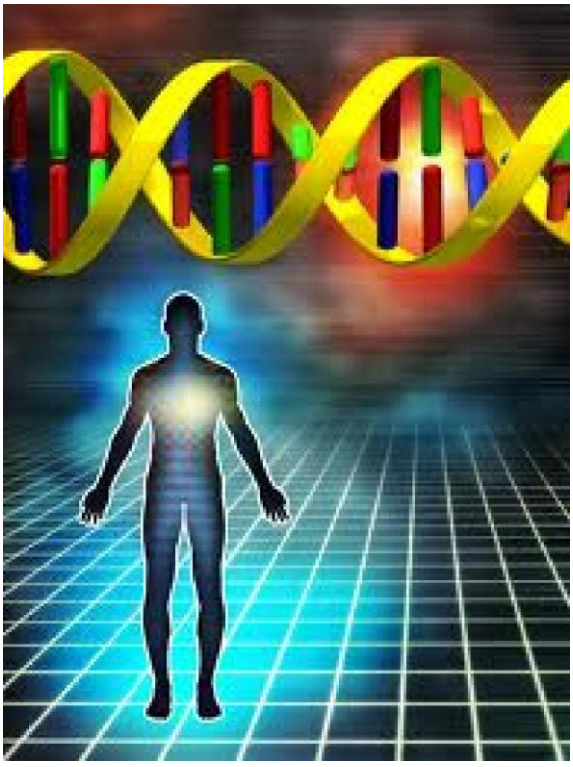
UPPSALA
UNIVERSITET

PXK gene: detailed annotation and new isoforms, associated with autoimmune disease
Systemic Lupus Erythematosus (SLE)



Sepideh Poorazizollahi

Degree project in biology, Master of science (2 years), 2012
Examensarbete i biologi 30 hp till masterexamen, 2012

Biology Education Centre and Department of Medical Biochemistry and Microbiology (IMBIM), Uppsala University
Supervisor: Sergey Kozyrev External
opponent: Maria Lembring

Contents

Abreviations:

SLE: Systemic lupus erythematosus

PXK: Phox homology domain containing serine/ threonine kinase

GWA: Genome Wide Association

EGFRs: epidermal growth factor receptors

MHC: major histocompatibility

LD plot: linkage disequilibrium plot

PBMCs: peripheral blood mononuclear cells

SNPs: single nucleotide polymorphisms

PCR: polymerase-chain reaction

RACE PCR: Rapid Amplification of cDNA ends

cDNA: coding DNA

EXO/SAP: Exonuclease enzyme (from *E.Coli*)/Shrimp Alkaline Phosphatase

Ct: Threshold cycle

Abstract:

Systemic lupus erythematosus (SLE) is classified as a prototypic systemic autoimmune disease characterized by a very diverse range of clinical manifestations. Some patients show skin rashes but more than one-half of the SLE patients have more severe complications of the disease including glomerulonephritis, arthritis, central nervous system vasculitis, interstitial lung disease and stroke. SLE patients are suffering from disregulation of adaptive and innate immune systems. The principal immunological event in SLE is extended autoantibody production. The surplus of various immune complexes depositing in different organs results in inflammation and tissue damage.

The disease can occur at nearly any age however, women in their reproductive ages are affected by the disease with the ratio of 9:1 to males (over 85% of patients are female). SLE is a complex disease which means that genetic factors and environmental factors such as sun exposure, certain drugs and viral infections contribute to the disease development.

Association of the PXK gene to SLE has been reported recently by GWA (Genome Wide Association) studies but the gene has not been well characterized yet. Our knowledge of functional annotation and causative variants of the gene is very insufficient.

In the current project; selective re-sequencing of some regions of the gene in a number of SLE patients and healthy individuals has been performed. A statistical genetic study on SNPs (single nucleotide polymorphisms) accomplished through genotyping study and statistical analysis. Software like SNPExpress, Sequencher, GraphPad and genome browsers like Ensembl, Ncbi and UCSC were mainly used in this experiment.

As the results of current project, two new isoforms of the gene has been introduced including Δ Ala-exon16 and Δ Ala-exon17. In addition, a correlation between genotype of SNP rs6772652 (A/G) and expression has been demonstrated in the current project.

Introduction:

Systemic lupus erythematosus (SLE):

Systemic lupus erythematosus (SLE) is a multisystem autoimmune disease [1]. Skin, blood cells, central nervous system and kidneys are influenced in SLE [1]. The disease course is characterized by periods of relapse and remission [2]. Clinical manifestations are varying remarkably between patients and may range from rashes and anemia to arthritis, nephritis and psychosis [1, 2]. The clinical manifestations of SLE are reported to be more severe and aggressive in pediatric-onset than that of adults [3].

The disease characterized by presence of antinuclear autoantibodies, hyper-activation of T and B cells, activation of complement and interferon, decreased ability of elimination of apoptotic cells and organ destruction [4, 5]. Both innate and adaptive immune systems are deregulated in SLE which leads to inefficient clearance of apoptotic debris; subsequently production of auto-antibodies which accumulate in tissues results in inflammation and organ damage [1, 2].

The disease is affecting women mostly in their reproductive years with a female: male ratio of 9:1 [2]. The prevalence of SLE is estimated at 10-40 per 100 000 in populations of northern European ancestry; in African and Hispanic American the prevalence is two- to fivefold higher [1]. The overall prevalence of SLE is 20 to 150 cases per 100 000 [6].

SLE is a complex disease which means genetic components and environmental factors are both contributing to its pathogenesis [6]. Genomic variations are interacting with each other and with environmental factors [5]. The concordance rate of 25% in monozygotic twins comparing to that of 2% for dizygotic pairs, is an obvious evidence of genetic contributions to SLE as well as the role of environmental factors [1]. Environmental factors such as sun exposure, smoking, viruses and certain drugs may trigger the disease [2].

Genetic studies on SLE:

Different approaches have been followed to discover and study genetic risk factors of SLE comprising candidate gene approach, pedigree-based linkage analysis and GWA (Genome-wide association) studies [4]. Regarding the basis of SLE which is lack of immune tolerance to self-components, many genes coding for proteins with regulatory functions in the immune system have been suggested as candidate genes for the disease [2]. Some of susceptible genes reported by candidate gene studies are *HLA-DR* and several complement components, *STAT4, PTPN22, TREX1* and *IRF5* [4].

Several genome scan studies also carried out by major scientific groups in USA and in Europe (Uppsala, Sweden) introducing many susceptibility loci for SLE [2]. The MHC (major histocompatibility) region has shown the strongest association in recent GWA studies of SLE [1]. Several other genes have been reported through GWA studies associated to SLE, comprising *PXK, BANK1, ATG5, BLK, ICA1* and *ICA1* [1]. Table 1 summarizes association evidence for some newly reported genes for SLE(data from Sestak *et al* 2011, reference 8).

Table 1: *List of susceptibility genes for SLE. (Data from Sestak et al 2011, reference 8).*

Gene	Chromosome	Gene	Chromosome
BANK1	4q24	*LYN*	8q12.1
BLK	8p23.1	*NMNAT2*	1a25
C1q	6p21.32	*KIAA1542/RHRF1*	11p15.5
C2	6p21.32	*LRRC18-WDFY4*	10q11.22
C4A/B	6p21.32	*LYN*	8q12.1
CRP	1q23.2	*NMNAT2*	1q25
ETS1	11q24.3	*PRDM1,ATG5*	6q21
FcGR2A-FcGR3A	1q23.2	*PTPN22*	1p13
FcGR3B	1q23.2	*PTTG1*	5q33.3
HIC2-UBE2L3	22q11.21	*PXK*	3p14.3
HLA-DR2 and DR3	6p21.32	*RASSGRP3*	2p22.3
IKZF1	7p12.2	*SLC15A4*	12q24.32
IL-10	1q32.1	*STAT1, STAT4*	2q32.3
IRAK1,MECP2	Xq28	*TNFA/P3*	6q22.3
IRF5	7q32	*TNFSF4*	1q25.1
ITGAM-ITGAX	16p11.2	*TNIP1*	5q33.1
JAZF1	7p15.2	*TREX1*	3p21.31
KIAA1542/PHRF1	11p15.5	*UHRF1BP1*	6p21.31
LRRC18-WDFY4	10q11.22	*XKR6*	8p23.1

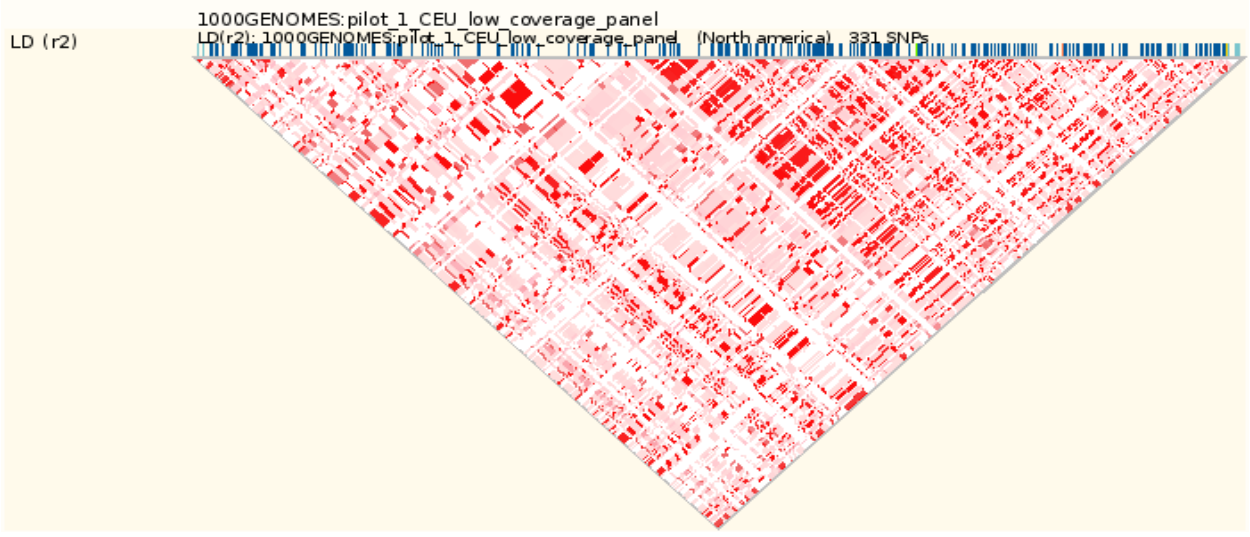
PXK gene:

Phox homology domain containing serine/ threonine kinase (PXK), positioned on chromosome 3p14.3, is highly expressed in brain, heart, skeletal muscle and peripheral blood lymphocytes [3]. Recently, PXK was reported to be involved in the ligand-induced internalization and degradation of epidermal growth factor receptors (EGFRs) [6]. Based on this data PXK could be a candidate gene as a cause of SLE [6]. Additionally, association of the gene has been supported by GWA studies with a very strong evidence of association ($P = 7.1 \times 10^{-9}$) for SNP rs6445975 in population of women with European ancectry [1, 5, 7, 8].

There are no other associated PXK variations reported for SLE so far. However, according to ongoing studies on PXK (unpublished); there are association evidences of some other SNPs of PXK gene. Bioinformatics studies and studying haplotype blocks of PXK suggested 6 potentially associated variations and regions of the gene to SLE.

The non-random association between occurrences of alleles at two loci is called linkage disequilibrium (LD) [9]. The importance of LD is to ease the process of identifying disease-susceptibility loci in GWA studies [10]. To evaluate the LD in practical data sets, graphical approaches have been developed. [10]. Two different pairwise LD statistics including r^2 and D'(reviewed in, *e.g.*, [11]) are being demonstrated by heat map representations [10]. In r^2or D' displays, the strength and distribution of the pairwise LD are indicated by color shading creating segments in the plot called LD block [10]. r^2 and D' LD plots for PXK extracted from ensembl are shown in Figure 1.

a)



b)

Figure 1: *Examples of LD plot of PXK gene. a)* r^2 *LD plot. b)* D' *LD plot. Data from Ensembl.*

Suggested SNPs from recent unpublished studies are shown in Table 2.

Table 2: *list of six SNPs in PXK, potentially associated to SLE.*

SNP	Location	Alleles
Rs9862378	Exon-1	G/T
Rs7610449	Exon-1	G/A
Rs11713310	Exon-1	G/A
Rs11710823	Exon-16	A/G
Rs4681851	Exon-16	C/G
Rs6772652	Exon-17	A/G

PXK has been reported in different GWA studies as a candidate gene for SLE, but there are not much information on gene annotation and function. In this experiment, focus was on PXK detailed annotation. Additionally, selectively re-sequencing and genotyping of some important regions of the gene and SNPs in the population of healthy individuals and patients was performed. Human peripheral blood mononuclear cells (PBMCs) DNA, also cDNA (coding DNA) from spleen, PBMC and thymus are subjected to transcript analysis and genotyping in current experiment.

Materials and methods:

Materials.

Cell lines Jurkat and Daudi DNA were used for characterization of 3' end of the gene by RACE PCR (Rapid Amplification of cDNA ends). Cell lines also were subjected to PCR (polymerase-chain reaction) for characterization of some gene regions, as well as the template in optimization PCRs to find optimum conditions to perform PCRs.

DNA extracted from PBMC of 192 healthy individuals selected from blood donors of Uppsala hospital and used for SNPs genotyping and sequencing. DNA was available from previous experiments and used for genotyping SNPs in current experiment.

cDNA for transcript analysis was prepared from total RNA purified form healthy donors as described in Kozyrev et al. 2008 [12]. RACE PCR (Rapid Amplification of cDNA ends) using one primer complementary to a region in the middle of the gene and the other primer complementary to 3' end of the gene was performed on PBMC cDNA and products were sequenced to characterize the 3' end of the gene.

Genomic DNA from 16 lupus patients was used for sequencing and SNP genotyping.

cDNA from PBMC, spleen and thymus of group of 16 lupus patients used for sequencing of regions of the gene and studying splice variants.

cDN A from human spleen and thymus was purchased from Clontech Laboratories, Inc.

Bioinformatic analysis of the PXK gene.

Bioinformatics annotation was performed by using on-line databases including: ENSEMBL address at http://www.ensembl.org/index.html, NCBI available at http://www.ncbi.nlm.nih.gov and USCS available at http://www.genome.ucsc.edu/cgi-bin/hgGateway/.

The gene expression analysis was performed with the program SNPexpress . This program contains gene expression values of 47294 transcripts from lymphoblastoid cell lines analyzed by microarray in 270 Individuals for 3.96 million SNPs from four populations comprising CEU (Utah residents with ancestry from northern and western Europe): 90, YRI (Yoruba in Ibadan, Nigeria): 90, CHB (unrelated Han Chinese

in Beijing): 45 and JPT (unrelated Japanese in Tokyo): 45 [15]. Each gene was represented on the Illumina Human WG-6 Expression BeadChip v1 array by one or more probes and data was normalized for each population separately (in order to preserve population-specific differences) [15]. In the second round, normalization was performed after pooling all four populations together, which made it possible to have direct comparisons across populations [15]. This software provides calculations of correlation between HapMap genotypes and transcript expression levels [15]. SNPexpress performs an expression quantitative trait locus (eQTL) analysis by visualizing the correlation between genotype of SNPs located in specific region of genome and a gene of interest [15]

Reading sequencing files of samples within this experiment was done by Sequencher4.8 software available at http://www.sequencher.com/. This software shows chromatograms of sequenced samples and aligns several samples at time both in reverse and forward directions.

LD plot was downloaded from Ensembl.

Statistical analysis.

The statistical analysis was performed with GraphPad, online available at http://www.graphpad.com/welcome.htm. P-values and average expressions of PXK for all individuals were calculated by GraphPad. The relative expression levels of PXK for individuals were calculated before, using TBP (TATA binding protein) as reference gene. TBP gene expression is constant in all samples and not affected by the experimental treatment during the study; these traits are characteristics of a reference gene.

Polymerase Chain reaction.

To amplify regions of gene containing our SNPs of interest from genomic DNA and cDNA of group of healthy individuals and patients, many Polymerase Chain Reactions were performed with different DNA polymerases, primers and annealing temperatures. Hifi DNA polymerase, Platinum® Taq DNA Polymerase and PCR buffer purchased from Invitrogen. DNA polymerase enzyme AmpliTaqGold, from Fermentas also used in this experiment. In some amplification Invitrogen PCR Enhancer, PCRx Enhancer was also used. PCR reactions were carried out following as in Table 3.

Table 3. *Basic protocol used for PCRs in this experiment.*

Platinum® Taq DNA Polymerase **Hifi DNA polymerase** **AmpliTaqGold DNA polymerase**	0.15 µl / PCR reaction
10X PCR Buffer, Minus Mg	2.5 µl / PCR reaction
Mg Cl_2 (50 mM)	0.8 µl / PCR reaction
dNTPs (5 mM)	1 µl / PCR reaction
Forward primer: *5'-GATTGGCCTGAGATAGTAAAGTCA-3'* Forw-PXK-int15 *5'-AATGAAGTGTGACTCCAGAGCCTACT-3'* Forw-Ex1B-PXK *5'-TAGAGCATGCACCATTTTGAACGTG-3'* Forw-ex17a *5'-GCTCTTGAAAATAGTGAAGAGCAT-3'* Forw-ex16-PXK *5'-GCTAAACTCCTGGACTCAAGCCAT-3'* Forw-prom2 *5'-GCTGAGAAGTTGATCCCAAGGT-3'* Forw-prom1 *5'-CTGCGAGGAGCAGGGAAGCGCA-3'* Forw-PXK-int1 *5'-GCTCTTGAAAATAGTGAAGAGCAT-3'* Forw-ex16-PXK *5'-TCACCA GCATCGAAGACTGACAAGAGC-3'* Forw-ex15	1 µl / PCR reaction
Reverse primer: *5'-TCTGTAGTCATTACTACATTGCCCAG-3'* Rev-PXK-int16close *5'-ACTTGCTATCATTTGTGCCTAAAGG-3'* Rev-rs7610449 in int2 *5'-CAAAGGAGAAGGTGGTTCTCCCGAGAG-3'* Rev-int17a *5'-GGTTTCACCCTAGTTACCAAGCAGTT-3'* Rev-int16-PXK *5'-TCTTCTAGTCCATATGGTGGGATCA-3'* Rev-prom2 *5'-TTCTGCGCTGGGTCGGCGCTA-3'* Rev-prom1 *5'-ACAAGTAGGCTCTGGAGTCACACT-3'* Rev-PXK-ex1B *5'-GGTTTCACCCTAGTTACCAAGCAGTT-3'* Rev-int16-PXK	1 µl / PCR reaction
PCR conditions: (different annealing temperatures)	95° for 5' 95° for 15" 58°-66°for15" } X 45 cycle 72° for 1' 72° for 5'
PCRx Enhancer (only used in some of the amplifications)	2.5µl and/or 1.25µl

In order to characterization of 3' and 5' ends of gene and transcripts, RACE (rapid amplification of cDNA ends) PCR was performed fallowing Clontech, Inc Marathon-Ready™ cDNAprotocol. Electrophoresis to validate PCR results was performed on 1% TAE agarose gels. DNA ladders used in electrophoresis comprising 100 bp DNA ladder, 3231L, BIOLABS, 1kb DNA ladder, N3232L, NEW ENGLAND BIOLABS Inc and 100 bp DNA ladder Generuler from Fermentas.

DNA purification:

When multiple PCR bands were present on a gel, every PCR-band was excised from the gel and DNA was purified using Qiagen Gel Extraction kit. The DNA concentration was measured by Nanodrop afterwards.

EXO/SAP treatment.

After PCR amplification and validation of sequence of interest on 1% agarose gel, PCR products were treated with Exo/SAP treatment. Treatment reagents including E.coli exonuclease enzyme and SAP (Shrimp Alkaline Phosphatase) used to clean up PCR product from leftover primers and dNTPs purchased from Fermentas (product numbers: EN0581 and EF0511). Exo/SAP treatment protocol illustrated in Table 4.

Table 4. *EXO/SAP treatment recipe. Special amounts of ingredients including SAP (Shrimp Alkaline Phosphatase), EXO1 and EXO buffer were mixed and added to each PCR product. Incubation time was 1 hour at 37° and 15 minutes at 85°.*

0.15 µl / PCR product	**SAP(Shrimp Alkaline Phosphatase)**
0.1 µl / PCR product	**EXO1**
2 µl / PCR product	**EXO buffer**

Genotyping.

The choice of genotyping technique was determined by the expected number of samples needed to be genotyped. The sequencing method of Sanger sequencing was applied in this experiment; the service is provided by Uppsala Genome Center Sanger sequencing service at Rudbeck laboratory, Uppsala. In that method reactions were carried out using AB BigDye Terminator v3.1 and separated on the ABI3730XL DNA Analyzer by capillary electrophoresis. Afterwards, Sequencher 4.8 software was used to read chromatograms. This method is proper for re-sequencing of specific regions of a gene and samples are prepared in 96-well plates which is totally convenient to be applied in this experiment. Each well contained 2 µl of EXO/SAP treated PCR products, 1 µl primer (4PM/ µl) and 15 µl of H_2O.

Results:

In order to understand the molecular details underlying the genetic association of the PXK gene with autoimmune disease SLE, the gene and the associated variants were firstly annotated using various public databases (Ensembl, Ncbi and UCSC). The human PXK gene coding for PX domain, containing serine/threonine kinase, is located on chromosome 3p14.3 covering the region of over 92 kb. 14 different transcripts were reported for the gene in Ensembl (Figure 2, Table 5), with varying number of exons. The full-length isoform contains 18 exons, while some truncated transcripts included 4-9.

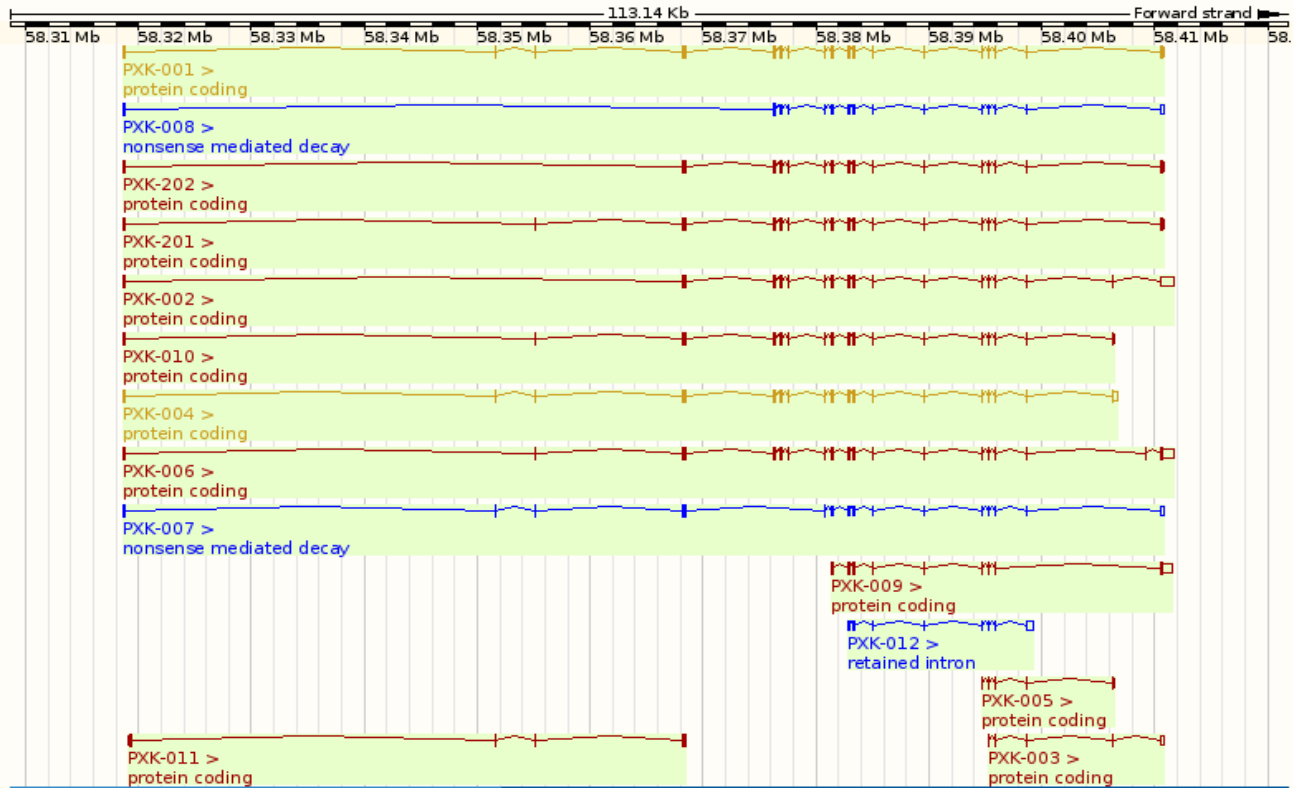


Figure 2: *Structure of PXK transcripts. There are 14 transcripts reported for this gene in Ensembl.*

Table 5: *PXK exons in different transcripts of the gene. Constitutive exons are the ones which are existing in all of transcripts and Alternative exons are missing in some of the transcripts. The ones indicated with a star(*) are located in 3'UTR (untranslated) region.*

	PXK-001	PXK-008	PXK-202	PXK-201	PXK-002	PXK-010	PXK-004	PXK-006	PXK-007	Exon characteristic
Exon 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	constitutive
Exon 2	yes	No	No	No	No	No	yes	Yes	No	Alternative
Exon 3	yes	No	No	yes	No	yes	Yes	yes	yes	Alternative
Exon 4	yes	No	Yes	Yes	Yes	Yes	Yes	yes	Yes	Alternative
Exon 5	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	No	Alternative
Exon 6	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	No	Alternative
Exon 7	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	No	Alternative
Exon 8	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	Constitutive
Exon 9	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 10	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 11	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 12	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 13	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 14	yes	Yes	Yes	Yes	Yes	Yes	yes	yes	Yes	constitutive
Exon 15	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 16	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	Yes	constitutive
Exon 17	yes	Yes	Yes	Yes	Yes	Yes	Yes	yes	yes	constitutive
Exon 17A	No	No	No	No	yes	Yes	Yes*	No	No	Alternative
Exon 17B	No	No	No	No	No	No	No	yes	No	Alternative
Exon 18	yes	Yes*	yes	yes	yes*	No	No	Yes*	Yes*	Alternative

According to Ensembl, 868 known variations have been reported for PXK with different characteristics based on their position. Table 6 briefly gives information on how the variations are distributed.

Table 6: *list of PXK gene variations reported in Ensembl and their characterization.*

Number of variants	Type	Description
0	Essential splice site	In the first 2 or the last 2 basepairs of an intron
0	Stop gained	In coding sequence, resulting in the gain of a stop codon
0	Stop lost	In coding sequence, resulting in the loss of a stop codon
0	Complex in/del	Insertion or deletion that spans an exon/intron or coding sequence/UTR border
0	Frameshift coding	In coding sequence, resulting in a frameshift
44	Non-synonymous coding	In coding sequence and results in an amino acid change in the encoded peptide sequence
49	Splice site	1-3 bps into an exon or 3-8 bps into an intron
0	Partial codon	Located within the final, incomplete codon of a transcript whose end coordinate is unknown
41	Synonymous coding	In coding sequence, not resulting in an amino acid change (silent mutation)
0	Regulatory region	In regulatory region annotated by Ensembl
0	Within mature miRNA	Located within a microRNA
651	Intronic	In intron
153	NMD transcript	Located within a transcript predicted to undergo nonsense-mediated decay
11	5 prime UTR	In 5 prime untranslated region
62	3 prime UTR	In 3 prime untranslated region
22	Within non-coding gene	Located within a gene that does not code for a protein
22	Upstream	Within 5 kb upstream of the 5 prime end of a transcript
15	Downstream	Within 5 kb downstream of the 3 prime end of a transcript
0	HGMD mutation	Mutation from the HGMD database - consequence unknown
0	Intergenic	More than 5 kb either upstream or downstream of a transcript

Unpublished data on PXK based on LD plot and other experiments suggesting some potentially associated SNPs with SLE. Six SNPs including rs9862378, rs7610449, rs11713310, rs11710823 and rs6772652 summarized in Table 2 in introduction section.

In total 25 SNPs were genotyped in population of 96 healthy blood donors of Uppsala hospital. Small DNA fragment containing the SNP was PCR-amplified and analyzed by sequencing.

PCRs were performed following protocol mentioned in Table 3 and gel electrophoresis pictures of some samples are provided in Figure 3.

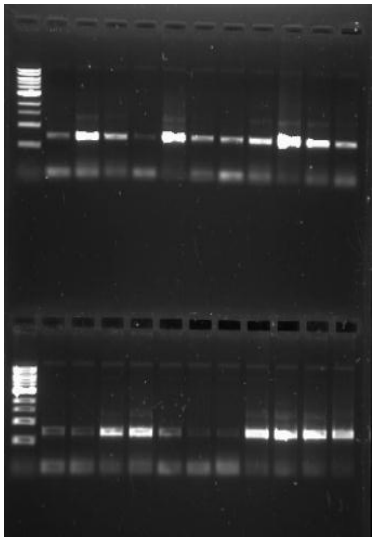
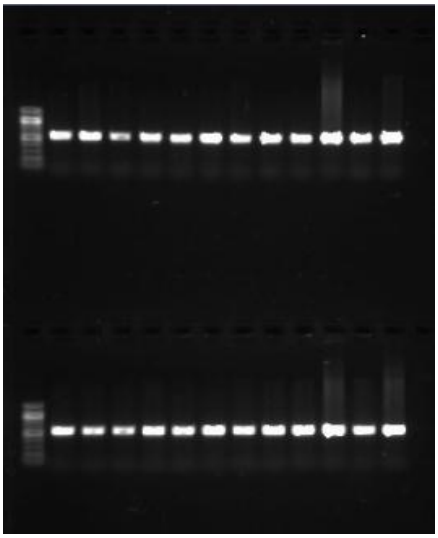
a)  b) 

Figure 3 a, b: *amplification of some regions in exon 1 and exon 16 of PXK gene (respectively pictures a and b) in some individuals from PBMC DNA of 96 healthy blood donors of Uppsala hospital. Ladder used in electrophoresis was 100bp Generuler on 1% agarose TAE gel. a) Amplification is not very good for some samples. b) Perfect amplification of the region of interest in all of the samples. (Ladder used in both gels 100bp Generuler)*

EXO/SAP treatment of amplified regions was performed according to the protocol illustrated in Table 4. PCR products were sent for sequencing to Uppsala Genome Center. Sequencing chromatograms on Sequencher software were analyzed (Figure 4).

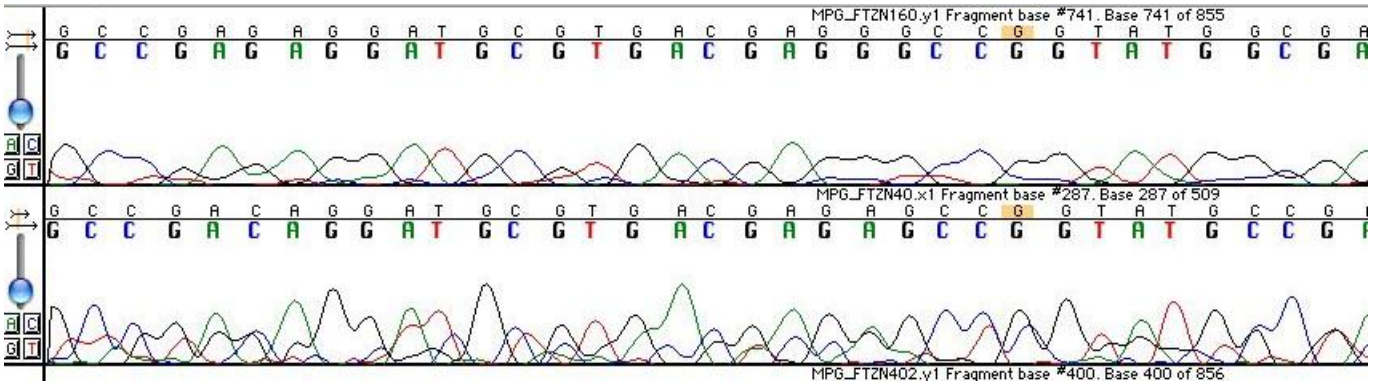


Figure 4: *Sequencing chromatograms. Data from Sequencher.*

In case of four SNPs of interest comprising rs4681851, rs7610449, rs6772652 and rs11710823, amplification was done successfully for most of the 96 individuals. Sequencing results showed some

heterozygotes in the population which were then subjected to expression study. Expression levels of all individuals for PXK gene was available (Table 7).

Table 8: *Relative expression of PXK/TBP values for some of the individuals used in this experiment.*

	PXK			TBP:52-TBP/TBP-3			
sample ID	*Ct1*	*Ct2*	*aver Ct*	*Ct1*	*Ct2*	*Aver Ct*	PXK/TBP
46	24.1	24.02	24.06	24.5	24	24.25	0.989796
49	24.4	24.4	24.4	24.2	24.1	24.15	0.997934
51	23.4	23.4	23.4	23	22.5	22.75	0.98913
56	24.3	24.6	24.45	22.9	23	22.95	1.002183
58	25.4	25.3	25.35	23.8	23.4	23.6	0.991597
60	24.5	24.08	24.29	22.6	21.8	22.2	0.982301
65	24.4	24.2	24.3	23.5	22.4	22.95	0.976596
68		26.6	26.6	21.8	20.7	21.25	0.974771
112	24.9	24.8	24.85	23.3	22.1	22.7	0.974249
113	25.2	25.9	25.55	22.2	21.6	21.9	0.986486
114	24.1	24	24.05	23	22.5	22.75	0.98913
115	24.6	24.7	24.65	23	22.6	22.8	0.991304
117	25.5	25.6	25.55	23.7	22.2	22.95	0.968354
119	24.6	24.7	24.65	23.1	21.6	22.35	0.967532
121	24.8	24.7	24.75	23	21.3	22.15	0.963043
123	24.7	24.3	24.5	22.6	22	22.3	0.986726
204	24.5	24.3	24.4	22.9	21.1	22	0.960699
206	24	23.8	23.9	22.6	21.7	22.15	0.980088
212	23.4	23.1	23.25	21.9	20.7	21.3	0.972603
215	24.6	24.3	24.45	22.4	20.7	21.55	0.962054
216	24.7	24.7	24.7	22.6	21.1	21.85	0.966814
217	24.3	24.3	24.3	22	20.8	21.4	0.972727
328	24.2	24.04	24.12	22.4	21.3	21.85	0.975446
330	25	24.9	24.95	23.6	22.7	23.15	0.980932
331	24.8	24.5	24.65	24.8	23.4	24.1	0.971774
332	25.5	25.2	25.35	22.6	21.4	22	0.973451
337	24.9	24.7	24.8	22.2	20.6	21.4	0.963964
352	24.7	24.7	24.7	22.4	20.6	21.5	0.959821
358	24.7	24.5	24.6	22.6	21.2	21.9	0.969027
460	24.4	24.1	24.25	22.4	21.8	22.1	0.986607

In statistical genetic study, average expression for each genotype was calculated. P values of each genotype versus two others were calculated using GraphPad. The P value is a measurement of statistical significance. The P value should be higher than the significance level indicated by α which is normally 0.05 or 0.01 (in this experiment α was 0.01).

Real-time PCR was performed in previous studies to quantify the expression levels of the PXK and TBP (TATA binding protein) genes. The Ct is threshold cycle which is an indicator of increase of the product in real-time polymerase chain reaction. The relative expression levels of PXK for individuals were calculated using TBP as a reference gene. Results provided in Figures 5, 6, 7 and 8, and Table 8.

Table 8: *Genotyping results of* rs6772652 (A/G), rs4681851 (G/C) *and* rs11710823 (A/G*) and statistic genetic analysis of their expression.*

SNP	Genotypes	Average expressions	Individuals	P values
rs6772652 A/G	GG	0.162453	53	GG vs AG: 0.01 GG vs AA: 0.73 AG vs AA: 0.59
	AG	0.253143	35	
	AA	0.183333	3	
rs4681851 G/C	CC	0.201268	71	CC vs GC: 0.88
	GC	0.196923	13	
	GG	0.13	1	
rs11710823 A/G	GG	0.201333	30	GG vs AG: 0.2848 GG vs AA: 0.1162 AG vs AA: 0.1689
	AG	0.2605	20	
	AA	0.106	5	

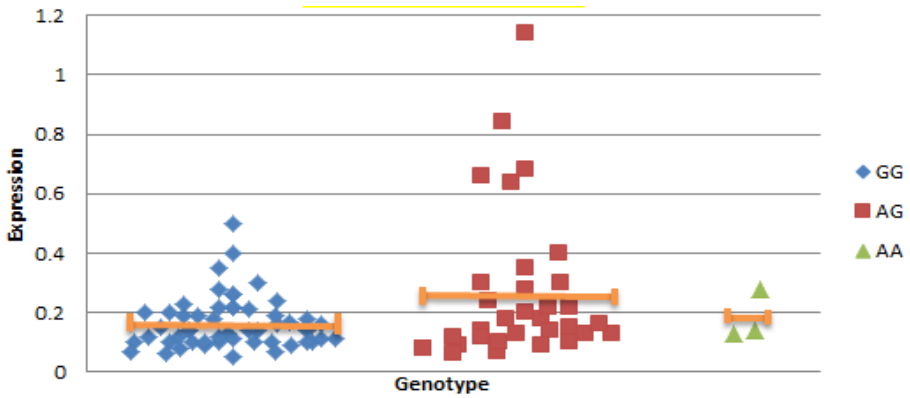


Figure 5: *SNP rs6772652 genotyping and expression study results. Distribution of genotypes in population regarding their expression level.*

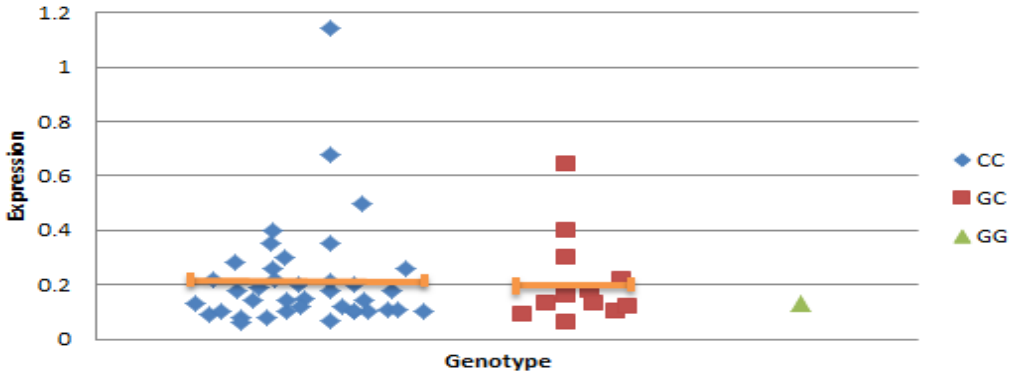


Figure 6: *SNP rs4681851 genotyping and expression study results. Distribution of genotypes in population regarding their expression level.*

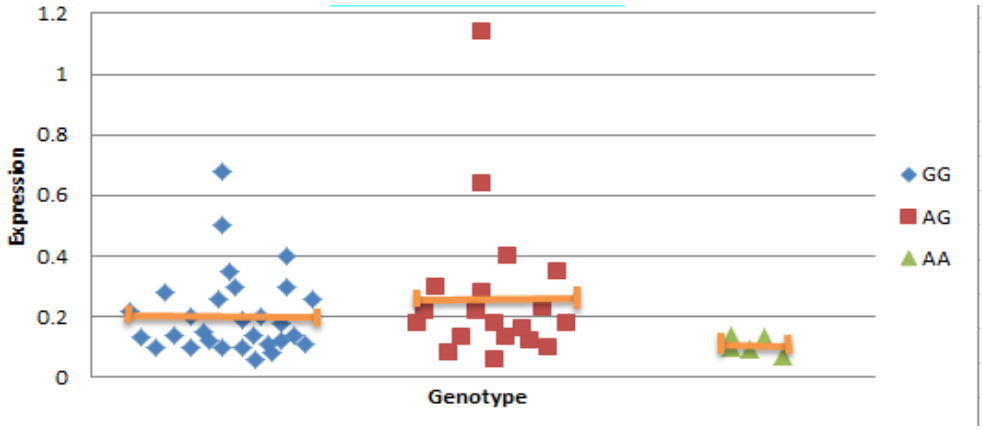


Figure 7: *SNP rs11710823 genotyping and expression study results. Distribution of genotypes in population regarding their expression level.*

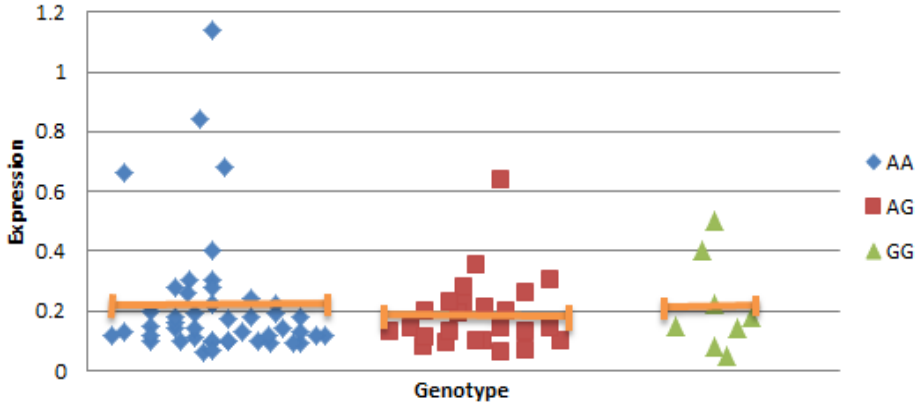


Figure 8: *SNP rs7610449 genotyping and expression study results. Distribution of genotypes in population regarding their expression level.*

SNPexpress was deeply searched for expression patterns of the SNPs in its database of PBMC cells. Figure 9 demonstrates data for SNPs rs11710823 extracted from SNPexpress.

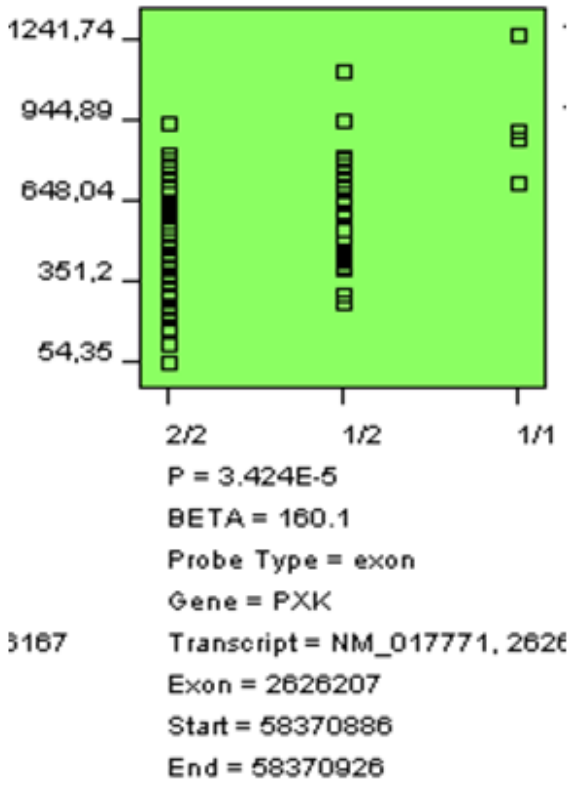
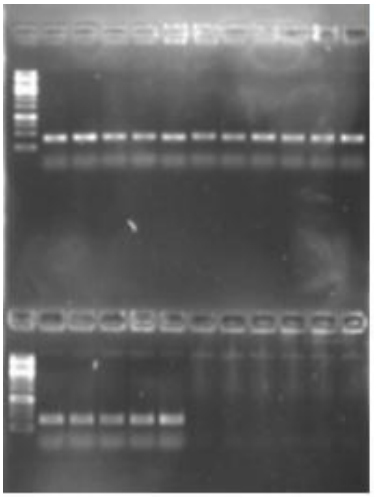


Figure 9: *expression data of SNP rs11710823 in PBMC cells. Expression levels of samples subjected to the analysis for homozygous and heterozygous alleles. 1/1 corresponds to homozygous for minor allele which is very high. This SNP located in exon 16 of PXK*

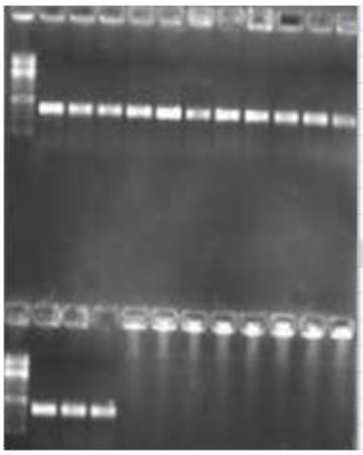
In order to study the splice variants of the gene, regions of exon 16 and promoter PCR amplified from spleen and PBMC cDNA of 16 very sick patients. To optimize the PCR conditions, pilot PCRs were run with different annealing temperatures of 63°C and 66°C, also different amounts of PCR enhancers. The templates for pilot PCRS were Daudi and Jurkat cell lines. Gel electrophoresis pictures of amplified regions are provided in Figure10. After EXO/SAP treatment on samples they were sent for sequencing at Uppsala Genome Center. Sequencher was used to read the chromatograms.

a)



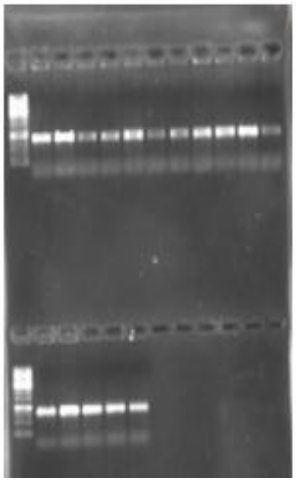
INDIVIDUALS ID	RS111374218 G/A	RS4681851 C/G	rs56384862 A/G
1472	G	C	G
1513	G	C	G
323-SGR	G	C/G	A/G
335-SGR	G	G	A/G
SLA003	G	C	G
318-SGR	G	C	G
344-SGR	G	C	G
SLA007	G	C	A
SLA27	G	C	A/G
SLA28	G	C	A/G
E14	G	G	A
G9	G	G	A
SL68	G	C/G	A
LP074	G	C	A
LP103	NO SEQUENCING DATA		
SLA009	G	C/G	A

b)



INDIVIDUALS ID	rs9815182 A/G	rs12637163 A/G	rs77380699 C/A	rs77009475 T/A
1472	A	A	C	T
1513	A	A	C	T
323-SGR	A/G	A/G	C	T
335-SGR	A	A	C	T
SLA003	A	A	C	T
318-SGR	A	A	C	T
344-SGR	A	A	C	T
SLA007	A	G	C	T
SLA27	NO SEQUENCING DATAT			
SLA28	A	A/G	C	T
E14	A	A	C	T
G9	A	A	C	T
SL68	A	A	C	T
LP074	A	A	C	T
LP103	A	A	C	T
SLA009	A	A	C	T

c)



INDIVIDUALS ID	rs78946071 G/T	rs76226186 C/T	rs114552058 G/C	rs79813245 G/A	rs73835170 A/G	rs13082844 G/A
1472	G	C	G	G	A	G
1513	G	C	G	G	A	G
323-SGR	G	C	G	G	A	G
335-SGR	G	C	G	G	A	G
SLA003	G	C	G	G	A	G
344-SGR	G	C	G	G	A	G
SLA27	G	C	G	G	A	G
SLA28	G	C	G	G	A	G
E14	G	C	G	G	A	G
G9	G	C	G	G	A	G
SL68	G	C	G	G	A	G
LP074	G	C	G	G	A	G
LP103	G	C	G	G	A	G
SLA009	G	C	G	G	A	G

Figure 10: *amplification of three regions in PBMC cDNA of 16 patients and their sequencing result. a) Gel picture of region amplified with primers forw-intr15/rev-intr16, without enhancer, annealing temperature of 63℃ and its sequencing reads of SNPs siting within that region. b) Gel picture of region amplified with primers forw-prom2/rev-prom2, 2.5µl PCR enhancer, annealing temperature of 63℃ ° and its sequencing reads of SNPs siting within that region. c) Gel picture of region amplified with primers forw-prom1/rev-prom1, with 1.2µl PCR enhancer, annealing temperature of 63° and its sequencing reads of SNPs siting within that region. (Ladder used in all gels 100bp Generuler)*

In most of the SNPs in the amplified regions there was no difference in genotypes of individuals to go for further studies. In some cases different alleles were detected.

Amplification of exon 1 region in Jurkat DNA was carried out. The strong upper band from first lane was separated as sequencing sample1, from the rest, upper bands and small lower bands separated as sequencing samples 2 and 3 respectively. The product was prepared for sequencing and result was analyzed. Gel picture and the sequencing result summarized in Figure 11 and Table 13.

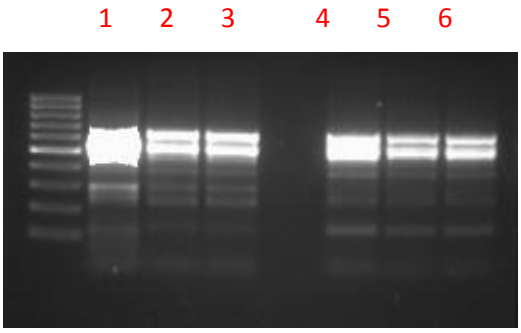


Figure 11: *Amplification of exon 1 region in Jurkat DNA. Primer set used in all lanes are the same to amplify exon 1; template was Jurkat DNA; lane 1: w/o PCR enhancer, lane 2: with 1.25µl PCR enhancer and lane 4, 5,6: with 2.5µl PCR enhancer.(Ladder 100bp Generuler)*

Table 9: *Sequencing results of extracted bands from gel demonstrated in Figure 11.*

	Description	Sequencing result
Sequencing sample1	Strong band from lane1	Non-specific band
Sequencing sample2	Second band from top of lane 2,3,4,5,6	Bad sequencing Un-readable
Sequencing sample3	Third band from top of lane 2,3,4,5,6	Bad sequencing Un-readable

Three other regions of exon1 to exon8, exon14 to 3'UTR and exon15 to exon17 were amplified in PBMC cDNA with three sets of primers comprising forw-exon1/rev-exon8, forw-exon14/rev-PXK-3'UTR and forw-exon15/rev-exon17. PCR products were purified from gel using Qiagen Gel Extraction kit and sequenced.

Additionally to characterize the 3' and 5' ends of the gene, RACE PCR (Rapid Amplification of cDNA Ends) on cDNA of PBMC was performed with primer set of Rev-exon8/AP2. Electrophoresis of RACE products showed four bands which were separated and purified from gel and sequenced after another run of PCR (re-amplifying).Gel pictures and sequencing result shown in Figure 12.

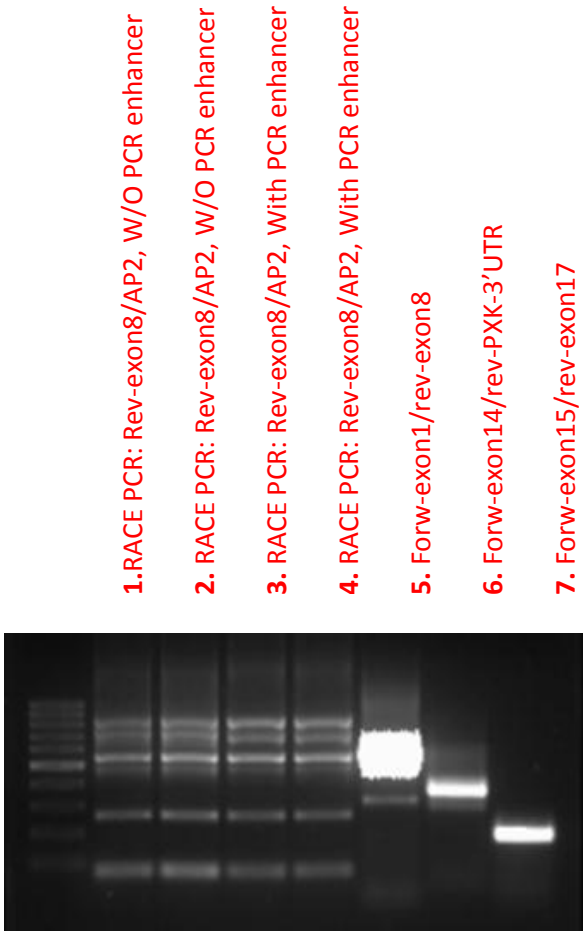


Figure 12: *Amplification of some regions of PBMC cDNA normal PCR and RACE PCR. Ladder: 100bp Generuler.*

PCR product extracted from gel and sent for sequencing in order described in Table 10 below; sequencing analysis result was also shown in same table.

Table 10: *Description of sequencing samples (PCR products extracted from gel showed in Figure 12) and their sequencing results.*

	Description	Sequencing result
Sequencing sample 1	First band from top of lanes 1, 2, 3 and 4 mixed	Bad sequencing, un-readable
Sequencing sample 2	Second band from top of lanes 1, 2, 3 and 4 mixed	Bad sequencing, un-readable
Sequencing sample 3	Third band from top of lanes 1, 2, 3 and 4 mixed	Bad sequencing, un-readable
Sequencing sample 4	Fourth band from top of lanes 1, 2, 3 and 4 mixed	Bad sequencing, un-readable
Sequencing sample 5	Fifth band from top of lanes 1, 2, 3 and 4 mixed	Bad sequencing, un-readable
Sequencing sample 6	Strong upper band from lane 5	Full length normal transcript
Sequencing sample 7	Second band from top of lane 5	Δ exon 2-4, normal transcript
Sequencing sample 8	Strong band from lane 6	Δ Ala-exon16 new isoform!
Sequencing sample 9	Strong band from lane 7	Δ Ala-exon16 new isoform!

As showed in table above five sequencing samples were un-readable so only samples 6-9 could be subjected to discussion.

Region around exon 16 and 17 within PXK gene in PBMC cDNA from a set of 10 healthy individuals was amplified using primer set covering from exon14 to rev-3'UTR. Two bands were detected in each sample and extracted from gel (Figure 13). Another run of re-amplification was done. In order to confirm the existence of different sized bands, re-amplified products were loaded on agarose gel (Figure 14) and then sent for sequencing.

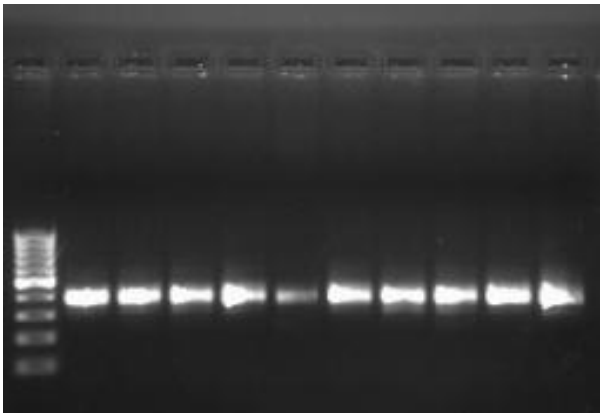


Figure 13: *Amplification of region of PXK from exon14 to 3'UTR in PBMC cDNA of 10 healthy individuals showing two bands in each lane. Upper bands are very strong but the smaller bands are week. (Ladder 100bp Generuler)*

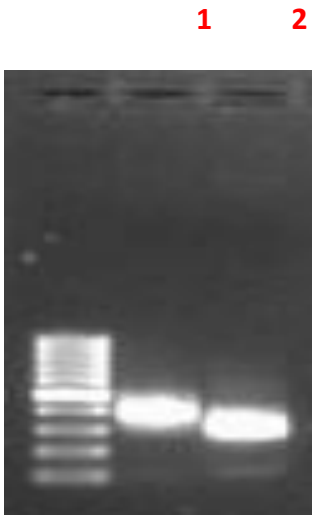


Figure 14: *Gel picture confirming of two bands with different size in amplification of region from exon 14 to 3'UTR showed in Figure 13. Lane 1: strong upper band and lane 2: lower band purified from gel showed in same figure. (Ladder 100bp Generuler)*

Figure 14 confirmed the existence of the second smaller band in amplification of region from exon 14 to 3'UTR. Sequencing result also confirmed this difference since smaller band did not contain exon 17 (Δ exon17). Sequencing results for longer band also showed a combination of two reads. One read was corresponding to normal transcript with exon16, 17 and 18; but the other read was missing Alanine amino acid from the beginning of the exon 17 (Δ Ala-exon17).

The sequence of Δ exon17 could be representing new isoform for PXK not reported before or it could be one of the previously reported truncated transcripts of the gene. More optimizations are needed to confirm this data.

Discussion:

The human PXK gene coding for PX domain containing serine/threonine kinase is located on chromosome 3p14.3 [3]. The full-length transcript contains 18 exons, while some truncated transcripts including 4-9 exons [4]. Some exons are missing in some of the transcripts (alternative exons) and some exons exist in all transcripts (constitutive exons). The list of alternative and constitutive exons in Table 5 was very helpful in transcript analysis**.**

The highest expression level of PXK is in brain, heart, skeletal muscle and peripheral blood lymphocytes [3]. In this experiment, spleen, thymus and PBMC were used with reasonable expression levels for the gene, making them applicable for this project.

Association of the PXK gene to SLE has been reported recently by several GWA studies by introducing one SNP (rs6445975) in population of women with European ancestry [1, 5, 7, and 8] but no data has been published on functional annotation of the gene. According to LD plots (Figure 1) and some recent unpublished data, there are evidences of association of some regions and variations of PXK to SLE. Regions of exon 1, exon 17 and especially of exon 16 of PXK and six SNPs listed in Table 2 were subjected to re-sequencing and genotyping in this experiment.

In amplification of regions of interest (containing six SNPs) in PBMC DNA of 96 healthy blood donors of Uppsala Hospital, different DNA polymerases have been used to get better efficiency on PCRs. DNA polymerases including Hifi, AmpliTaqGold and Platinum® Taq were used in the amplifications. Many rounds of optimization had to be done with different annealing temperatures and PCR enhancers. This step was very time consuming despite its simplicity. To clean up the PCR products for sequencing, EXO/SAP treatment was done on samples. The purpose of this step was to get rid of excess of primers and dNTPs. Even with purified PCR products, reading of sequencing results was not very good and the reads were unclear in many samples. The reason could be un-specific product from PCR, primers and dNTPs leftovers even after EXO/SAP treatment and also bad sequencing itself. Despite of all obstacles and limitations for this step, sequencing analysis for 4 out of 6 of SNPs of interest was done and genotyping data was evaluated statistically (Table 8). rs6772652 (A/G) showed P value of 0.01 for GG vs. AG genotypes, which is statistically significant. This is an evidence of correlation between genotypes and expression of alleles in the population for this SNP. To optimize this data extended experiment is valuable. However during the current project further efforts on genotyping more individuals were not successful. For the rest of SNPs, expressions did not show significant P values so there is no correlation between their expression and genotypes.

According to SNPexpress, rs11710823 which was mapped in exon 16 showed high expression for minor alleles (A/A) with very good P value (Figure 9). However current experiment did not support this data. Genotyping result did not show significant correlation for that SNP. One limitation of using SNPexpress and other bioinformatics tools is insufficient data for some variations or genes which are not well characterized yet; and in many cases data extracted from one database does not agree with others. Another limitation was using old version of gene browsers as their reference so it was not so easy to find the right region by mapping the probes used in their experiment.

The promoter of a gene is recognized by regulatory elements, so it is an important region in regulation of gene expression. Regarding this, and according to unpublished data which suggested exon 16 as associated region to pathogenesis, regions of promoter and exon16 of PXK in spleen and PBMC cDNA of group of 16 very sick individuals were amplified (Figure 10). Genotyping of variations mapped in amplified regions did not show any difference in most of the cases so it could not be subjected to further experiments. Finding optimum PCR condition for this part of experiment was also hard because of some regions of very high CG content.

DNA amplification of exon 1 in Jurkat cell line showed multiple bands on gel electrophoresis (Figure 11) which were extracted from gel, but sequencing result was not good. It was expected to observe some evidences of new isoforms and splicing patterns in this region and afterwards it could be investigated in population, but sequencing result was not reliable (Table 9).

RACE PCR (Rapid Amplification of cDNA Ends) on cDNA of PBMC was performed (Figure 12) with the purpose of characterizing the 3' end of the gene. Four amplified bands were purified from gel and sequenced after another run of PCR (re-amplifying). The purpose of re-amplification was to generate more DNA to improve the sequencing. Sequencing was unsuccessful for RACE products (Table 10).

Three other regions of exon1 to exon8, exon14 to 3'UTR and exon15 to exon17 were amplified in PBMC cDNA (Figure 12). PCR products were sequenced (Table 10). This observation introduced new isoform of Δ Ala-exon16. Δ Ala-exon16 refers to a transcript that lacks three nucleotides corresponding to an alanine amino acid in the beginning of exon16. This isoform was not reported before.

PBMC cDNA from a set of 10 healthy individuals were amplified from exon14 to 3'UTR (Figure 13). It was confirmed that there were two PCR products with different size (Figure 14). Sequencing result also confirmed this difference since smaller fragment did not contain exon 17 (Δ exon17). That Δ exon17 fragment could be truncated transcript PXK-009 which lacks exon 17, or it could be new isoform longer than PXK-009. Some efforts were made on clarifying the existence of this new isoform but data was not supportive.

Sequencing results for bigger amplified region revealed new transcript of the gene which was missing the alanine residue from the beginning of the exon 17 (Δ Ala-exon17).

In brief, the main achievements of current experiment are the confirmation of a correlation between expression and genotype of SNP rs6772652 (A/G), introducing susceptibility variation rs11710823 according to SNPexpress database but not supporting this data during experiment and Introducing two noble isoform for PXK lacking an alanine residue from the beginning of exon 16 (Δ Ala-exon16) and exon17 (Δ Ala-exon17).

Acknowledgments:

I had the honor to work under careful supervision of Sergey Kozyrev, who was a great teacher for me during this project and showed me the right way to face obstacles during the experiment.

Thanks to staff and members of Department of Medical Biochemistry and Microbiology (IMBIM) Uppsala University for providing me a perfect environment to do my degree project.

Many thanks to my coordinator Hakan Rydin and Lars Liljas for all of the support and helps they gave me during my education in Uppsala University.

I should say a very special thanks to my mother and dear sisters for being absolutely supportive in whole my life.

References:

1. M L Budarf, P Goyette, G Boucher, J Lian, R R Graham, J O Claudio, T Hudson, D Gladman, A E Clarke, J E Pope, C Peschken, C D Smith, J Hanly, E Rich, G Boire, S G Barr, M Zummer, GenES Investigators, P R Fortin, J Wither and J D Rioux. A targeted association study in systemic lupus erythematosus identifies multiple susceptibility alleles. Genes and Immunity (2011) 12, 51–58.
2. J Castro, E Balada, J Ordi-Ros , M Vilardell-Tarrés. The complex immunogenetic basis of systemic lupus erythematosus. Autoimmunity Reviews (2008)7, 345–351.
3. JL Huang, KW Yeh, TC Yao, YL Huang, HT Chung, LS Ou, WI Lee and LC Chen. Pediatric lupus in Asia. Lupus. 2010 Oct; 19(12):1414-8.
4. H-S. Lee and S-C. Bae. What can we learn from genetic studies of systemic lupus erythematosus? Implications of genetic heterogeneity among populations in SLE. Lupus. (2010) 19, 1452-9.
5. The International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN), John B Harley, M E Alarcón-Riquelme, L A Criswell, C O Jacob, R P Kimberly, K L Moser, B P Tsao, T J Vyse & C D Langefeld. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nature Genetics (2008) 40, 204-210.
6. B Yu, Q Wu, Y Chen, P Li, Y Shao, J Zhang, Q Zhong, X Peng, H Yang, X Hu, B Chen, M Guan, W Zhang and J Wan. Polymorphisms of PXK are associated with autoantibody production, but not disease risk, of systemic lupus erythematosus in Chinese mainland population. (2011) Lupus 20, 23.
7. M Suarez-Gestal et al. Replication of recently identified systemic lupus erythematosus genetic associations: a case-control study. Arthritis Res Ther. 2009; 11(3):R69. Epub 2009 May 14.
8. A L Sestak, B G Fürnrohr, J B Harley, J T Merrill, B Namjou. The genetics of systemic lupus erythematosus and implications for targeted therapy. Ann Rheum Dis. 2011 Mar; 70 Suppl 1:i37-43.
9. J M. VanLiere and N A. Rosenberg. Mathematical properties of the r2 measure of linkage disequilibrium. Theor Popul Biol. 2008 August; 74(1): 130–137.
10. N Kumasaka, Y Nakamura, N Kamatani.The Textile Plot: A New Linkage Disequilibrium Display of Multiple-Single Nucleotide Polymorphism Genotype Data. PloS One. 2010; 5(4): e10207.
11. B Devlin, N Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29: 311–322.
12. S Kozyrev et al. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. Nature Genetics 40, 211 - 216 (2008).
13. K Holm, E Melum, A Franke and T H Karlsen. SNPexp - A web tool for calculating and visualizing correlation between HapMap genotypes and gene expression levels. BMC Bioinformatics 2010, 11:600