Published in final edited form as:
Transfusion. 2016 December ; 56(12): 3047–3054. doi:10.1111/trf.13801.

Full length nucleotide sequence of *ERMAP* alleles encoding Scianna (SC) antigens

Kshitij Srivastava[1], **Eunah Lee**[1], **Eric Owens**[1], **Pairaya Rujirojindakul**[2], and **Willy A. Flegel**[1]

[1]Department of Transfusion Medicine, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA [2]Department of Pathology, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand

Abstract

Background—Scianna (SC) blood group system comprises 2 anthithetical antigens, Sc1 and Sc2, and 5 additional antigens. The antigens reside on a glycoprotein encoded by the erythroblast membrane-associated protein (*ERMAP*) gene. For the common *ERMAP* alleles, we determined the full length nucleotide sequence that encodes the Scianna glycoprotein.

Study design and methods—Blood donor samples from 5 populations were analyzed including 20 African Americans, 10 Caucasians, 10 Thai, 5 Asians and 5 Hispanics for a total of 100 chromosomes. An assay was devised to determine the genomic sequence of the *ERMAP* gene in 1 amplicon, spanning 21.4 kb and covering exon 2 to 12 and the intervening sequence (IVS). All alleles (confirmed haplotypes) were resolved without ambiguity.

Results—Among 50 blood donors, we found 80 single nucleotide polymorphisms (SNPs), including 6 novel SNPs, in 21,308 nucleotides covering the coding sequence of the *ERMAP* gene and including the introns. The non-coding sequences harbored 75 SNPs (68 in the introns; and 7 in the 3′UTR). No SNP indicative of a non-functional allele was detected. The nucleotide sequences for 48 *ERMAP* alleles (confirmed haplotypes) were determined by allele-specific PCR and sequencing in 100 chromosomes.

Address for correspondence: Willy A Flegel MD, Laboratory Services Section, Department of Transfusion Medicine, NIH Clinical Center, National Institutes of Health; Bethesda MD 20892, USA, waf@nih.gov, Phone : (301) 594-7401, FAX : (301) 496-9990.

Supporting Information
Additional Supporting Information may be found in the online version of this article:

Conflict of interest disclosure: The authors declared having no competing financial interest relevant to this article.

Statement of Disclaimer: The views expressed do not necessarily represent the view of the National Institutes of Health, the Department of Health and Human Services, or the U.S. Federal Government.

Authorship contribution: WAF conceived the study. KS designed and performed experiments. EL and EO completed the experimental part of this study. Data were analyzed by KS and WAF who also wrote the manuscript.

Web resources
dbSNP database, Build ID: 138 Phase I (http://www.ncbi.nlm.nih.gov/SNP/)
Hardy-Weinberg equilibrium calculator (http://www.tufts.edu/~mcourt01/Documents/Court%20lab%20-%20HW%20calculator.xls)
MaCH, version 1.0 (http://www.sph.umich.edu/csg/abecasis/MACH/index.html)
PredictSNP, version 1.0 (http://loschmidt.chemi.muni.cz/predictsnp/)
Primer3 software, version 0.4.0 (http://bioinfo.ut.ee/primer3-0.4.0/)
SeattleSNPs (http://pga.gs.washington.edu)

Conclusions—We documented 48 *ERMAP* alleles of 21,308 nucleotides each. The 2 nucleotide sequences available in GenBank for *ERMAP* alleles of similar length have not been found in our 100 chromosomes. Alleles determined without ambiguity can be used as templates to analyze next generation sequencing data, which will enhance the reliability in clinical diagnostics.

Introduction

The human *ERMAP* gene (erythroblast membrane-associated protein; MIM#609017) is located on chromosome 1 (1p34.2), 17.5 Mbp centromeric to *RHCE*, and encodes a single pass transmembrane adhesion/receptor glycoprotein.[1] The ERMAP glycoprotein is highly expressed on erythroid tissues and carries the 7 Scianna blood group antigens.[2–5] It is also weakly expressed on leukocytes; and found in the thymus, lymph nodes, spleen and bone marrow of adults and in fetal liver.[1]

The Scianna blood group system (SC; ISBT 013) comprises 2 antithetical antigens, the high-frequency antigen Sc1[6] and the low-frequency antigen Sc2.[7] Five other antigens, Sc3,[8] Sc4 (Rd),[3] Sc5 (STAR),[9] Sc6 (SCER)[10] and Sc7 (SCAN)[10] also reside on the ERMAP protein. Observations of allo-and auto-antibodies against Sc antigens and their clinical implications have been reviewed.[11]

The *ERMAP* RefSeqGene sequence, NG_008749.1, consists of 12 exons. The first 2 exons are non-coding. The sixth nucleotide in exon 3 represents the 'A' of the start codon of the coding sequence (CDS). The 475 amino acid protein is encoded by exons 3 to 12 from either a 3,485 bp (NM_001017922.1) or a 3,381 bp (NM_018538.3) mRNA transcript. Thus in the case of the *ERMAP* gene, nucleotide sequencing of genomic DNA starting in intron 2 and including exon 12 covers the entire CDS.

A comprehensive population-based collation of *ERMAP* alleles and their protein products was missing, because online databases such as dbSNP[12] and 1000 Genome Project[13] lack phase information. The allele (confirmed haplotype) information will be useful in determining the evolutionary history of the *ERMAP* gene. We describe the nucleotide variations in a large number of *ERMAP* alleles. The *ERMAP* alleles found in 100 chromosomes from 50 random individuals of 5 populations were resolved without ambiguity.

Materials and Methods

Blood samples

Blood samples from 20 African American, 10 Caucasian, 5 Hispanic and 5 Asian blood donors were collected at random in the NIH Blood Bank, along with 10 samples of β-thalassaemia major patients[14] (REC:57-0148-05-1 approval by the Institutional Ethics Committee of the Faculty of Medicine, Prince of Songkla University, Thailand). Genomic DNA was extracted from EDTA anticoagulated whole blood (EZ1 DNA blood kit on the BioRobot EZ1 Workstation; Qiagen, Valencia, CA). The explorative study was restricted to 50 samples, similar to previous comparable approaches.[15,16] We over-represented individuals with African descent as they are known to carry many polymorphic alleles, and we also included a representative number of Caucasian and Asian samples. The 10 Asian samples came from a recently published study.[14]

ERMAP gene amplification

For our study, we designed a sequencing approach capturing the whole 3,381 bp NM_018538.3 mRNA transcript including the non-coding exon 2, but excluding the non-coding exon 1. A 21,406 nucleotide stretch of *ERMAP* gene was amplified as a single amplicon using 100 ng of genomic DNA. The amplification was done using the long range Taq polymerase (LongAmp Taq DNA Polymerase; New England Biolabs, Ipswich, MA, USA) along with the primers 5′-GTGCTCCATGAGTCAAGCGATTAC-3′ and 5′-TACCTTCCCCACAACTCCTCATTC-3′ (Eurofins MWG Operon; Huntsville, AL). Thermocycling conditions were: initial denaturation at 94 °C for 2 min; 35 cycles of 94 °C for 30 sec, 62.9 °C for 30 sec, 68 °C for 22 min and a final extension at 68 °C for 5 min (DNA Engine Tetrad 2 Peltier Thermal Cycler; Bio-Rad, Hercules, CA). This primary amplicon covered the 1,428 nucleotides of the CDS in the exons 3 to 12 and, in addition, 1,993 bp of intron 1, 116 bp of the non-coding exon 2 and 5 non-coding nucleotides of exon 3, 1,803 bp of the 3′-UTR and 16,061 bp of the introns 2 to 11.

Nucleotide sequencing

The primers for sequencing were designed using Primer3 (Table S1).[17] The primary amplicons were purified and sequenced as previously described[18] with extensive confirmatory resequencing. Nucleotide sequences were aligned (CodonCode Aligner; CodonCode, Dedham, MA) to NCBI RefSeq NG_008749.1 and nucleotide positions defined using the first nucleotide of the CDS of NM_001017922.1. The genotype sequence of all 50 samples was determined for 21,308 nucleotides.

Determination of haplotypes

The unphased genotype data obtained from sequencing the samples was used as input in the Markov Chain based haplotyper MaCH 1.0[19] software to infer the *ERMAP* alleles (haplotypes). Briefly, the software starts by randomly generating a pair of haplotypes, compatible with observed genotypes, for each sampled individual. These initial haplotypes are then refined using Hidden Markov Model (HMM)-based iterations that describe the haplotype pair as an imperfect mosaic of the other haplotypes. After a number of iterations, typically 20 to 100 steps, the consensus haplotypes are constructed by merging the haplotypes sampled in each round.

Confirmation of haplotypes

Complete homozygosity or heterozygosity at a single site allowed umambiguous assignment of a haplotype.[20] Allele-specific PCR and subsequent sequencing of the PCR products was used to construct haplotype structure in samples with more than one heterozygous site. Briefly, 10 allele-specific PCR primers (Table S1) were designed for the first and last heterozygous sites found in the amplicons of 34 individuals. Long range allele-specific PCRs, nested in the primary 21,406 nucleotide amplicon, were carried out and all variant positions between the first and last heterozygous sites were sequenced in an allele-specific way.

Computational modeling of amino acid substitutions

PredictSNP was used to predict the functional impact of non-synonymous amino acid substitutions.[21] We used PredictSNP to determine a consensus prediction for a given SNP. PredictSNP is a metaserver that combines experimental annotations from Protein Mutant Database and UniProt with the predicted outcomes from 6 *in silico* prediction tools: MAPP (Multivariate Analysis of Protein Polymorphism), PhD-SNP (Predictor of human Deleterious SNP), PolyPhen-1 (Polymorphism Phenotyping-1), PolyPhen-2 (Polymorphism Phenotyping-2), SIFT (Sorting Intolerant From Tolerant) and SNAP (Screening for Non-Acceptable Polymorphisms).

Statistical description

95% confidence intervals (CI) for allele frequencies were calculated using the Poisson distribution.[22]

Results

A random survey in 50 individuals was performed to describe the genetic variability of the *ERMAP* gene and to determine a large number of alleles (confirmed haplotypes). We determined the genomic sequence of 21,308 nucleotides of the *ERMAP* gene in each individual. Among the 1,065,400 nucleotides such sequenced, we observed a total of 80 positions with SNPs (Fig. 1).

Nucleotide variations

Among the 80 SNPs observed, 5 occurred in the CDS, 7 in 3′UTR and 68 in the 11 introns (Table S2). In the CDS, 2 SNPs were non-synonymous and 3 were synonymous. No non-sense or splice site mutation was detected, and 6 intronic SNPs were novel, not previously documented in the dbSNP database.

Genotype patterns and haplotype determination

In the 50 individuals, 46 distinct genotype patterns were observed (Table S3). One African American, Caucasian and Thai individual each was observed as being homozygous for an *ERMAP* allele, allowing us the unambiguous assignment of 3 distinct alleles (x, y and z, respectively). Another 2 Thai individuals were heterozygous for 2 of the distinct alleles (y and z). In another 10 individuals, the predicted haplotypes *in trans* to 1 of the 3 distinct alleles were amplified with allele-specific PCR primers and confirmed by sequencing. In the remaining 35 individuals, the 2 alleles were amplified by allele-specific primers, and all heterozygous sites were sequenced to determine the alleles. Among the 100 chromosomes studied, we determined 48 alleles (Fig. 2) and deposited 21,308 nucleotides for each of the 48 alleles in the GenBank database (Table 1).

ERMAP alleles

All 48 alleles detected carried the SNP indicative of the common SC:1 phenotype. No SNPs characteristic of the other 6 Scianna phenotypes were observed (SC:2, SC:4, SC:-3, SC:-5, SC:-6 and SC:-7). No SNP encoding a non-sense mutation or a frame-shift mutation was observed.

Comparison with previous *ERMAP* sequences

There were 2 sequences in the nucleotide databases exceeding in length the DNA stretch analyzed by us (Table 3), including the RefSeqGene sequence for the *ERMAP* gene (NG_008749.1). Both of the published *ERMAP* sequences were not observed as alleles in our study. Among the 7 and 19 nucleotides that differed from any of our 48 alleles, 8 SNPs in RefSeqGene sequence (NG_008749.1) and 1 SNP in DQ090843.1 may be very rare in humans, if they occur at all (Table 3).

Effect on protein structure

Computational modeling predicted structural changes induced by the 2 non-synonymous SNPs, p.Ala4Val (rs35757049) and p.His26Tyr (rs33953680), to be neutral (Table 2).

Computerized haplotype prediction

With the genotype information (Table S3) as input, the MaCH software predicted 54 haplotypes. Using allele-specific PCR and sequencing, 48 *ERMAP* alleles were identified (Table 1). Out of these 48 alleles, only 42 alleles (87.5%) were correctly predicted by MaCH, as confirmed by our allele-specific PCR (Table S4). In 6 individuals, the 12 alleles predicted by MaCH, each calculated by the software to occur only once among the 100 chromosomes, were not confirmed. In these 6 individuals however, 6 alleles (12.5%), not predicted by MaCH, were identified by our allele-specific PCR.

Discussion

The aim of this study was to determine alleles (confirmed haplotypes) of *ERMAP* gene in general population. We sequenced 21,308 nucleotides of the *ERMAP* gene and identified 80 SNPs and 48 alleles in 50 individuals from 5 populations. This is the first study to systematically categorize SNPs found at the *ERMAP* gene locus into defined alleles.

The dbSNP database[12] lists more than 1700 nucleotide variations for the *ERMAP* gene. In the present study, we observed 80 SNPs in 50 individuals from 5 populations. Many variants described in the dbSNP database, although not observed in our study, may not be polymorphic in the populations we studied or so rare that our screening panel lacked adequate power to detect them. We also did not detect any variant associated with a non-functional ERMAP protein, although these have been reported in literature and online databases such as Ensembl.[23] Interestingly, the 68 intronic variations were found to be non-randomly distributed along the length of *ERMAP* gene, being mostly concentrated in introns 2, 4, 7 and 11. The relative dearth of sequence variations in other introns indicate that these regions may have high functional constraints, which are yet to be discerned.

A large number of *ERMAP* alleles identified in our samples were observed with a low frequency (Table 1). Many of these rare alleles are population specific and more prevalent in our African American cohort, correlating with its more diverse genetic background.[24] Because we defined alleles by a large region spanning more than 21 kb, most of the individuals have unique alleles with limited population overlap. All these features, if well consolidated in suitable databases, will be useful for precision medicine using high throughput technologies such as next generation sequencing (NGS).[25]

The MaCH algorithm inferred 54 alleles, of which only 42 were confirmed by using our assay with allele-specific PCR and nucleotide sequencing from 1 chromosome (Table S4). The algorithm computed 12 alleles that were not present and missed 6 alleles that were recognized by our assay (Fig. 2). These prediction errors always concerned rare alleles with 1 observation only (mean frequency 1%). Thus, depending on the ethnicity of the population, computerized allele prediction approaches may replace physical sequencing approaches for determining common alleles. However, the rare alleles must be molecularly haplotyped due to problems related to imputations of untyped SNPs, as illustrated by our results. Data sets like the current one may aid in validating complicated allele prediction.

The 1000 Genome Project lists many more alleles of the *ERMAP* gene which are inferred using the *in silico* algorithms.[13] Due to the large number of samples tested, currently encompassing 2,504 samples, it is expected that most of these inferred alleles, though rare, are still correct due to haplotype redundancy. In the present study, we used a small sample size of 50 individuals (100 chromosomes) to experimentally identify alleles (confirmed haplotypes) for the *ERMAP* gene locus. The *ERMAP* alleles that are unambiguously identified in our study can be applied in refining the 1000 Genome data on the *ERMAP* gene. The comprehensive data set presented here will lend itself to analyze the evolutionary relationships of *ERMAP* alleles, such as constructing a phylogeny tree. This approach has previously been applied, for example, to predict an intermediate allele of the *RHD* gene,[26] later confirmed by observation[27] and dubbed DFV,[28] and also to several intermediate alleles of the *ACKR1* gene.[29] Comprehensive data on actually observed and on inferred, possibly extant, alleles can be used as templates and will such facilitate the analysis of next generation sequencing (NGS) data.

The prediction of an amino acid substitution to affect protein structure using PredictSNP for the 2 non-synonymous variants p.Ala4Val and p.His26Tyr turned out to be neutral with more than 65% prediction accuracy (Table 2). The result should be interpreted with caution as there is no known protein structure available for ERMAP.

Further studies may overcome 2 limitations possibly relevant for Scianna antigen prediction. Because we excluded the non-coding exon 1 and the associated promoter, any variation was missed which occurred in the 5′UTR of the longer *ERMAP* transcript NM_001017922.1. Moreover, all the samples tested had the common SC*01 allele and were predicted to have the SC:1 phenotype. Although this prediction may well be correct, it could not be serologically confirmed. Mutations in the untested regions of the promoter might affect the expression of *ERMAP* transcripts, even abolishing the expression of ERMAP on the RBC surface. The amino acid mutation closest to the SC:1 SNP, was 32 amino acid positions removed. This or any other amino acid substitution may still affect the expression of the SC:1 antigen.

We developed and applied an assay to genotype and phase 21,308 nucleotides of the *ERMAP* gene using genomic DNA. The present study is the first to experimentally verify the haplotypes of a clinically relevant large stretch of the *ERMAP* gene in the general US population and in a set of 10 individuals from Thailand.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Harvey G. Klein for critical review and Elizabeth J. Furlong for English editing. Eunah Lee and Eric Owens participated in the study during their Summer Internship Program at NIH in 2014. This work was supported by the Intramural Research Program (project ID Z99 CL999999) of the NIH Clinical Center.

References

1. Su Y-Y, Gordon CT, Ye T-Z, et al. Human ERMAP: An Erythroid Adhesion/Receptor Transmembrane Protein. Blood Cells, Molecules, and Diseases. 2001; 27:938–49.
2. Lewis M, Kaita H, Chown B. Scianna blood group system. Vox Sang. 1974; 27:261–4. [PubMed: 4415694]
3. Wagner FF, Poole J, Flegel WA. Scianna antigens including Rd are expressed by ERMAP. Blood. 2003; 101:752–7. [PubMed: 12393480]
4. Xu H, Foltz L, Sha Y, et al. Cloning and characterization of human erythroid membrane-associated protein, human ERMAP. Genomics. 2001; 76:2–4. [PubMed: 11549310]
5. Su YY, Gordon CT, Ye TZ, et al. Human ERMAP: an erythroid adhesion/receptor transmembrane protein. Blood Cells Mol Dis. 2001; 27:938–49. [PubMed: 11783959]
6. Schmidt RP, Griffitts JJ, Northman FF. A new antibody, anti-Sm, reacting with a high incidence antigen. Transfusion. 1962; 2:338–40. [PubMed: 13908792]
7. Anderson C, Hunter J, Zipursky A, et al. An antibody defining a new blood group antigen, Bu-a. Transfusion. 1963; 3:30–3. [PubMed: 14012816]
8. Nason SG, Vengelen-Tyler V, Cohen N, et al. A high incidence antibody (anti-Sc3) in the serum of a Sc:-1,-2 patient. Transfusion. 1980; 20:531–5. [PubMed: 7423592]
9. Hue-Roye K, Chaudhuri A, Velliquette RW, et al. STAR: a novel high-prevalence antigen in the Scianna blood group system. Transfusion. 2005; 45:245–7. [PubMed: 15660834]
10. Flegel WA, Chen Q, Reid ME, et al. SCER and SCAN: two novel high-prevalence antigens in the Scianna blood group system. Transfusion. 2005; 45:1940–4. [PubMed: 16371048]
11. Brunker PA, Flegel WA. Scianna: the lucky 13th blood group system. Immunohematology. 2011; 27:41–57. [PubMed: 22356519]
12. Sherry ST, Ward M-H, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–11. [PubMed: 11125122]
13. The Genomes Project C. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
14. Rujirojindakul P, Flegel WA. Applying molecular immunohaematology to regularly transfused thalassaemic patients in Thailand. Blood Transfus. 2014; 12:28–35. [PubMed: 24120606]
15. Wagner FF, Moulds JM, Tounkara A, et al. RHD allele distribution in Africans of Mali. BMC Genet. 2003; 4:14. [PubMed: 14505497]
16. Toomajian C, Kreitman M. Sequence variation and haplotype structure at the human HFE locus. Genetics. 2002; 161:1609–23. [PubMed: 12196404]

17. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3–new capabilities and interfaces. Nucleic Acids Res. 2012; 40:e115. [PubMed: 22730293]
18. Srivastava K, Almarry NS, Flegel WA. Genetic variation of the whole ICAM4 gene in Caucasians and African Americans. Transfusion. 2014; 54:2315–24. [PubMed: 24673173]
19. Li Y, Willer CJ, Ding J, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816–34. [PubMed: 21058334]
20. Chen Q, Srivastava K, Ardinski SC, et al. Full-length nucleotide sequences of 30 common SLC44A2 alleles encoding human neutrophil antigen-3. Transfusion. 2016; 56:729–36. [PubMed: 26437811]
21. Bendl J, Stourac J, Salanda O, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014; 10:e1003440. [PubMed: 24453961]
22. Sachs, L. Angewandte Statistik - Anwendung statistischer Methoden. 7th. Vol. 1992. Springer-Verlag; Berlin: p. 446-7.
23. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. Nucleic Acids Research. 2014; 42:D749–D55. [PubMed: 24316576]
24. Campbell MC, Tishkoff SA. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annual review of genomics and human genetics. 2008; 9:403–33.
25. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. bioRxiv. 2015
26. Wagner FF, Ladewig B, Angert KS, et al. The DAU allele cluster of the RHD gene. Blood. 2002; 100:306–11. [PubMed: 12070041]
27. Noizat-Pirenne F, Lee K, Pennec PY, et al. Rare RHCE phenotypes in black individuals of Afro-Caribbean origin: identification and transfusion safety. Blood. 2002; 100:4223–31. [PubMed: 12393640]
28. Flegel WA, Von Zabern I, Doescher A, et al. DCS-1, DCS-2, and DFV share amino acid substitutions at the extracellular RhD protein vestibule. Transfusion. 2008; 48:25–33. [PubMed: 17900276]
29. Schmid P, Ravenell KR, Sheldon SL, et al. DARC alleles and Duffy phenotypes in African Americans. Transfusion. 2012; 52:1260–7. [PubMed: 22082243]

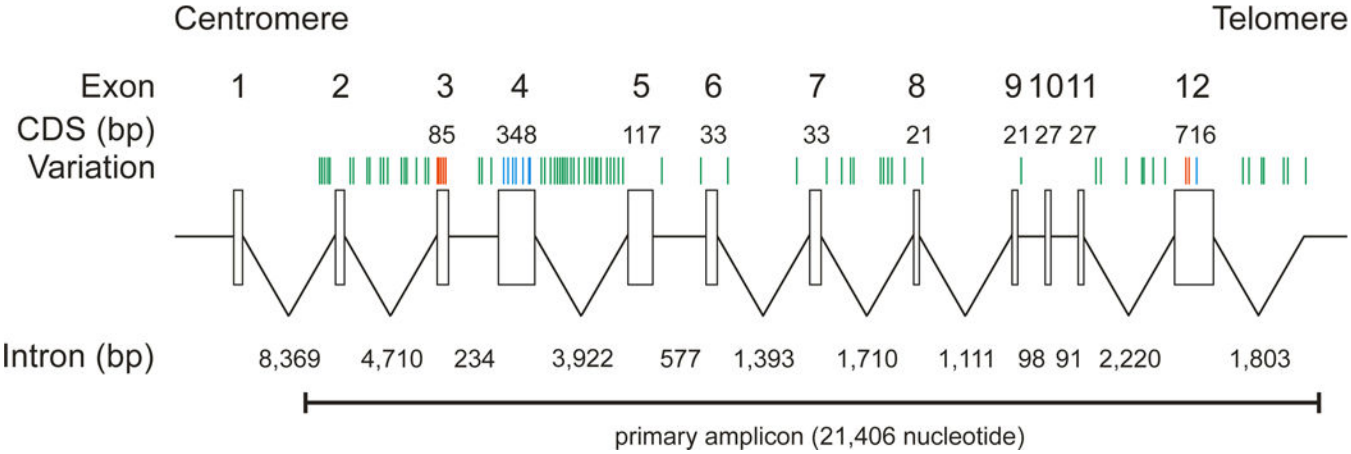


Figure 1. Genetic variations in the *ERMAP* gene

The *ERMAP* gene is located at chromosome 1p34.2 and consists of 12 exons encoding a protein of 475 amino acids. The exons (▯) are shown schematically along with their 5′-UTR, 3′-UTR and 11 introns (——). The primary 21,406 nucleotide amplicon covered the complete 1,428 nucleotide coding sequence (CDS) in the exons 3 to 12. It also encompassed 1,941 bp of intron 1, 116 bp of the non-coding exon 2, 1,757 bp of the 3′-UTR and 16,066 bp of the introns 2 to 11. The positions of the 80 variations (SNPs) are indicated (|) in the exons (red) and introns (green). The 7 additional SNPs in the exons indicative for the Scianna phenotypes SC:1, SC:2, SC:-3, SC:4, SC:-5, SC:-6 and SC:-7 were not observed in this study (blue).

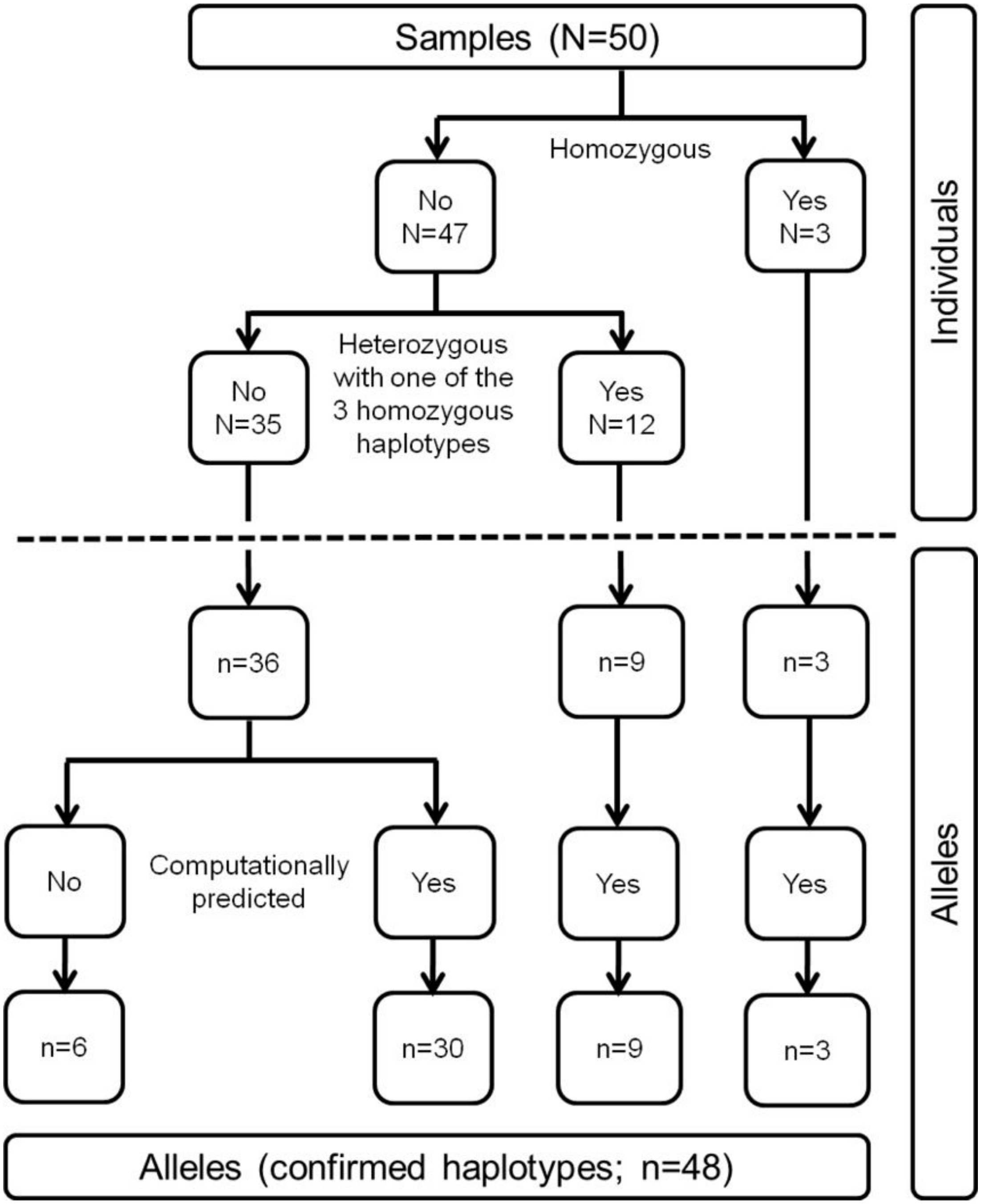

Samples (N=50)
Homozygous
No
N=47
Yes
N=3
Heterozygous with one of the 3 homozygous haplotypes
No
N=35
Yes
N=12
Individuals
n=36
n=9
n=3
Computationally predicted
No
Yes
Yes
Yes
n=6
n=30
n=9
n=3
Alleles
Alleles (confirmed haplotypes; n=48)

Figure 2. Flow diagram of *ERMAP* alleles identified in this study

Based on our nucleotide sequencing results, the 50 samples that were entered into the study were split in 3 groups of 3, 12 and 35 samples. In the group of 3 homozygous samples, 3 alleles were found. In the group of 12 samples, 9 alleles were confirmed. In the group of 35 samples, 36 alleles were confirmed, 30 of which had been computationally predicted and 6 of which were only found based on additional nucleotide sequencing by nested PCR. All 48 alleles (see Table 1) had thus been confirmed by physical nucleotide sequencing from single chromosomes.

Table 1

ERMAP alleles found in this study

Allele number	Allele (confirmed haplotype)*	GenBank Number	Observations (n)
1	ATTGGCACCAGGCCGCCGCCCTGCTTAAGCCCTGGCGTGGTACTCGTCACGGTCCGCCGGGGCCGGATTAAA	KX265235	8
2	-C----G----------A---G----T---T---A-A---C-T-TC---G--------A--A----------	KX265236	12
3	---G-A-------A------------	KX265189	7
4	G-----G---A--------TTG--G--G--T-----------------T-C--------------------G	KX265190	5
5	-C----G--------------G-------------T-------CT----G-A-------A------------	KX265191	5
6	-C----G----------A---G--------TG--A-A---C---TC---G--------A--A----------	KX265192	4
7	----------------------------A-----------------------C----------G---C-GG-	KX265193	3
8	------G--------------G-------------T------------------------------------	KX265194	3
9	-C----G----------A---G---GT---T---A-A---C-T-TC---G--------A--A----------	KX265195	3
10	-C----G----------A---G---GT---T---A-A-----T-TC---G--------A--A----------	KX265196	3
11	-C----G----------A---G----T---T---A-A---C-T-TC---G-A-------A------------	KX265197	3
12	---------------------G--	KX265198	2
13	---TTC----------------	KX265199	2
14	--A-------------	KX265200	2
15	-C----G----------A---G--------T---A-A---C-T-TC---G--------A--A----------	KX265201	2
16	-C----G----------A---G--------T-----A---C---T----G-A-------A------------	KX265202	2
17	-C----G----------A---G----T---T---A-A---C-T-TC--------------------------	KX265203	2
18	G-----G---A--------TTG--G--G--T---	KX265204	2
19	-C----G--------------G-------------T------------------------------------	KX265205	1
20	-----------A--A-------------	KX265206	1
21	--------------------------------------A---------T-C-------------A------G	KX265207	1
22	------G--------------G--------T--------A-GT------G--------A--A----------	KX265208	1
23	------G--------------G--------T--------A-GT----------TTC----------------	KX265209	1
24	G-----G---A--------TTGAAG-----T--------------------A-------A------------	KX265210	1
25	------GT---------A---G--------T-C-A-A-A--GT-T---------------------------	KX265211	1

Allele number	Allele (confirmed haplotype)*	GenBank Number	Observations (n)
26	-CC---G-----G----AT--G----T---T---ATA-----T-TC--------------------------	KX265212	1
27	-----TG------TC--A---G----T---T-----------T-T----G--------A-AA----------	KX265213	1
28	------G----------A---G---G----T-C-A-A----GT-T---T-C--------------------G	KX265214	1
29	------G----------A---G---G----T-C-A-A-A--GT-T----G--------A--A----------	KX265215	1
30	------G--G-------A---G--------T-C-A-A-A--GT-T---TGC--------------------G	KX265216	1
31	---G--------A--A----------	KX265217	1
32	--C---G-----G--T-AT--G----T---------A------CTC---G-A-----G-A--G--A------	KX265218	1
33	------G--------------G---------------------------G-A-------A------------	KX265219	1
34	G-----G---AA-------TTG--G--G--T--A---C------T--TT-C-----A-A------------G	KX265220	1
35	------G----------A---G--------T----------GT-T---------------------G-C---	KX265221	1
36	G-----G---A--------TTG--G--G-TT---	KX265222	1
37	G-----G---A--------TTG--G--G--T------------------G--------A--A----------	KX265223	1
38	G-----G---AA-------TTG--G--G--T--A---C------T--TT-C-----A-A------------G	KX265224	1
39	----T-G-------------TG--G-TG--T---------C-------------------------------	KX265225	1
40	---A----------------------T-------------------C--G-A-------A------------	KX265226	1
41	-C----G----------A---G--------T---A-A------CT-------------A-------G-----	KX265227	1
42	-C----G--------------G-------------T-------CT----G----------------G-----	KX265228	1
43	--------T--G--------A--A----------	KX265229	1
44	------G---------T---TG--G--G--T---------C------T------------------------	KX265230	1
45	------G--------------G-------------T------T-----------------------------	KX265231	1
46	------G---A--------TTG--G-----T------------------G-A-------A------------	KX265232	1
47	-C----G----------A---G---G----T---A-A---C-T-TC---G------A-A--A----------	KX265233	1
48	-C----G----------A---G----T---T---A-A---C-T-TC---G--------AA-A----------	KX265234	1

[*] The nucleotides at the 72 SNP positions with variations are shown in 5′ to 3′ orientation (Table S2). The remaining 21,236 nucleotide positions that we determined had no variation relative to our reference sequence KX265235.

Table 2

ERMAP allele frequencies

	Allele frequency							
	Observations (n)							
GenBank Number	African American	Caucasian	Thai	Hispanic	Asian	All	Mean[†]	95% CI[‡]
KX265235	4	2	1	0	1	8	8%	3.3–14.9
KX265236	3	1	7	0	1	12	12%	6.7–20.3
KX265189	0	2	3	1	1	7	7%	3.3–13.8
KX265190	2	1	0	1	1	5	5%	1.9–11.2
KX265191	1	2	0	1	1	5	5%	1.9–11.2
KX265192	4	0	0	0	0	4	4%	1.4–9.6
KX265193	3	0	0	0	0	3	3%	0.8–8.1
KX265194	1	0	2	0	0	3	3%	0.8–8.1
KX265195	1	0	0	2	0	3	3%	0.8–8.1
KX265196	0	3	0	0	0	3	3%	0.8–8.1
KX265197	0	0	3	0	0	3	3%	0.8–8.1
KX265198	2	0	0	0	0	2	2%	0.4–6.7
KX265199	2	0	0	0	0	2	2%	0.4–6.7
KX265200	1	1	0	0	0	2	2%	0.4–6.7
KX265201	1	1	0	0	0	2	2%	0.4–6.7
KX265202	1	0	0	0	1	2	2%	0.4–6.7
KX265203	0	1	1	0	0	2	2%	0.4–6.7
KX265204	0	0	0	2	0	2	2%	0.4–6.7
KX265205	1	0	0	0	0	1	1%	0.1–5.3
KX265206	1	0	0	0	0	1	1%	0.1–5.3
KX265207	1	0	0	0	0	1	1%	0.1–5.3
KX265208	1	0	0	0	0	1	1%	0.1–5.3
KX265209	1	0	0	0	0	1	1%	0.1–5.3
KX265210	1	0	0	0	0	1	1%	0.1–5.3
KX265211	1	0	0	0	0	1	1%	0.1–5.3
KX265212	1	0	0	0	0	1	1%	0.1–5.3

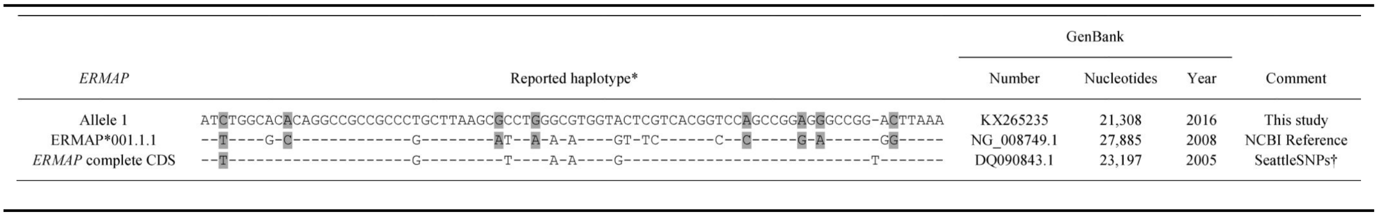
	Allele frequency							
	Observations (n)							
GenBank Number	African American	Caucasian	Thai	Hispanic	Asian	All	Mean[†]	95% CI[‡]
KX265213	1	0	0	0	0	1	1%	0.1–5.3
KX265214	1	0	0	0	0	1	1%	0.1–5.3
KX265215	1	0	0	0	0	1	1%	0.1–5.3
KX265216	1	0	0	0	0	1	1%	0.1–5.3
KX265217	1	0	0	0	0	1	1%	0.1–5.3
KX265218	1	0	0	0	0	1	1%	0.1–5.3
KX265219	0	1	0	0	0	1	1%	0.1–5.3
KX265220	0	1	0	0	0	1	1%	0.1–5.3
KX265221	0	1	0	0	0	1	1%	0.1–5.3
KX265222	0	1	0	0	0	1	1%	0.1–5.3
KX265223	0	1	0	0	0	1	1%	0.1–5.3
KX265224	0	1	0	0	0	1	1%	0.1–5.3
KX265225	0	0	1	0	0	1	1%	0.1–5.3
KX265226	0	0	1	0	0	1	1%	0.1–5.3
KX265227	0	0	1	0	0	1	1%	0.1–5.3
KX265228	0	0	0	1	0	1	1%	0.1–5.3
KX265229	0	0	0	1	0	1	1%	0.1–5.3
KX265230	0	0	0	1	0	1	1%	0.1–5.3
KX265231	0	0	0	0	1	1	1%	0.1–5.3
KX265232	0	0	0	0	1	1	1%	0.1–5.3
KX265233	0	0	0	0	1	1	1%	0.1–5.3
KX265234	0	0	0	0	1	1	1%	0.1–5.3
Total (n)	40	20	20	10	10	100		

[†] Number of observed alleles (n)/Total number of alleles.

[‡] 95% confidence interval (CI), Poisson distribution, two sided.[22]

Table 3

The prevalent *ERMAP* alleles compared with the 2 published alleles of similar length



ERMAP	Reported haplotype[*]	GenBank			Comment
		Number	Nucleotides	Year	
Allele 1	ATCTGGCACACAGGCCGCCGCCCTGCTTAAGCGCCTGGGCGTGGTACTCGTCACGGTCCAGCCGGAGGGCCGG-ACTTAAA	KX265235	21,308	2016	This study
ERMAP*001.1.1	--T----G-C-------------G--------AT--A-A-A----GT-TC------C--C-----G-A------GG-----	NG_008749.1	27,885	2008	NCBI Reference
ERMAP complete CDS	--T--------------------G---------T----A-A----G---------------------------T-------	DQ090843.1	23,197	2005	SeattleSNPs[†]

[*] The 8 SNPs in NG_008749.1 and 1 SNP in DQ090843.1 (grey), which were not observed in any of the 48 confirmed haplotypes (see Table 1): c. −121−795T>C; rs6600425, c.−6+1835C>A; rs1471747, c.433+1239A>G; rs11210725, c.433+1548A>G; rs10789427, c.617−398C>A; rs11210728, c.712+250G>A; rs1466549, c.713−606A>G; rs11210730 and c.*939G>C; rs10789428. One nucleotide insertion (c.*484_*485insT; rs55986603) in DQ090843.1 (black) was also not observed among any of the 48 alleles.

[†] SeattleSNPs. NHLBI HL66682 Program for Genomic Applications

n.a. — not applicable

Table 4

Functional significance of non-synonymous SNPs predicted by PredictSNP

	Variation		Computational analysis results	
dbSNP reference no	**Nucleotide change**[*]	**Amino acid substitution**[†]	**Classification**	**Expected accuracy (%)**[‡]
rs35757049	c.11C>T	p.Ala4Val	Neutral	68
rs33953680	c.76C>T	p.His26Tyr	Neutral	65

[*] relative to NCBI Reference Sequence NM_001017922.1

[†] relative to NCBI Reference Sequence NP_001017922.1

[‡] normalized confidence as calculated by the software (PredictSNP)