, J.T., Xu, Z., Smith, G.J., Mohler, J.L., Fontham, E.T., and Taylor, J.A. (2013). Genetic polymorphism and prostate cancer aggressiveness: a case-only study of 1,536 GWAS and candidate SNPs in African-Americans and European-Americans. Prostate *73*, 11–22.

Bergaya, S., Hilgers, R.H., Meneton, P., Dong, Y., Bloch-Faure, M., Inagami, T., Alhenc-Gelas, F., Levy, B.I., and Boulanger, C.M. (2004a). Flow-dependent dilation mediated by endogenous kinins requires angiotensin AT2 receptors. Circ. Res. *94*, 1623–1629.

Bergaya, S., Matrougui, K., Meneton, P., Henrion, D., and Boulanger, C.M. (2004b). Role of tissue kallikrein in response to flow in mouse resistance arteries. J. Hypertens. *22*, 745–750.

Bernard, D., Pourtier-Manzanedo, A., Gil, J., and Beach, D.H. (2003). Myc confers androgen-independent prostate cancer cell growth. J. Clin. Invest. *112*, 1724–1731.

Bharaj, B.B., Luo, L.Y., Jung, K., Stephan, C., and Diamandis, E.P. (2002). Identification of single nucleotide polymorphisms in the human kallikrein 10 (KLK10) gene and their association with prostate, breast, testicular, and ovarian cancers. Prostate *51*, 35–41.

Borgono, C.A. and Diamandis, E.P. (2004). The emerging roles of human tissue kallikreins in cancer. Nat. Rev. Cancer *4*, 876–890.

Collins, F.S., Brooks, L.D., and Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. *8*, 1229–1231.

Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Cruchaga, C., Karch, C.M., Jin, S.C., Benitez, B.A., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., et al. (2014). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature *505*, 550–554.

Deng, Z., Wan, M., Cao, P., Rao, A., Cramer, S.D., and Sui, G. (2009). Yin Yang 1 regulates the transcriptional activity of androgen receptor. Oncogene *28*, 3746–3757.

Eissa, A. and Diamandis, E.P. (2008). Human tissue kallikreins as promiscuous modulators of homeostatic skin barrier functions. Biol. Chem. *389*, 669–680.

Euskirchen, G.M., Rozowsky, J.S., Wei, C.L., Lee, W.H., Zhang, Z.D., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M.B., et al. (2007). Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. Genome Res. *17*, 898–909.

Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science *329*, 841–845.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91–100.

Goard, C.A., Bromberg, I.L., Elliott, M.B., and Diamandis, E.P. (2007). A consolidated catalogue and graphical annotation of dbSNP polymorphisms in the human tissue kallikrein (KLK) locus. Mol. Oncol. *1*, 303–312.

Helgason, H., Sulem, P., Duvvari, M.R., Luo, H., Thorleifsson, G., Stefansson, H., Jonsdottir, I., Masson, G., Gudbjartsson, D.F., Walters, G.B., et al. (2013). A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nat. Genet. *45*, 1371–1374.

Hudson, M.E. and Snyder, M. (2006). High-throughput methods of regulatory element discovery. BioTechniques *41*, 673, 675, 677 passim.

International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Katz, B.A., Liu, B., Barnes, M., and Springman, E.B. (1998). Crystal structure of recombinant human tissue kallikrein at 2.0 A resolution. Protein Sci. *7*, 875–885.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kote-Jarai, Z., Amin Al Olama, A., Leongamornlert, D., Tymrakiewicz, M., Saunders, E., Guy, M., Giles, G.G., Severi, G., Southey, M., Hopper, J.L., et al. (2011). Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. Hum. Genet. *129*, 687–694.

Kruglyak, L. and Nickerson, D.A. (2001). Variation is the spice of life. Nat. Genet. *27*, 234–236.

Ladiges, W., Kemp, C., Packenham, J., and Velazquez, J. (2004). Human gene variation: from SNPs to phenotypes. Mutat. Res. *545*, 131–139.

Laxmikanthan, G., Blaber, S.I., Bernett, M.J., Scarisbrick, I.A., Juliano, M.A., and Blaber, M. (2005). 1.70 A X-ray structure of human apo kallikrein 1: structural changes upon peptide inhibitor/substrate binding. Proteins *58*, 802–814.

Liang, G., Chen, X., Aldous, S., Pu, S.F., Mehdi, S., Powers, E., Xia, T., and Wang, R. (2012). Human kallikrein 6 inhibitors with a para-amidobenzylanmine P1 group identified through virtual screening. Bioorg. Med. Chem. Lett. *22*, 2450–2455.

Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R., et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature *451*, 994–997.

Margolius, H.S., Horwitz, D., Pisano, J.J., and Keiser, H.R. (1976). Relationships among urinary kallikrein, mineralocorticoids and human hypertensive disease. Fed. Proc. *35*, 203–206.

Menez, R., Michel, S., Muller, B.H., Bossus, M., Ducancel, F., Jolivet-Reynaud, C., and Stura, E.A. (2008). Crystal structure of a ternary complex between human prostate-specific antigen, its substrate acyl intermediate and an activating antibody. J. Mol. Biol. *376*, 1021–1033.

Ng, P.C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res. *11*, 863–874.

Paliouras, M. and Diamandis, E.P. (2006). The kallikrein world: an update on the human tissue kallikreins. Biol. Chem. *387*, 643–652.

Parikh, H., Wang, Z., Pettigrew, K.A., Jia, J., Daugherty, S., Yeager, M., Jacobs, K.B., Hutchinson, A., Burdett, L., Cullen, M., et al. (2011). Fine mapping the KLK3 locus on chromosome 19q13.33 associated with prostate cancer susceptibility and PSA levels. Hum. Genet. *129*, 675–685.

Parsa, A., Kao, W.H., Xie, D., Astor, B.C., Li, M., Hsu, C.Y., Feldman, H.I., Parekh, R.S., Kusek, J.W., Greene, T.H., et al. (2013). APOL1 risk variants, race, and progression of chronic kidney disease. N. Engl. J. Med. *369*, 2183–2196.

Pathak, M., Wong, S.S., Dreveny, I., and Emsley, J. (2013). Structure of plasma and tissue kallikreins. Thromb. Haemost. *110*, 423–433.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Riegman, P.H., Vlietstra, R.J., Klaassen, P., van der Korput, J.A., Geurts van Kessel, A., Romijn, J.C., and Trapman, J. (1989). The prostate-specific antigen gene and the human glandular kallikrein-1 gene are tandemly located on chromosome 19. FEBS Lett. *247*, 123–126.

Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., et al. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. *41*, D56–63.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Slim, R., Torremocha, F., Moreau, T., Pizard, A., Hunt, S.C., Vuagnat, A., Williams, G.H., Gauthier, F., Jeunemaitre, X., and Alhenc-Gelas, F. (2002). Loss-of-function polymorphism of the human kallikrein gene with reduced urinary kallikrein activity. J. Am. Soc. Nephrol. *13*, 968–976.

Sucheston, L.E., Bensen, J.T., Xu, Z., Singh, P.K., Preus, L., Mohler, J.L., Su, L.J., Fontham, E.T., Ruiz, B., Smith, G.J., et al. (2012). Genetic ancestry, self-reported race and ethnicity in African Americans and European Americans in the PCaP cohort. PLoS One *7*, e30950.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Tsui, K.H., Lin, Y.F., Chen, Y.H., Chang, P.L., and Juang, H.H. (2011). Mechanisms by which interleukin-6 regulates prostate-specific antigen gene expression in prostate LNCaP carcinoma cells. J. Androl. *32*, 383–393.

Waeckel, L., Potier, L., Richer, C., Roussel, R., Bouby, N., and Alhenc-Gelas, F. (2013). Pathophysiology of genetic deficiency in tissue kallikrein activity in mouse and man. Thromb. Haemost *110*, 476–483.

Wang, J., Schmitt, E.S., Landsverk, M.L., Zhang, V.W., Li, F.Y., Graham, B.H., Craigen, W.J., and Wong, L.J. (2012). An integrated approach for classifying mitochondrial DNA variants: one clinical diagnostic laboratory's experience. Genet. Med. *14*, 620–626.

Yang, S.Z. and Abdulkadir, S.A. (2003). Early growth response gene 1 modulates androgen receptor signaling in prostate carcinoma cells. J. Biol. Chem. *278*, 39906–39911.

Yousef, G.M., Borgono, C.A., White, N.M., Robb, J.D., Michael, I.P., Oikonomopoulou, K., Khan, S., and Diamandis, E.P. (2004). In silico analysis of the human kallikrein gene 6. Tumour Biol. *25*, 282–289.

Supplemental Material: The online version of this article (DOI: 10.1515/hsz-2014-0136) offers supplementary material, available to authorized users.