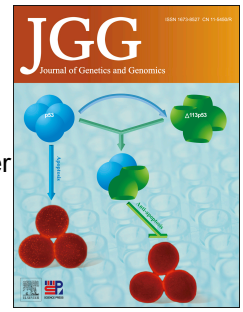
Accepted Manuscript



Serological Targeted Analysis of an ITIH4 Peptide Isoform: A Preterm Birth Biomarker and Its Associated SNP Implications

Zhou Tan, Zhongkai Hu, Emily Y. Cai, Cantas Alev, Ting Yang, Zhen Li, Joyce Sung, Yasser Yehia El-Sayed, Gary M. Shaw, David K. Stevenson, Atul J. Butte, Guojun Sheng, Karl G. Sylvester, Harvey J. Cohen, Xuefeng B. Ling

PII: S1673-8527(15)00098-3

DOI: 10.1016/j.jgg.2015.06.001

Reference: JGG 373

To appear in: *Journal of Genetics and Genomics*

Received Date: 31 March 2015

Revised Date: 2 June 2015

Accepted Date: 5 June 2015

Please cite this article as: Tan, Z., Hu, Z., Cai, E.Y., Alev, C., Yang, T., Li, Z., Sung, J., El-Sayed, Y.Y., Shaw, G.M., Stevenson, D.K., Butte, A.J., Sheng, G., Sylvester, K.G., Cohen, H.J., Ling, X.B., Serological Targeted Analysis of an ITIH4 Peptide Isoform: A Preterm Birth Biomarker and Its Associated SNP Implications, *Journal of Genetics and Genomics* (2015), doi: 10.1016/j.jgg.2015.06.001.

Serological Targeted Analysis of an ITIH4 Peptide Isoform: A Preterm Birth Biomarker and Its Associated SNP Implications

Over 11% of all pregnancies in the US result in preterm birth, greatly contributing to perinatal morbidity and mortality (Goldenberg and Rouse, 1998). Preterm birth etiologies remain largely unknown, and effective prevention methods have yet to be developed. The use of biofluid (e.g., serum or urine) for the analysis of the naturally occurring peptidome (MW < 4000) as a source of biomarkers has been reported for different diseases (Villanueva et al., 2006; Ling et al., 2010a, b, c, 2011). Mass spectrometry-based profiling of naturally occurring peptides can provide an extensive inventory of serum peptides derived from either high-abundant endogenous circulating proteins or cell and tissue proteins (Liotta and Petricoin, 2006). These peptides are usually soluble and unaffected by endogenous proteases or peptidases, and can be directly analyzed by liquid chromatography-mass spectrometry (LCMS) without additional manipulation (e.g., tryptic digestion). Esplin et al. (2011) used serum peptidomic patterns to identify patients with preterm birth. A peptide (QLGLPGPPDVPDHAAYHPF) derived from inter-alpha-trypsin inhibitor heavy chain 4 (ITIH4) was identified as a preterm birth biomarker from a predominantly (~75%) African-American cohort. The serum level of ITIH4 peptide was shown to decrease in pregnancies that resulted in preterm births, and had a sensitivity of 65.0% and specificity of 82.5% in discriminating preterm birth from term delivery, although the biological activity of the parent protein or the fragment identified is unknown.

Examination of the NCBI SNP database of common gene variations and population genetics data from 1000 genomes project (Genomes Project et al., 2010) revealed that there is a single nucleotide polymorphism (SNP, variant rs2276814) in *ITIH4* (position 2065) where a single coding nucleotide differs from A of amino acid codon cAa to T of cTa, resulting in an amino acid change from glutamine (Q) to leucine (L). As shown in Fig. S1 and Table S1, African American or Sub-Saharan African subjects have comparable probabilities of "A" or "T" allele, and therefore, similar chances of glutamine (Q) or leucine (L) at ITIH4 protein position 669. In contrast, European, Asian and Hispanic American subjects are predominantly homozygous for the "T" allele, and therefore, carry leucine at position 669 of ITIH4 ("L" isoform).

Within the European, Asian and Hispanic American populations, we postulated that the ITIH4 derived serum peptide should be in its "L" isoform (protein ITIH4 669, nucleotide level A→T, protein level Q→L). This expected "L" isoform peptide sequence should be LLGLPGPPDVPDHAAYHPF, which should share an almost identical sequence as the preterm birth serum peptide biomarker with "Q" isoform (QLGLPGPPDVPDHAAYHPF) but with the first amino acid changed from L to Q.

LCMS based quantitative proteomic analysis is a powerful method for selective quantification of specific proteins/peptides in very complex mixtures. LCMS method coupled with stable isotope dilution (SID) provides both absolute structural specificity for the analyte and relative or absolute measurement of analyte concentration (Addona et al., 2009). Unlike the label-free quantitative method, which is more error-prone due to systematic variations among individual runs and stochastic nature of the indices used for calculation (Kito and Ito, 2008), SID based peptide assay is the gold standard for absolute protein quantitation *via* mass spectrometry (Fig. S2). We postulated that targeted mass spectrometry (also known as data-dependent acquisition), if focusing on "Q" isoform's specific mass/charge ratio (m/z) and specific liquid chromatography time as reported previously (Esplin et al., 2011), would fail to read out the ITIH4 "L" isoform peptide.

We hypothesized that the ITIH4 "L" isoform peptide could be detected in the sera of Asian, European, or Hispanic backgrounds, and further that this "L" isoform serum peptide would be also a biomarker associated with preterm birth, similar to the reported "Q" isoform (Esplin et al., 2011). To test these hypotheses, we firstly applied quantitative proteomic methods to identification of the ITIH4 "L" isoform peptide in normal pregnancy sera. Fig. 1A shows a representative LCMS ion chromatogram from the term delivery pregnancy samples (ProMedDX, USA, http://www.promeddx.com). The ProMedDX sera were from women of uncomplicated pregnancies (age, 26.1 ± 6.86 years; 11 were of Hispanic origin and 3 African Americans) at various gestational ages. Each of the peaks in these chromatograms was formed by the elution of serum peptides. The "L" ITIH4 peptide, eluting at average 15.9 min (standard deviation, 0.23 min) from the 60-min HPLC column, was identified as doubly charged with m/z of 1006.26 and the sequence MS/MS spectral profile was found to match to the peptide sequence LLGLPGPPDVPDHAAYHPF (Fig. 1B).

Next, we conducted quantitative proteomic analysis to determine the association between the ITIH4 "L" isoform serum peptide and the preterm birth. The LCMS ion chromatogram and the matching MS/MS spectrum of ITIH4 "L" isoform in Stanford samples are consistent with those in the ProMedDX samples. The information of Stanford samples, cases (n = 11, preterm birth) and controls (n = 14, term delivery) was summarized in Table S2. Serum samples were collected at Stanford University Medical Center under IRB approved protocols. Our study cohort consisted of predominantly Asians, Europeans, or their mixture (case: 9.1% African-American, 27.3% Asian, 18.2% Caucasian,

and 45.5% Hispanic; control: 0% African American, 14.3% Asian, 7.1%

Caucasian, 71.4% Hispanic, and 7.1% Pacific islanders). The ProMedDX samples were used for "L" isoform identification, while the Stanford samples were used for the following quantitative study. As summarized in Fig. 1C, the peaks in the chromatogram were formed by the elution of "L" isoform peptide from C18 column (see Materials and Methods for details) at the HPLC 16th time point. The normalized "L" isoform peptide serum concentration in women with preterm birth was found to be 3-fold less when compared with women who had term deliveries. Scatter plot analysis (Fig. 1D) revealed that the abundance of ITIH4 "L" isoform decreased significantly (both Student's *t*-test and Mann-Whitney U test, $P < 0.001$) in preterm birth subjects relative to the controls. As summarized in Fig. 1E, we analyzed the normalized ITIH4 peptide serum abundances in the term delivery pregnancy sera (ProMedDX collection) of different gestational ages. The comparative analysis of ITIH4 "L" peptide serum abundance between early (<35 weeks) and late (≥35 weeks) gestational age groups revealed no statistically significant difference (Student's *t*-test, $P = 0.747$). The level of ITIH4 "L" isoform peptide does not change with gestational ages in term delivery pregnancy sera. These supported our hypotheses that 1) ITIH4 "L" isoform serum peptide can be detected in a pregnancy cohort predominantly of Asian, Hispanic and Caucasian origin; 2) reduction of the serum level of ITIH4 "L" isoform is associated with preterm birth in our cohort. The ROC AUC (area under the curve) as a measure of the classification performance is 90.2% (95% confidence interval (CI), 88.3%–91.0%). The cut point performance of positive predictive value (PPV) is 0.76 (95% CI, 0.65–0.84), with test sensitivity of 0.78 (95% CI, 0.64–0.88) and specificity of 0.80 (95% CI, 0.62–0.84).

The proteomic observation in the current study is congruent with the allele frequency of SNP records, which shows that most subjects of European, Asian and/or Hispanic American origins have "L" at ITIH4 669 position (Fig. S1 and Table S1). In addition, our observations of the lack of effect of gestational age on level of the "L" peptide in women with full term pregnancies indicate that the reduction of the ITIH4 "L" isoform abundance in women with preterm birth is indeed associated with preterm birth pathophysiology, but not gestational age.

Our findings of the serum ITIH4 "L" isoform peptide point out the importance of the ethnic variation when applying the ITIH4 serum peptide as a preterm birth biomarker. Population genetic analysis of the available SNP data indicates that most European, Asian and Hispanic American subjects have "L" at ITIH4 669. Targeted analysis of the serum ITIH4 "Q" isoform peptide in these race/ethnic subjects would be expected to reveal minimum detection. Therefore, the "Q" peptide based algorithm (Esplin et al., 2011) would misclassify these race/ethnic subjects in the high-risk group for preterm birth regardless of their physiological conditions. This could result in the wrong inference in a future trial to establish the clinical utility of ITIH4 serum peptide for prediction of gestational timing of delivery.

We believe that the association of the ITIH4 "L" isoform with preterm birth complements the previous findings of the "Q" isoform as a preterm birth biomarker. Specifically, in patients whose genotypes are heterozygous, quantifying either "L" or "Q" peptide alone and comparing with normal ITIH4 serum peptide level may lead to misdiagnosis. Consideration of both ITIH4 peptide isoforms can help to avoid false positives arising from the analysis of serum levels of a non-detectable ITIH4 peptide isoform target in the corresponding population with a particular race/ethnic background. To comprehensively apply the ITIH4 serum peptide biomarker for preterm birth analysis in all race/ethnic backgrounds, we proposed a multiple-stage procedure (Fig. S3). At the first stage, blood cells are processed and SNP genotyping is performed to determine whether the subject's ITIH4 669 protein position is in "Q" or "L" isoform. At the second stage, upon the genotyping results, the targeted analysis of ITIH4 "Q" (homozygous genotype), "L" (homozygous genotype), or the mixed two (heterozygous genotype) in the patient serum is conducted. Finally, the serum quantity of ITIH4 peptide isoform(s) would be used, in combination with other protein markers (Esplin et al., 2011), to gauge the patient's risk of preterm birth.

Our examination of the ITIH4 "L" isoform as a biomarker for preterm birth is carried out in a small cohort consisting of samples with different gestational age, which is a limitation of the study. Conducting a future prospective trial of our proposed multiple-stage procedure (Fig. S3) may lead to a clinically applicable test for preterm birth in patients of diverse race/ethnic backgrounds.

Due to the SNP (rs2276814) at position 2065 of *ITIH4*, there are "Q" or "L" peptide isoforms (Q/LGLPGPPDVPDHAAYHPF). Our targeted analysis showed that the reduction of the "L" peptide's serum quantity is associated with the clinical state of preterm birth ($P < 0.001$) in a cohort of uncomplicated pregnancies and pregnancies complicated by preterm birth subjects. We concluded that "L" peptide is a potential biomarker predictive of preterm birth in the cohort consisting predominantly of Europeans, Asians and Hispanic Americans.

ACKNOWLEDGMENTS

The authors thank Kenneth Lau for MALDI mass spectrometric analysis. The authors also thank scientists at the Pediatric Proteomics group and the March of Dimes

Prematurity Research Center at Stanford University for critical discussions. This work was supported in part by March of Dimes, and the Lucile Packard Foundation for Children's Heath. This work was also supported in part by the National Natural Science Foundation of China (NSFC) to ZT (No. 31201697)

SUPPLEMENTARY DATA

Supplementary data associated with this article can be found at http://dx.doi.org/...

Zhou Tan[a,b,1], *Zhongkai Hu*[b,1], *Emily Y. Cai*[b,1], *Cantas Alev*[e], *Ting Yang*[b], *Zhen Li*[b], *Joyce Sung*[c], *Yasser Yehia El-Sayed*[c], *Gary M. Shaw*[d], *David K. Stevenson*[d], *Atul J. Butte*[d], *Guojun Sheng*[e], *Karl G. Sylvester*[b], *Harvey J. Cohen*[d], *Xuefeng B. Ling*[b,*]

[a]*Institute of Developmental and Regenerative Biology, Hangzhou Normal University, Hangzhou 310029, China*
[b]*Department of Surgery, Stanford University, Stanford, CA 94305, US*
[c]*Department of Obstetrics and Gynecology, Stanford University, Stanford, CA 94305, US*
[d]*Department of Pediatrics, Stanford University, Stanford, CA 94305, US*
[e]*Lab for Early Embryogenesis*[2]*, RIKEN Center for Developmental Biology, Chuo-Ku, Kobe, Hyogo 650-0047, Japan*

*Corresponding author. Tel: +1 650 427 9198, fax: +1 650 723 1154
E-mail address: bxling@stanford.edu (X. B. Ling)
[1]These authors contributed equally to this work.

REFERENCES

Addona, T.A., Abbatiello, S.E., Schilling, B., Skates, S.J., Mani, D.R., Bunk, D.M., Spiegelman, C.H., Zimmerman, L.J., Ham, A.J., Keshishian, H., Hall, S.C., Allen, S., Blackman, R.K., Borchers, C.H., Buck, C., Cardasis, H.L., Cusack, M.P., Dodder, N.G., Gibson, B.W., Held, J.M., Hiltke, T., Jackson, A., Johansen, E.B., Kinsinger, C.R., Li, J., Mesri, M., Neubert, T.A., Niles, R.K., Pulsipher, T.C., Ransohoff, D., Rodriguez, H., Rudnick, P.A., Smith, D., Tabb, D.L., Tegeler, T.J., Variyath, A.M., Vega-Montoto, L.J., Wahlander, A., Waldemarson, S., Wang, M., Whiteaker, J.R., Zhao, L., Anderson, N.L., Fisher, S.J., Liebler, D.C., Paulovich, A.G., Regnier, F.E., Tempst, P., Carr, S.A., 2009. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. Nat. Biotechnol. 27, 633-641.

Esplin, M.S., Merrell, K., Goldenberg, R., Lai, Y., Iams, J.D., Mercer, B., Spong, C.Y., Miodovnik, M., Simhan, H.N., van Dorsten, P., Dombrowski, M., Eunice Kennedy Shriver National Institute of Child, H., Human Development Maternal-Fetal Medicine Units, N., 2011. Proteomic identification of serum peptides predicting subsequent spontaneous preterm birth. Am. J. Obstet. Gynecol. 204, 391 e391-398.

Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

Goldenberg, R.L., Rouse, D.J., 1998. Prevention of premature birth. N. Engl. J. Med. 339, 313-320.

Kito, K., Ito, T., 2008. Mass spectrometry-based approaches toward absolute quantitative proteomics. Curr. Genomics 9, 263-274.

Ling, X.B., Lau, K., Deshpande, C., Park, J.L., Milojevic, D., Macaubas, C., Xiao, C., Lopez-Avila, V., Kanegaye, J., Burns, J.C., Cohen, H., Schilling, J., Mellins, E.D., 2010a. Urine peptidomic and targeted plasma protein analyses in the diagnosis and monitoring of systemic juvenile idiopathic arthritis. Clin. Proteomics 6, 175-193.

Ling, X.B., Lau, K., Kanegaye, J.T., Pan, Z., Peng, S., Ji, J., Liu, G., Sato, Y., Yu, T.T., Whitin, J.C., Schilling, J., Burns, J.C., Cohen, H.J., 2011. A diagnostic algorithm combining clinical and molecular data distinguishes Kawasaki disease from other febrile illnesses. BMC Med. 9, 130.

Ling, X.B., Mellins, E.D., Sylvester, K.G., Cohen, H.J., 2010b. Urine peptidomics for clinical biomarker discovery. Adv. Clin. Chem. 51, 181-213.

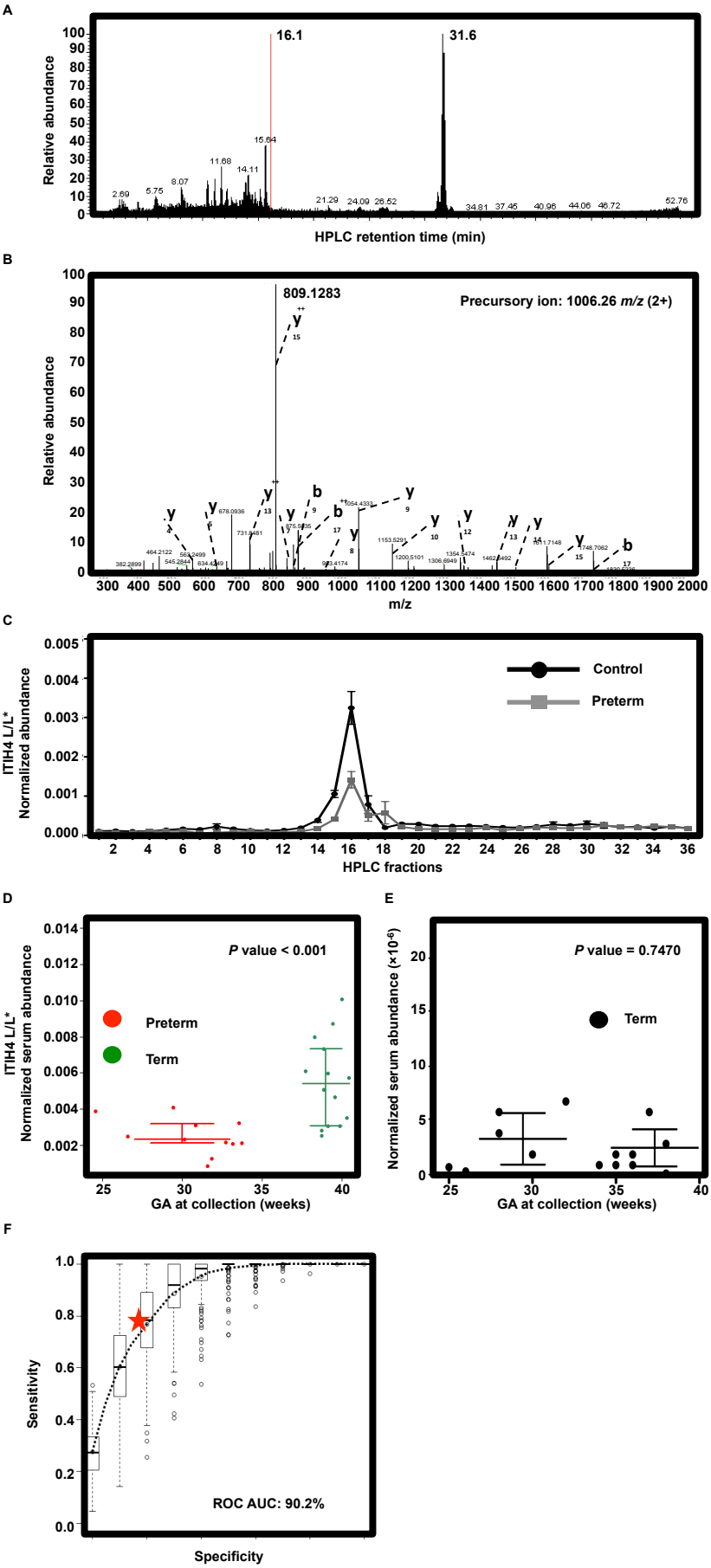
Ling, X.B., Sigdel, T.K., Lau, K., Ying, L., Lau, I., Schilling, J., Sarwal, M.M., 2010c. Integrative urinary peptidomics in renal transplantation identifies biomarkers for acute rejection. J. Am. Soc. Nephrol. 21, 646-653.

Liotta, L.A., Petricoin, E.F., 2006. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. J. Clin. Invest. 116, 26-30.

Villanueva, J., Shaffer, D.R., Philip, J., Chaparro, C.A., Erdjument-Bromage, H., Olshen, A.B., Fleisher, M., Lilja, H., Brogi, E., Boyd, J., Sanchez-Carbayo, M., Holland, E.C., Cordon-Cardo, C., Scher, H.I., Tempst, P., 2006. Differential exoprotease activities confer tumor-specific serum peptidome patterns. J. Clin. Invest. 116, 271-284.

Figure Legend

Fig. 1. The ITIH4"L" isoform serum peptide is a potential biomarker associated with preterm birth.
A: Representative LCMS ion chromatogram. The red vertical line indicates that the ITIH4 "L" peptide precursor ion with the matching MS/MS spectrum was found eluting at 16.1 min. **B:** An MS/MS spectrum of precursory ion m/z 1006.26 (2+) that matched to the ITIH4 "L" peptide with sequence LLGLPGPPDVPDHAAYHPF. **C:** ITIH4 "L" isoform peptide chromatogram derived from preterm birth cases (gray) and healthy pregnancy controls (black). The peaks in the chromatogram were formed by the elution of "L" isoform peptides at the HPLC 16th time point, and the ionic intensities of ITIH4 "L" isoform peptide were normalized with the stable isotope labeled spiked-in ITIH4 "L*" isoform peptide. **D:** Scatter plot analysis of each subject's normalized serum abundance as a function of the baby gestational age at the time of sample collection. Violet red represents preterm birth cases; sea green represents full term pregnancy controls. **E:** ITIH4 "L" isoform peptide abundance in term delivery pregnancy sera of different gestational ages. Comparative analysis of ITIH4 "L" peptide serum abundances between early (< 35 weeks) and late (≥ 35 weeks) GA revealed no statistically significant difference between the two groups (Student's *t* test, $P = 0.747$). **F:** ROC analysis of the ITIH4 "L" isoform serum peptide as a preterm birth biomarker. The plotted ROC curve is the vertical average of the 500 bootstrapping runs, and the box and whisker plots show the vertical spread around the average. The red star denotes the cut point with the optimal sensitivity and specificity of the assay.


A
Relative abundance
HPLC retention time (min)
16.1
31.6
B
Relative abundance
m/z
809.1283
Precursory ion: 1006.26 m/z (2+)
C
ITIH4 L/L* Normalized abundance
HPLC fractions
Control
Preterm
D
ITIH4 L/L* Normalized serum abundance
GA at collection (weeks)
P value < 0.001
Preterm
Term
E
Normalized serum abundance (×10⁻⁶)
GA at collection (weeks)
P value = 0.7470
Term
F
Sensitivity
Specificity
ROC AUC: 90.2%

Targeted Analysis of an ITIH4 Peptide Isoform: A Serum Preterm Birth Biomarker and Population SNP Implications

Zhou Tan, Zhongkai Hu, Emily Y. Cai, Cantas Alev, Ting Yang, Zhen Li, Joyce Sung, Yasser Yehia El-Sayed, Gary M. Shaw, David K. Stevenson, Atul J. Butte, Guojun Sheng, Karl G. Sylvester, Harvey J. Cohen, Xuefeng B. Ling

Supplementary Data

Materials and Methods

Specimen collection and ethics

To identify the ITIH4 peptide sequences in "normal" term pregnancies, we procured 14 serum specimens from ProMedDX Inc. (Norton, MA, USA, http://www.promeddx.com). The ProMedDX sera were from women of uncomplicated pregnancies (mean of age, 26.1; age standard deviation, 6.86; 11 were of Hispanic origin and 3 African Americans) at various gestational ages. We confirmed that all of the ProMedDX specimens were collected under Institutional Review Board (IRB) approved protocols by qualified investigator sites. These sites conducted ProMedDX studies according to 21 CFR, ICH/GCP guidelines and HIPAA Privacy Regulations. Informed consent was obtained from every subject, unless this requirement had been determined by the IRB not to apply and had been waived.

To determine the association between the ITIH4 "L" isoform serum peptide and the preterm labor, case (n = 11, preterm birth) and control (n = 14, term delivery) cohorts were constructed to match age, ethnicity, and whether mothers were in labor at the time of collection (Table S2). All serum samples were collected at Stanford University Medical Center under IRB approved protocols. Written informed consent was obtained from each Stanford collection subject. Our study cohort consisted of predominantly Asians, Europeans, or their Mixed (case: 9.1% African American, 27.3% Asian, 18.2% Caucasian, and 45.5% Hispanic; control: 0% African American, 14.3% Asian, 7.1% Caucasian, 71.4% Hispanic, and 7.1% Pacific islanders). There were expected statistical differences (P < 0.001) between cases and controls in gestational age at delivery and at collection, time gap between collection and delivery, and baby birth weight.

LC-MS/MS analysis to identify the ITIH4 "L" isoform in term delivery sera

Serum peptides from the ProMedDX collection were prepared as previously described (Ling et al., 2010). Lyophilized human serum peptide samples were reconstituted in 2% acetonitrile with 0.1% formic acid and separated on a Paradigm MS4 liquid chromatography system (Michrom BioResources, Auburn, CA, USA) with a 60 min linear gradient of 5%–95% buffer A to B (buffer A: 2% acetonitrile with 0.1% formic acid in H_2O, buffer B: 90% acetonitrile with 0.1% formic acid in H_2O) at a flow rate of 2 μL/min using a 0.2×50 mm 3μ 200Å Magic C18AQ column (Michrom BioResources). Each randomized sample run was followed by a 60 min wash run. The fractionated peptides were directly applied to an LTQ ion trap mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) equipped with a Fortis

tip mounted nano-electrospray ion source (AMR, Tokyo, Japan). The electrospray voltage was set at 1.8kV. Each full MS scan with a mass range of 400–2000 *m/z* was followed by two data-dependent scans of the two most abundant ions observed in the first full MS scan. MS/MS spectra were generated for the highest peak in each scan with the relative collision energy for MS/MS set to 35%. Raw MS/MS data were preprocessed, as previously described (Griffin et al., 2010), before further statistical analysis. Peptide protein identification was searched against the human SwissProt database as previously described. At first, the intensity values of the same peptides in the same proteins were summed up across different fractions for each sample. Therefore, each peptide in one sample has one intensity value, which was later normalized by the total intensity value of all peptides found in the sample.

Quantitative proteomics based on stable isotope dilution

The 25 μL aliquot of patient serum from the Stanford collection was mixed with 75 μL of 100% methanol, and was vortex mixed for about 30 min. After centrifugation at 3000 rpm (~1700 *g*) for 10 min, supernatants were transferred into a 96 well plate. The extracted samples' peptide concentrations were quantified by 2,4,6-Trinitrobenzene Sulfonic Acid (TNBSA or TNBS) kit (TS-28997, Thermo Fisher, CA, USA).

For absolute quantification method using stable isotope labeled synthetic marker analogues, we chose stable isotope labeled (with five ^{13}C-labeled and one ^{15}N-labeled for each proline) "L" isoform ITIH4 peptide synthesized by AnaSpec Inc. (USA). Therefore, the synthetic labeled peptide had a total mass difference of 30 atomic mass units from the endogenous serum peptide. As shown in Fig. S2, stable isotope labeled peptide was added as a quantification reference in defined amounts to the serum peptide samples prior to the liquid chromatography (C18 column) mass spectrometric analysis using ABI5800 matrix-assisted laser desorption/ionization (MALDI) TOF (Time of Flight). Each sample's endogenous ITIH4 "L" isoform peptide abundance was normalized to the isotope reference peptide for subsequent statistical analysis.

Statistical analyses

Patient demographic data was analyzed using the "Epidemiological calculator" (R epicalc package). Student's *t* test was performed to calculate *P* values for continuous variables, and Fisher's exact test was used for comparative analysis of categorical variables. Testing was performed using Student's *t* test (two tailed) and Mann-Whitney U test (two tailed). Given the limited sample size in this study, bootstrapping method, in conjunction with the ROC curve analysis (Zweig and Campbell, 1993; Sing et al., 2005), was used to estimate the diagnostic performance of the ITIH4 "L" isoform peptide. The cut point along the ROC curve was determined as previously described (Zweig and Campbell, 1993) to obtain the optimal sensitivity and specificity of the assay.

REFERENCES

Griffin, N.M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J.A., Schnitzer, J.E., 2010. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat. Biotechnol. 28, 83-89.

Ling, X.B., Mellins, E.D., Sylvester, K.G., Cohen, H.J., 2010. Urine peptidomics for clinical biomarker discovery. Adv. Clin. Chem. 51, 181-213.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940-3941.

Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561-577.

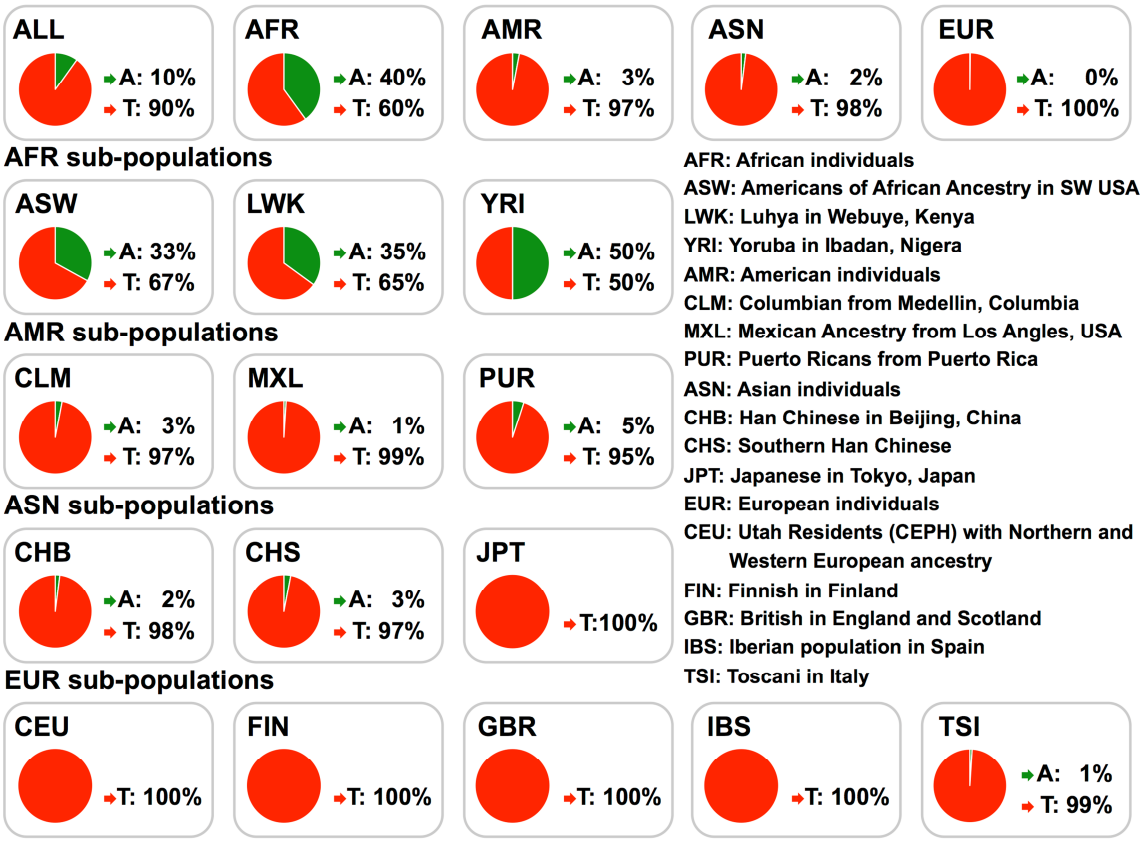
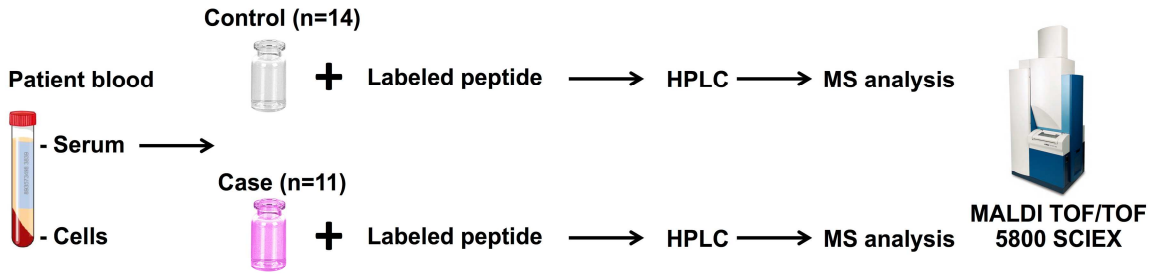


Fig. S1. Population genetics data analysis.
The SNP (rs2276814) allele frequencies of ITIH4 “L” and “Q” isoforms were extracted from the 1000 genomes project data. African American or Sub-Saharan African subjects have comparable probabilities of “A” or “T” allele at ITIH4 SNP position (Chromosome 3:52819464). In contrast, European, Asian and Hispanic American subjects predominantly are homozygous for the “T” allele.



$$\text{Normalized ITIH4 “L” isoform peptide quantity} = \frac{\text{Endogenous peptide MS signal}}{\text{Labeled peptide MS signal}}$$

Fig. S2. Experimental design diagram of the LCMS based targeted serum peptide analysis using stable isotope dilution (SID) method.

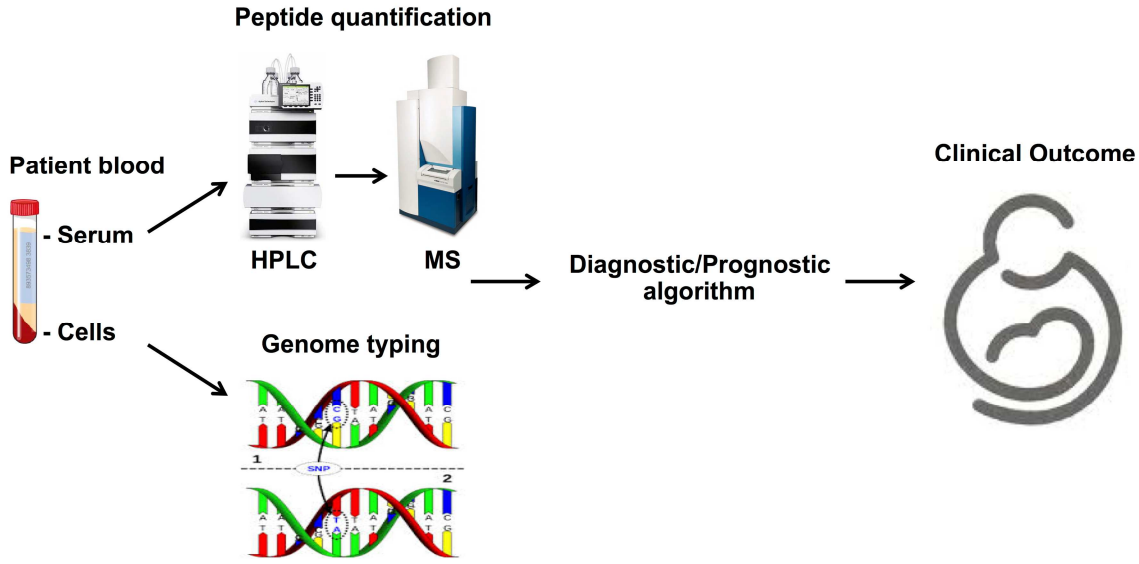


Fig. S3. A proposed multiple-stage procedure for the ITIH4 serum peptide biomarker analysis.
At the first stage, the blood cell genomic DNA will be extracted for genotyping of the ITIH4 669 allele. At the second stage, either the ITIH4 "L", or "Q", or both isoforms will be quantified by mass spectrometric based method. Finally, the serum quantity of ITIH4 peptide isoform(s) would be used, in combination with other protein markers, to determine the preterm birth risk of the assayed subject.

Table S1. Population diversity

	Sample ascertainment				Genotype Detail				Allele	
ss#	Population ID	Individual group	Chrom. sample cont.	Source	A	A/A	A/T	T/T	HWP	A	T
ss117081502	YRI		2	IG		1				1	
ss163407181	YRI	Sub-Saharan African	2	IG		1				1	
ss202512419	BANTU		1	IG	1					1	
ss220133049	pilot_1_YRI_low_coverage_panel		118	AF						0.475	0.525
ss23310746	AFD_EUR_PANEL	European	48	IG				1			1
	AFD_AFR_PANEL	African American	6	IG		0.087	0.478	0.435	0.752	0.326	0.674
	AFD_CHN_PANEL	Asian	48	IG				1			1
ss3213446	JBIC-allele		1502	AF						0.003	0.997
ss342135152	ESPN_Cohort_Populations		4550	GF		0.044	0.188	0.767	0.001	0.138	0.862
ss44404965	HapMap-CEU	European	116	IG				1			1
	HapMap-HCB	Asian	90	IG		0.022		0.978	0.001	0.022	0.978
	HapMap-JPT	Asian	88	IG				1			1
	HapMap-YRI	Sub-Saharan African	120	IG		0.167	0.617	0.217	0.1	0.475	0.525
	ENSEMBL_Watson		2	IG				1			1
	ENSEMBL_Venter		2	IG				1			1
ss491835989	CSAgilent		1225	GF		0.002	0.018	0.98		0.011	0.989
ss68861109	HapMap-CEU	European	120	IG				1			1
	HapMap-HCB	Asian	90	IG			0.022	0.978	1	0.011	0.989
	HapMap-JPT	Asian	90	IG				1			1
	HapMap-YRI	Sub-Saharan African	120	IG		0.167	0.617	0.217	0.1	0.475	0.525

Data are adapted from dbSNP website http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?type=rs&rs=2276814. ss#, submitted SNP number; Chrom. sample cont., ascertainment sample size; Source, population diversity source type; IG, individual genotype; AF, allele frequency; GF, genotype frequency; HWP, Hardy-Wenburger probability.

Table S2. Summary of patient characteristics of Stanford samples

Characteristic	Case (n = 11)	Control (n = 14)	P value Case *vs.* Control
Age (year)			0.811
Mean	29.1	28.5	
SD	4.9	6.8	
Race			0.451
African-American	1 (9.1%)	0 (0%)	
Asian	3 (27.3%)	2 (14.3%)	
Caucasian	2 (18.2%)	1 (7.1%)	
Hispanic	5 (45.5%)	10 (71.4%)	
Pacific Islander	0 (0%)	1 (7.1%)	
GA* at delivery (week)			< 0.001
Mean	32.74	39.3	
SD	3.32	0.94	
GA at collection (week)			< 0.001
Mean	30.74	39.21	
SD	2.93	0.77	
Gap between collection and delivery (week)			< 0.001
Mean	2	0.08	
SD	2.58	0.5	
Birth weight (g)			< 0.001
Mean	2069.9	3533.9	
SD	663.7	306.1	
Labor			0.288
Yes	8 (72.7%)	13(92.9%)	
No	3 (27.3%)	1(7.1%)	

*GA, gestational age.