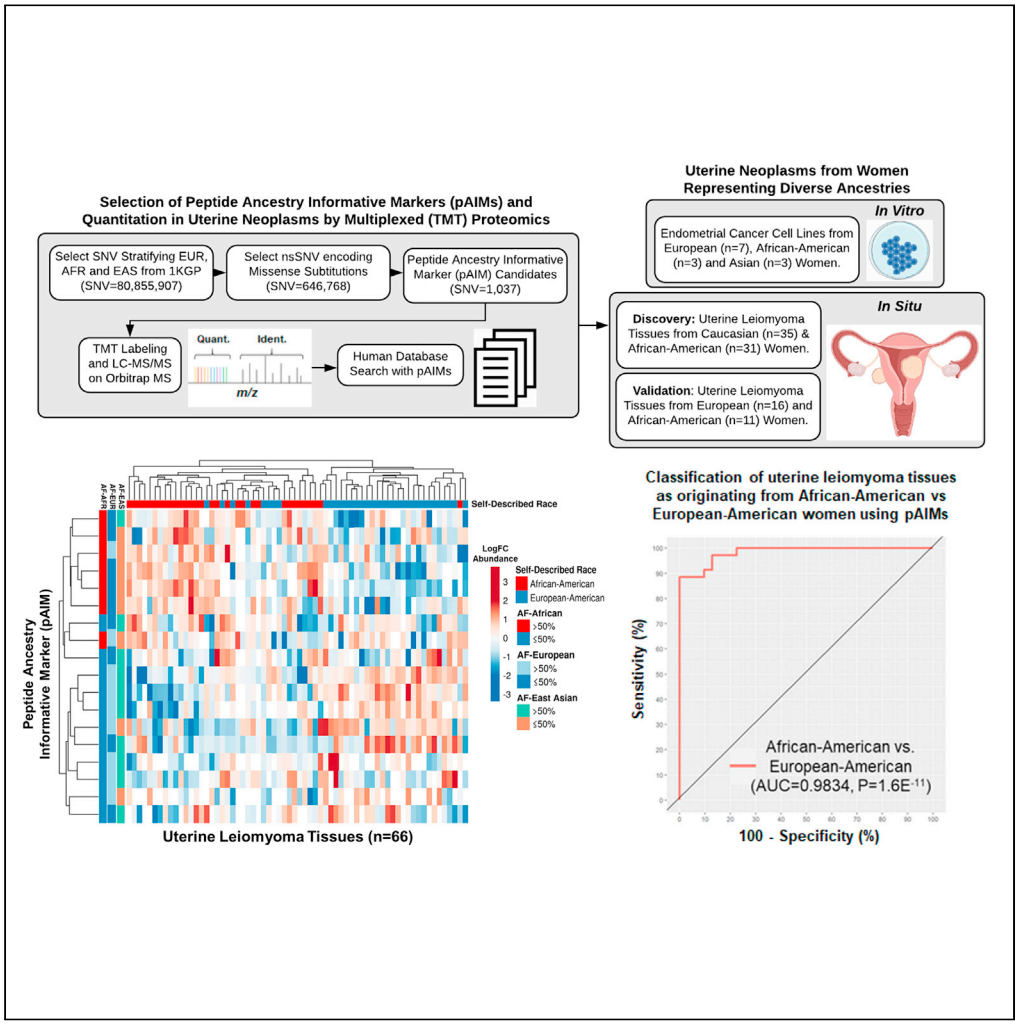
iScience

CellPress
OPEN ACCESS

Article

Peptide ancestry informative markers in uterine neoplasms from women of European, African, and Asian ancestry



Nicholas W. Bateman, Christopher M. Tarney, Tamara S. Abulez, ..., G. Larry Maxwell, Thomas P. Conrads, Timothy D. O'Connor

batemann@whirc.org

Highlights

pAIMs encode substitutions from European, African, and East Asian populations

In silico analysis shows ~20 pAIMs can determine ancestry similarly as >260K SNPs

pAIMs can estimate population-level ancestry in proteomic data from human tissues

Bateman et al., iScience *25*, 103665
January 21, 2022
https://doi.org/10.1016/j.isci.2021.103665

iScience

CellPress
OPEN ACCESS

Article

Peptide ancestry informative markers in uterine neoplasms from women of European, African, and Asian ancestry

Nicholas W. Bateman,[1,2,3,16,17,*] Christopher M. Tarney,[1] Tamara S. Abulez,[1,3] Brian L. Hood,[1,3] Kelly A. Conrads,[1,3] Ming Zhou,[4] Anthony R. Soltis,[3,5] Pang-Ning Teng,[1,3] Amanda Jackson,[1] Chunqiao Tian,[1,3] Clifton L. Dalgard,[5,6] Matthew D. Wilkerson,[2,3,5,6] Michael D. Kessler,[7] Zachary Goecker,[8] Jeremy Loffredo,[1] Craig D. Shriver,[2] Hai Hu,[2,9] Michele Cote,[10] Glendon J. Parker,[8] James Segars,[11] Ayman Al-Hendy,[12] John I. Risinger,[13] Neil T. Phippen,[1,2] Yovanni Casablanca,[1,2] Kathleen M. Darcy,[1,2,3] G. Larry Maxwell,[1,2,4] Thomas P. Conrads,[1,2,4] and Timothy D. O'Connor[7,14,15,16]

SUMMARY

Characterization of ancestry-linked peptide variants in disease-relevant patient tissues represents a foundational step to connect patient ancestry with disease pathogenesis. Nonsynonymous single-nucleotide polymorphisms encoding missense substitutions within tryptic peptides exhibiting high allele frequencies in European, African, and East Asian populations, termed peptide ancestry informative markers (pAIMs), were prioritized from 1000 genomes. *In silico* analysis identified that as few as 20 pAIMs can determine ancestry proportions similarly to >260K SNPs (R^2 = 0.99). Multiplexed proteomic analysis of >100 human endometrial cancer cell lines and uterine leiomyoma tissues combined resulted in the quantitation of 62 pAIMs that correlate with patient race and genotype-confirmed ancestry. Candidates include a D451E substitution in GC vitamin D-binding protein previously associated with altered vitamin D levels in African and European populations. pAIMs will support generalized proteoancestry assessment as well as efforts investigating the impact of ancestry on the human proteome and how this relates to the pathogenesis of uterine neoplasms.

INTRODUCTION

Proteogenomics aims to integrate protein-level measurements with companion transcriptome and genome sequencing data (Gupta et al., 2007) to elucidate complex systems-level relationships between protein and transcript-level expression including identification of proteins encoding missense substitutions that may impact protein function (Ng and Henikoff, 2003; Gupta et al., 2007). Understanding the impact of proteogenomic alterations on disease risk and patient outcome is a priority for population-level investigations (Suhre et al., 2020), particularly those focusing on cancer, such as efforts by The Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Mertins et al., 2016), and the recently initiated Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) research network (Fiore et al., 2017). Many historic efforts have poor representation of US minorities among patient cohorts analyzed (Spratt et al., 2016), and there remains a paucity of systems-level molecular analyses within powered cohorts to elucidate race-related differences in disease etiology. In the context of gynecologic disease, patients with uterine leiomyomas (fibroids) and endometrial cancers experience significant racial disparities in disease incidence and outcome (Deshmukh et al., 2017; Tarney et al., 2018; Felix et al., 2011; Allard and Maxwell, 2009; Farley et al., 2007; Maxwell et al., 2006; Mahdi et al., 2014). Among possible molecular mechanisms that may give rise to racial disparities in gynecologic disease, one such manifestation could be from nonsynonymous single-nucleotide polymorphisms (nsSNPs) encoding missense substitutions that may alter protein function(s). Defining ancestry-linked peptide variants in disease-relevant cells and tissues will augment proteogenomic discovery efforts that aim to define molecular drivers underlying racial disparities in gynecologic diseases.

[1]Gynecologic Cancer Center of Excellence, Department of Gynecologic Surgery and Obstetrics, Uniformed Services University and Walter Reed National Military Medical Center, 8901 Wisconsin Avenue, Bethesda, MD 20889, USA

[2]The John P. Murtha Cancer Center, Uniformed Services University and Walter Reed National Military Medical Center, 8901 Wisconsin Avenue, Bethesda, MD 20889, USA

[3]Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., 6720A Rockledge Dr., Suite 100, Bethesda, MD 20817, USA

[4]Department of Obstetrics and Gynecology, Inova Fairfax Medical Campus, 3300 Gallows Road, Falls Church, VA 22042, USA

[5]The American Genome Center; Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA

[6]Department of Anatomy Physiology and Genetics, Uniformed Services University, 4301 Jones Bridge Road, Bethesda, MD 20814, USA

[7]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[8]University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

[9]Chan Soon-Shiong Institute of Molecular Medicine at

Continued

Ancestry informative markers (AIMs) comprise SNPs that are highly differentiated within discrete ancestral populations (Enoch et al., 2006; Kosoy et al., 2009; Galanter et al., 2012). Most of these markers occur within non-protein coding regions, but with large-scale sequencing now available for a diversity of ancestries (Mathias et al., 2016; Kessler et al., 2016), especially in protein coding regions (Tennessen et al., 2011, 2012), it is now possible to assess AIMs that are detectable through proteomics data alone. To this end, efforts to date have identified genetically variant peptides differentiating individuals of European and African descent (Parker et al., 2016) as well as Asian populations (Feng Lei et al., 2019), and these have been explored for forensic human identification, such as through proteomic analyses of human hair and bone samples. We leveraged population-level genotyping data from the 1000 Genomes Project (Genomes Project et al., 2015), an effort focused on cataloging human genetic variation around the world (https://www.internationalgenome.org/), to identify SNPs encoding missense substitutions that exhibit altered allele frequencies with European, African, and East Asian ancestries and prioritized candidates occurring within tryptic peptides observable by routine, bottom-up proteomic workflows to explore these within disease-relevant models and tissues from uterine neoplasms. Although these candidates are not tissue dependent and represent a generalized set of markers for proteoancestry assessment in diverse tissues and cell lines, we show here the ability to ancestrally characterize endometrial cancer cell line models as well as uterine fibroid tissues from women representing diverse racial and ancestral backgrounds using multiplexed proteomic approaches. Our study further defines a foundational set of ancestry-linked variant peptides within uterine tissues that will support ongoing efforts to investigate germline as well as somatic proteogenomic alterations underlying ancestry-linked disease biology and how this may further relate to racial disparities in the pathogenesis of uterine neoplasms.

RESULTS

Selection and *in silico* analyses of peptide ancestry informative markers

We selected >640K SNPs from the 1000 Genomes Project (Genomes Project et al., 2015) encoding nonsynonymous substitutions and further filtered them to those exhibiting ≥50% allele frequency differences between individuals of European (i.e., FIN, GBR, IBS, and TSI; N = 404), African (i.e., ESN, GWD, LWK, MSL, and YRI; N = 504), and East Asian (i.e., CDX, CHB, CHS, and KHV; N = 400) descent. Candidates were also filtered to prioritize missense substitutions occurring in tryptic peptides (≥6 and ≤40 amino acids in length) that would be observable within "bottom-up," multiplexed analyses of disease-relevant cell lines and tissues collected from individuals representing diverse ancestries (Workflow Figure 1A, Table S1A). This filtering approach resulted in 1,037 peptide ancestry informative markers (pAIMs) mapping to 831 unique proteins spanning all 22 autosomes and exhibiting the greatest frequencies on Chr1 (12.3%) and Chr19 (7.5%) (Table S1A). Most of these parent proteins encode a single pAIM (Table S1A) with a subset of 136 candidates encoding ≥2 pAIMs as well as 3 candidates encoding >5 pAIMs: filaggrin (10 pAIMs), cardiomyopathy-associated protein 5 (7 pAIMs), and HLA class II histocompatibility antigen, DP alpha 1 chain precursor (6 pAIMs). Functional enrichment analyses of proteins encoding pAIMs revealed significant enrichment of cellular pathways regulating the cellular matrisome as well as keratinocyte differentiation and sensory perception (Table S1B). Unique pAIMs exhibiting ≥50% allele frequencies were most prevalent in African populations (344 total sites) followed by East Asian (229) and European (79) ancestries, and East Asian as well as European populations exhibited nearly 2-fold greater conservation (208) of shared pAIMs relative to African populations (92 and 85, respectively) (Figure 1B). Fixation index (Fst) values for nonsynonymous single-nucleotide polymorphisms (SNPs) encoding pAIMs exhibited an average of 0.312 ± 0.1 consistent with Fst threshold expectations for identity-informative SNPs used in forensic DNA sequencing (King et al., 2018) (Table S1A). The clinical significance and known disease pathogenicity of SNPs encoding pAIMs within the ClinVar resource was assessed for 1,031 subset pAIMs mapping to the Ensemble Variant Effect Predictor tool (McLaren et al., 2016). Most of these pAIMs had no previous associations with clinical significance. We did, however, identify 87 pAIMs of unknown clinical significance that map to putative functional protein domains (Figure 2A, Table S1A). The remaining candidates were largely undocumented for disease association in ClinVar with 7 pAIMs candidates correlating with altered drug responses or being risk factors for disease. This latter subset included a substitution in kinesin family member 6 that is associated with altered response to statin treatment (Ruiz-Iruela et al., 2018) (W719R, rs20455) and has higher allele frequencies within African (86%) and East Asian (53%) populations, or as being risk factors for several diseases (Table S1A). Risk factors included a substitution in aurora kinase A (F31I, rs2273535), which has higher allele frequencies within East Asians (67%) and is associated with an increased risk of developing colon cancer within this population (Xu et al., 2014). We also assessed the impact of pAIM substitutions on protein function via *in silico* prediction analyses using SIFT (Sim et al., 2012), PolyPhen-2

Windber, Windber, PA 15963, USA

[10]Wayne State University, Detroit, MI 48202, USA

[11]Johns Hopkins University Medical Center, Baltimore, MD 21218, USA

[12]The University of Illinois College of Medicine, Chicago, IL 60612, USA

[13]Department of Obstetrics and Gynecology, Michigan State University, East Lansing, MI 48824, USA

[14]Program in Personalize and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[15]Marlene and Stewart Greenebaum Comprehensive Cancer, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[16]These authors contributed equally

[17]Lead contact

*Correspondence: 3289 Woodburn Rd, Suite 375, Annandale, VA 22003; batemann@whirc.org

https://doi.org/10.1016/j.isci.2021.103665

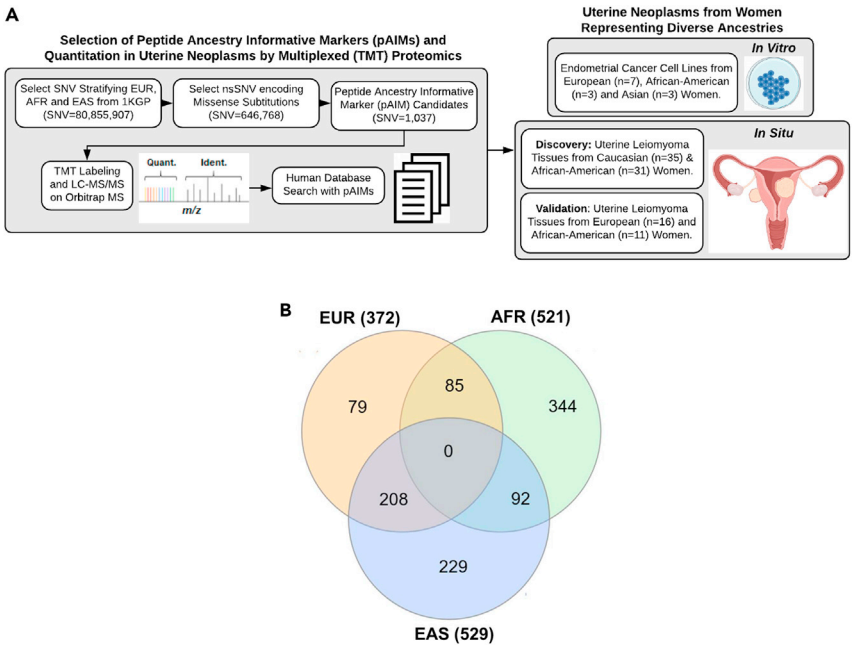


Figure 1. Selection of Peptide Ancestry Informative Markers (pAIMs) and Workflow for Analyses in Cell Lines Models and Tissues of Uterine Neoplasms collected from Women of Diverse Ancestries

(A) Greater than 640K single-nucleotide polymorphisms (SNPs) were filtered to prioritize 1,037 missense substitutions occurring within lysine or arginine-terminating tryptic peptides, ≥ 6 and ≤40 amino acids in length, so-called peptide ancestry informative markers (pAIMs). We then investigated pAIMs in endometrial cancer cell line models as well as uterine leiomyoma tissues from women representing diverse ancestries.

(B) Comparison of pAIMs exhibiting ≥50% allele frequencies within individuals of European (EUR), African (AFR), or East Asian (EAS) ancestry.

(Adzhubei et al., 2013), and PROVEAN (Choi et al., 2012) and identified 23 pAIMs that may be deleterious (Figure 2B). Although many of these candidates have unknown clinical significance, five candidates have been shown to not be pathogenic for disease, i.e., categorized by ClinVar as benign or likely benign (Figure 2C, Table S1A). These candidates include a substitution in ATP-binding cassette transporter sub-family C member 11 (G180R, rs17822931) that has high allele frequencies within East Asian populations (75%) and is significantly correlated with Axillary Osmidrosis in Chinese Han populations (Ren et al., 2017) as well as increased risk for breast cancer in Japanese women (Ishiguro et al., 2019). Further analyses of these 23 putatively deleterious pAIMs revealed they exhibit higher allele frequencies in East Asian versus European and African populations (Figure 2D). We also utilized as a test set of populations CEU, a European population from Utah (EUR); ASW, an admixed African American population from the South West (AA); ACB, an admixed African Caribbean population (AFR); and JPT, a Japanese population from East Asia (EAS) to assess the accuracy of pAIMs to classify ancestral proportions relative to standard genotype estimates *in silico*. We apply a random sampling analysis of nsSNPs encoding pAIMs and compared them to using 266,403 LD-pruned SNPs by standard estimates approaches. We find that as few as 20 variants can recapitulate genome-wide ancestry proportions, using standard approaches (Alexander and Lange, 2011; Alexander et al., 2009; Kessler et al., 2016) (error is increased due to smaller feature set, but the average of 500 random sampling assessments remains robust, $R^2 > 0.99$) (Figure 3).

Quantitation of peptide ancestry informative markers in endometrial cancer cell lines

We determined the ancestry of thirteen endometrial cancer cell line models by standard estimates using comparison of whole-genome sequencing (WGS) data with reference populations (Table 1). Most EC models were established from women of European descent (n = 7), with a subset corresponding to women of African (n = 3) and East Asian (n = 3) descent. We performed quantitative, bottom-up proteomic analyses employing a multiplexed tandem-mass tag (TMT) approach and quantified 133,473 total peptides that included 43 high-confidence pAIM variant peptides (Table S2), where MS2 ions flanking the pAIM

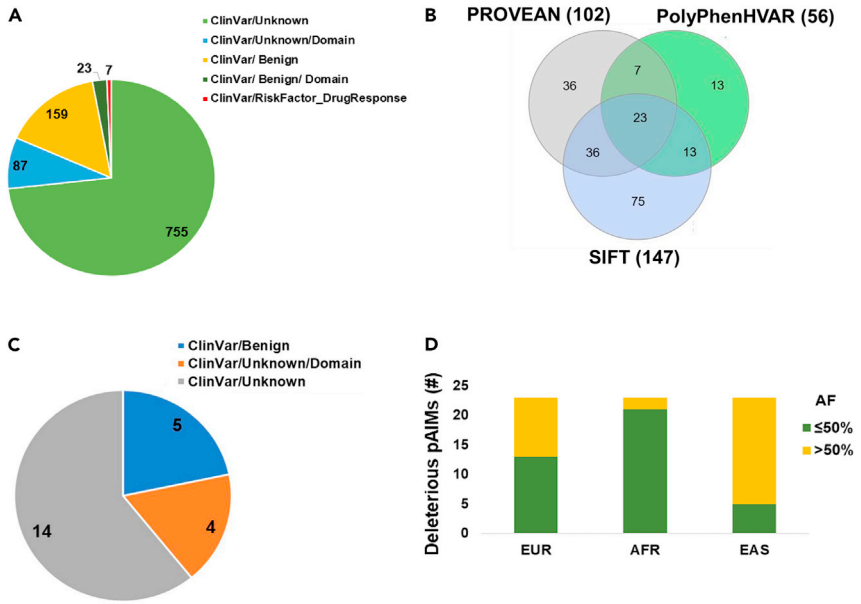


Figure 2. *In silico* analyses of Peptide Ancestry Informative Markers (pAIMs) to assess Clinical Significance and Localization to Protein Domains as well as to Predict Impact on Protein Function

(A) Clinical significance (ClinVar) and protein domain (Uniprot) localization of 1,031 pAIMs.

(B) Comparison of pAIMs predicted to be deleterious to protein function by PROVEAN, PolyPhen, and SIFT functional prediction tools.

(C) Clinical significance (ClinVar) and protein domain (Uniprot) localization of pAIMs predicted to be deleterious to protein function.

(D) Allele frequency of pAIMs predicted to be deleterious to protein function in European (EUR), African (AFR), or East Asian (EAS) populations.

substitution of interest were confirmed using SpectrumAI (Zhu et al., 2018). Of these pAIM candidates, 30 are of unknown clinical significance, 12 are undocumented for disease association in ClinVar, and 1 candidate (rs8010699) is of uncertain significance for disease. This latter variant encodes a substitution in nesprin-2 (H3309R) that is associated with Emery-Dreifuss muscular dystrophy and has further been correlated with expression of the tumor suppressor cyclin-dependent kinase inhibitor 1 (CDK1/p21) in HBV-related hepatocellular carcinomas (Han et al., 2016). Investigation of the allele frequencies of pAIM candidates identified in endometrial cancer cell line models showed most candidates encoded missense substitutions from nsSNPs exhibiting higher allele frequencies (>50%) in both European and East Asian populations (24), with subsets uniquely exhibiting higher allele frequencies in African (11), East Asian (7), or European populations (1) (Figure 4A). As most ancestry-specific pAIMs quantified corresponded to African or East Asian populations, we estimated the ancestry of cell lines confirmed to be of African or East Asian ancestry using standard genotype estimates (Figures 4B, S1A, and S1B). Our analysis shows that the predominant ancestry proportion measured for each cell line based on abundance of pAIMs candidates reflects the ancestry confirmed by standard genotype estimates. We further investigated the abundance of 17 pAIMs that exhibited >50% allele frequency within European ancestry populations and found that these candidates were significantly elevated in cell lines from individuals of European versus African ancestry (Mann-Whitney U [MWU] Median = +1.003 median fold difference, $p = 0.0333$) as well as East Asian ancestry (East Asian ancestry Median = +0.57 median fold difference, $p = 0.0167$) (Figure S2A). We further found that these European ancestry-correlated pAIMs were most abundant in European ancestry cell lines that were homozygous and heterozygous for the pAIM variant allele (+/+ and −/+ genotype) relative to those homozygous for the reference allele (−/−) (MWU $p = 0.0006$). Another subset of 21 pAIMs quantified exhibited >50% allele frequency within African ancestry populations and were significantly elevated in cell lines from individuals of African versus European (MWU $p = 0.0333$) ancestry and trended as elevated in African versus East Asian ancestry cell lines (+1.0676 median fold difference) (Figure S2B). These African ancestry-associated pAIMs were most abundant in African cell lines homozygous and heterozygous for the pAIM variant allele (+/+ and −/+ genotype) relative to those homozygous for the reference allele (−/−), i.e., +/+ versus

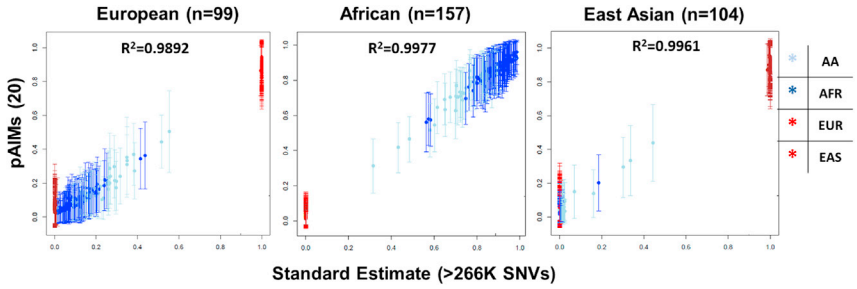


Figure 3. Comparison of African (AFR), African-American (AA), European (EUR), or East Asian (EAS) Global Ancestry Classification using Peptide Ancestry Informative Markers (pAIMs) with Standard Estimates

Random sampling of pAIMs in European, African, and East Asian reference populations from the 1000 genome project revealed that a minimum of 20 pAIM-encoding nonsynonymous single-nucleotide polymorphisms (SNPs) can classify global ancestry with accuracy comparable to standard estimates using >266,000 SNPs, i.e., $R^2 = 0.9943 \pm 0.005$.

−/− cell lines = +1.69 median fold difference. We also quantified 17 pAIMs exhibiting >50% allele frequency within East Asian populations and found that these are elevated in cell lines from individuals of East Asian versus European ancestry (+0.4058 median fold difference) and African (+0.84563 median fold difference) ancestry (Figure S2C). These East Asian ancestry-related pAIMs further tracked with genotype, i.e., +/+ pAIM allele versus −/− cell lines = +1.99 median fold difference.

Quantitation of peptide ancestry informative markers in uterine leiomyoma tissues

We assessed pAIMs within multiplexed (TMT) quantitative proteomic analysis of formalin-fixed, paraffin-embedded uterine leiomyoma tissues from self-described European-American (n = 35) or African-American (n = 31) women (Table 2, Discovery). We quantified 69,056 total peptides that included 33 high-confidence pAIMs (Table S3). Of these candidates, 24 are of unknown clinical significance, whereas 9 have been reported as not being pathogenic for disease (Table S1A, ClinVar). Several of these candidates map to putative functional protein coding domains including intermediate filament domains within keratin 3 (R375G, >50% AF in Europeans) and keratin 13 (A187V, >90% in Europeans and East Asians) as well as within the dihydroxyacetone (Dha)-binding domain of DHA kinase (A185T, >90% in Europeans and East Asians) (Table S1A). We further assessed the performance of a subset of 18 pAIMs quantified within >50% of all uterine leiomyoma tissues to distinguish patient ancestry stratified by self-described race. Unsupervised hierarchical cluster analyses revealed that the abundance of pAIMs exhibiting high allele frequencies within European or African populations trended with self-described European or African-American race (Figure 5A). We also assessed performance of these candidates using a partial least squares (Le Cao et al., 2011) approach and identified, based on principle component analysis, these candidates served to explain 21% and 7% of the variance between African-American and European-American patient tissues (Figure 5B) and can further classify African-American from European-American patients with high accuracy (AUC = 0.9834, $P = 1.6 \times 10^{-11}$, Figure 5C). We also quantified pAIM abundance by multiplexed (TMT), quantitative proteomic analyses within an independent cohort of fresh-frozen uterine leiomyoma tissues from women of European (n = 16) or African (n = 10) ancestry confirmed using standard genotype estimates from companion WGS analyses, as recently described (Bateman et al., 2021) (Table 2, Validation, Table S4). We quantified a total of 61,283 peptides among which 15 pAIMs were quantified in this cohort and 10 were co-quantified with discovery analyses (Table S4). This validated subset was predominated by candidates of unknown clinical significance (ClinVar) and included a substitution in GC vitamin D-binding protein, i.e., (D451E, >50% Europeans). Assessment of pAIMs with allele frequencies >50% in African populations were significantly more abundant in African ancestry patients than pAIMS with allele frequencies of >50% in European ancestry populations (+1.0381 median fold difference, MWU p = 0.003) (Figure 6A). Conversely, we found that pAIMs with allele frequencies >50% in European ancestry populations were significantly more abundant in European ancestry patients than pAIMS with allele frequencies >50% in African populations (+1.74 median fold difference, MWU $p < 0.0001$ (Figure 6B). Furthermore, as observed in the endometrial cancer cell lines, we find that pAIMs abundances directly correlate with patients homozygous and heterozygous for the pAIM variant allele (+/+ and −/+ genotype) relative to those homozygous for the reference allele (−/−), i.e., African and European ancestry +/+ versus −/− patients ($p < 0.0001$).

Table 1. Endometrial cancer cell lines analyzed by whole-genome sequence and quantitative proteomic analyses

Cell line	Histology	Ancestry (standard Estimate)
ACI181	Endometrioid	African
ACI80	Endometrioid	African
NCIEC1	Endometrioid	African
HEC1A	Adenocarcinoma NOS	East Asian
ISHIKAWA	Adenocarcinoma NOS	East Asian
SNGM	Adenocarcinoma NOS	East Asian
ACI52	Endometrioid	European
ACI61	Endometrioid	European
ACI68	Endometrioid	European
AN3CA	Adenocarcinoma NOS	European
KLE	Adenocarcinoma NOS	European
MFE296	Adenocarcinoma NOS	European
RL952	Adenocarcinoma NOS	European

Cell lines include both primary and commercial cell lines, and global ancestry was determined by standard estimates using single-nucleotide variant-derived ancestry informative markers measured by whole-genome sequence analyses.

We further prioritized two pAIMs, i.e., a V237A substitution in protein Serpin Family A Member 1 (SERPINA1, K.DTEEEDFHVDQ(V/A)TTVK) with an allele frequency (AF) = 0.64 in African versus AF = 0.18 in European populations and an A19D substitution in Serine/threonine-protein phosphatase CPPED1 (CPPED1, R.TL(A/D)AFPAEK) with an AF = 0.75 in African versus AF = 0.15 in European populations for targeted analysis using a parallel reaction monitoring assay (Peterson et al., 2012). Stable isotope standard (SIS) peptides were synthesized containing heavy isotope-labeled lysine residues and spiked into fibroid FFPE tissue digests collected from a cohort of African-American (n = 3) and European-American women (n = 3) previously assessed by multiplexed quantitative proteomic analysis (Table S3). Targeted analysis of heavy SIS peptides revealed consistent abundance trends and high dot product similarity scores (dotp) (Schilling et al., 2012) for the top five most abundant y-ions quantified for each peptide relative to library spectra across tissue digests from African-American and European-American women, i.e., 11.1% covariance (CV) in abundance and average dotp = 0.92 $\pm$ 0.01 for the SERPINA1-V237A variant (Heavy SIS, Figures 7A and S3, Table S5) and 10.6% CV in abundance and average dotp = 0.91 $\pm$ 0.01 for the CPPED1-A19D variant (Heavy SIS, Figure 7B, Table S5). Endogenous SERPINA1-V237A and CPPED1-A19D peptides were further co-analyzed in tissue digests collected from African-American women from revealed high relative abundances and dot product similarity scores, i.e., SERPINA1-V237A peptide average dotp = 0.94 $\pm$ 0.01 (Figure 7A) and CPPED1-A19D peptide average dotp = 0.91 $\pm$ 0.03 (Figures 7B and S3). Assessment of endogenous peptides in tissue digests collected from European-American women revealed peptide variants were largely below signal to noise or not detected and exhibited low dotp scores, i.e., SERPINA1-V237A average dotp = 0.33 $\pm$ 0.05 (Figure 7A) and CPPED1-A19D average dotp = 0.1 $\pm$ 0.15 (Figure 7B), suggesting these peptides were not present in these patient samples. Endogenous peptide abundance trends observed by targeted analysis are consistent with abundance trends quantified by multiplexed proteomic analysis for these patient tissue samples (Table S3).

DISCUSSION

Comprehensive analyses of proteogenomic alterations correlating with patient ancestry are necessary to better understand the relationship between racial disparities underlying disease pathogenesis. In gynecologic malignancies, uterine neoplasms exhibit significant racial disparities, specifically increased incidence of uterine leiomyomas (Baird et al., 2003), as well as more aggressive endometrial cancers in African-American versus European-American patients (American Cancer Society, 2016; Rocconi et al., 2016; Tarney et al., 2018). We and others have described race-specific molecular alterations correlating with disease outcome in endometrial cancers (Bateman et al., 2017; Maxwell et al., 2000; Dubil et al., 2018). As a foundational step to support our ongoing investigations of molecular alterations underlying racial disparities in gynecologic malignancies, we have investigated 1,037 nsSNPs exhibiting high allele frequencies within individuals of European, African, and East Asian populations that encode missense substitutions occurring within

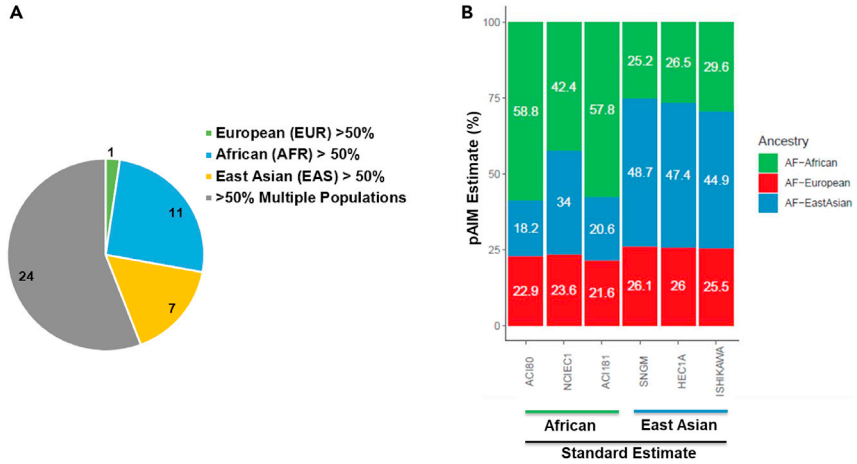


Figure 4. Peptide ancestry informative markers (pAIMs) estimate the ancestry of endometrial cancer cell lines established from women of African or East Asian ancestry

(A) pAIMs quantified in endometrial cancer cell lines encoding missense substitutions corresponding to nonsynonymous SNP exhibiting higher (>50%) allele frequencies uniquely in European, African, or East Asian populations or across multiplex populations.

(B) Estimates of ancestry based on the abundance of pAIMs (43 total quantified) correlated with population-level allele frequencies in European, African, and East Asian populations for cell lines confirmed to be established from women of African or East Asian ancestry using standard genotype estimates from companion whole-genome sequencing data.

putative tryptic peptides, i.e., termed peptide ancestry informative markers (pAIMS) in cell line models and tissues derived from uterine neoplasms collected from women representing diverse ancestral backgrounds. We found that many pAIMs encode substitutions of unknown clinical significance (ClinVar), occur within functional protein domains, and include subsets that are predicted to be deleterious to protein function. We further find that several pAIM variants are linked to an increased risk of developing cancer, such as the F31I substitution in aurora kinase A associated with colon cancer development in East Asian populations (16) and candidates that further track with cancer cell signaling, such as a substitution in nesprin-2 (SYNE2, H3309, rs8010699) that correlates with expression of cyclin-dependent kinase inhibitor 1 (CDK1/p21) in HBV-related hepatocellular carcinomas (Han et al., 2016). We also find that several candidates of unknown clinical significance occur within known functional protein domains including multiple immunoglobulin-like domains, such as in the major histocompatibility complex, class II, DP alpha 1 (HLA-DPA1) protein. These candidates further represent ancestrally linked, variant peptides encoding substitutions that may alter functional protein domain interactions as well as protein complex architecture that require further investigation.

Our efforts have quantified and validated pAIMs in uterine neoplasms from >100 women representing diverse ancestries and consist of a foundational set of ancestry-linked variant peptides observable by multiplexed proteomic analyses of uterine cells and tissues. Many pAIMs quantified are of unknown clinical significance in ClinVar and may warrant further investigation to better understand possible functional roles in the pathogenesis of uterine disease. Validated candidates were largely undocumented for disease pathogenesis in ClinVar but were associated with altered biological functions. One such candidate encodes a substitution in GC vitamin D-binding protein, i.e., D451E, rs7041, exhibiting a higher allele frequency within individuals of European (58%) versus African (7.3%) descent. rs7041 has been shown to correlate with altered levels of GC within European and African Americans (Powe et al., 2013) and further to impact vitamin D metabolism in pregnant women (Ganz et al., 2018). Comparison of cell line and tissue analyses revealed quantitation of pAIMs encoded within Protein-L-isoaspartate(D-aspartate) O-methyltransferase (V178I, >80% AF in African and East Asians), Serine/threonine-protein phosphatase CPPED1 (A19D, >75% AF in Africans), and AH receptor-interacting protein (Q228K, >99% AF in European and East Asians) across all samples. Protein-L-isoaspartate(D-aspartate) O-methyltransferase (PCMT1) regulates methyl esterification of L-isoaspartyl and D-aspartyl residues and has been shown to repair damage to aspartate residues that can occur in an age-dependent manner (DeVry and Clarke, 1999). Interesting, the I178 variant of PCMT1 has been shown to result in increased catalytic activity

Table 2. Uterine leiomyoma (ULM) tissues analyzed by quantitative proteomic (Discovery and Validation) as well as whole-genome sequencing (Validation) analysis

Discovery - ULM tissues (n = 66)	
Ancestry (Self-Described)	#
European-American	35
African-American	31
Validation - ULM tissues (n = 26)	
Ancestry (standard Estimate)	#
European	16
African	10

Global ancestry was determined by standard estimates for the validation cohort using single-nucleotide variant-derived ancestry informative markers measured by whole-genome sequence analyses.

and a more thermostable version of PCMT1, whereas the V178 variants exhibit greater substrate affinity. Serine/threonine-protein phosphatase CPPED1 has been shown to function as a tumor suppressor in bladder cancer via inactivation of Protein Kinase B (AKT) through dephosphorylation of S473 leading to inhibition of cell cycle progression and promotion of apoptosis (Zhuo et al., 2013). Although the A19D ancestry-linked substitution we have observed is not predicted to impact CPPED1 function, we have previously reported higher mutational frequencies in the gene encoding tumor suppressor phosphatase and tensin homolog (PTEN) in European-American versus African-American patients with endometrial cancer, a regulator of AKT signaling, that correlates with improved outcome in European American women (Maxwell et al., 2000). This intersection suggests that regulation of AKT signaling associated protein machinery may be linked to patient ancestry and warrants further investigation of these relationships and their impact on disease pathogenesis.

Our analyses provides proof of concept that estimates of European, African, and East Asian ancestry can be determined by routine proteomic analysis of disease-relevant patient tissues. We find that pAIMs exhibiting higher allele frequencies within European, African, or East Asian ancestry individuals are quantified at significantly greater abundances in cell lines and tissue samples derived from individuals of similar self-described ancestry, i.e., European-American or African-American, and further exhibited abundance trends correlating with patient genotype. However, our *in silico* analysis investigating the identification of the minimum number of ancestry-specific pAIMs that are necessary to predict ancestry in comparison with standard estimates reported in Figure 3 underscores that further refinement and selection of optimal pAIMs in tissues from uterine neoplasms is necessary to prioritize pAIM candidates that are reproducibly observed and exhibit maximal predictive accuracy to estimate European, African, and East Asian ancestry. Our findings also underscore that quantitation of pAIMs is feasible in formalin-fixed, paraffin-embedded as well as fresh frozen human tissue samples. These data extend previous efforts showing that genetically variant peptides can provide estimates of global ancestry in human hair and bone and underscore that these estimates can be determined using multiplexed proteomic data similar to that being generated for population-scale proteogenomic studies focused on improving our understanding of cancer. We further show that we can accurately estimate global ancestry by quantifying as few as 20 population-specific pAIMs in comparison with genotype analyses of ~266K genetic markers using standard estimates ($R^2 > 0.99$). Although our focus has been to characterize pAIMs in uterine neoplasms, these candidates are not tissue dependent and represent a generalized set of markers for proteoancestry assessment in diverse tissues and cell lines. To that end, comparison of proteins encoding pAIMs with large-scale proteomic analyses of diverse human cell and tissue types (Kim et al., 2014) suggests that many of these protein targets are ubiquitously expressed across various tissues, including diverse immune cell types (data not shown), and that quantitation of pAIMs in a wide range of biological samples is likely possible. We further provide proof-of-concept data for the targeted quantitation of two pAIM candidates, i.e., SERPINA1-V237A and CPPED1-A19D, that exhibit higher allele frequencies with African versus European populations in fibroid tissue digests from African-American and European-American women. These analyses verified the relative abundance trends observed for these candidates by multiplexed, quantitative proteomic analysis and further showed these candidates exhibited elevated abundance in African-American women but were largely not detected within European-American women, consistent with the known allele frequencies for these variants in these

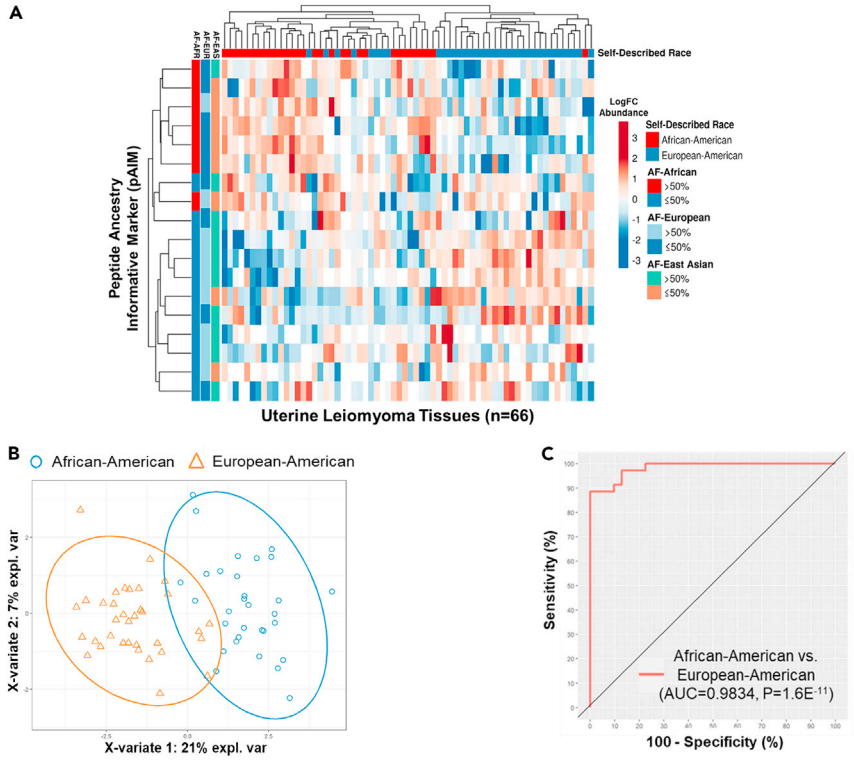


Figure 5. Quantitation of Peptide Ancestry Informative Markers (pAIMs) within Uterine Leiomyoma (ULM) Tissues Collected from a Cohort of Women of Self-Described African or European-American race

(A) Data reflect unsupervised cluster analyses of 18 pAIMs quantified across n66 patient tissue samples; heatmap reflects clustering by Euclidean distance and complete linkage. AF = allele frequency.

(B) Principle component analyses of 18 pAIMs, principal component 1 and principal component 2 that explain 21% and 7% of the total variance, respectively.

(C) Receiver operator curve assessing the performance of 18 pAIMs to classify African-American (n31) versus European-American (n35) patients.

populations. These targeted analyses represent Tier 3 (Carr et al., 2014) targeted mass spectrometry assays that provide proof of concept and require further development and optimization to satisfy expectations for quantitative performance as Tier 1 or 2 targeted MS assays. These findings will support future efforts focused on optimizing tissue-specific, stable isotope-labeled pAIMs panels that can be co-quantified during primary clinical sample analyses.

We have characterized ancestry-linked peptide variants in endometrial cancer cell lines and patient tissues representing a foundational step toward investigating relationships linking patient ancestry with the proteome, genome, and disease pathogenesis. To this end, recent investigations have revealed that clinical serum biomarkers can exhibit altered performance directly as a product of differences in patient genetic ancestry, underscoring the need to integrate proteomics with genetic ancestry assessments at the level of investigational analyses such as biomarker discovery (Sjaarda et al., 2020). Racial disparities in the incidence and outcome of uterine neoplasms persist following adjustments for socioeconomic variables, such as patient access to care (Oliver et al., 2011), and has fueled efforts to define ancestry-linked molecular alterations that may drive these differences. Our ability to discern these relationships will emerge through the assessment of more ancestrally diverse populations of women suffering from gynecologic disease as well as within disease-relevant *in vitro* and *in vivo* models. We recently showed that historic cell line models of cancer have been predominantly established from individuals of European ancestry (Kessler et al., 2019), noting that cancers impacted by racial disparities, such as endometrial cancer, often lacked cell line models from non-European individuals. We confirm here three endometrial cancer cell line models, i.e., NCI-EC1, ACI-80, and ACI-181, are derived from women of African ancestry representing *in vitro* models to investigate ancestry-linked disease biology in endometrial cancers. This

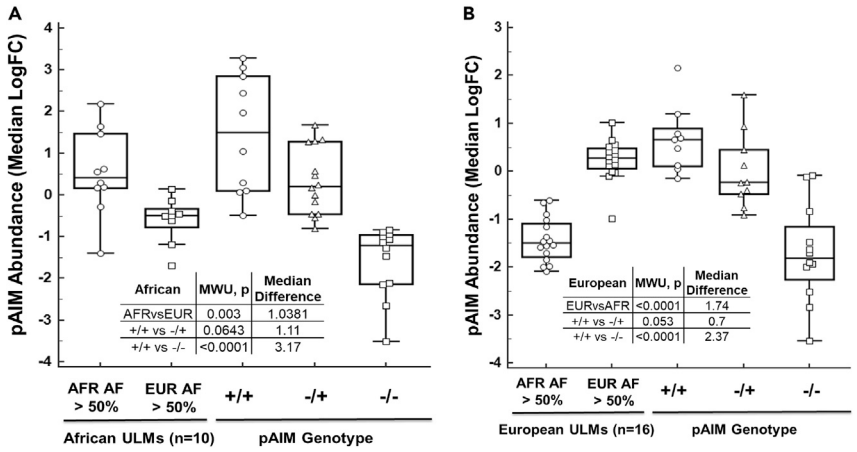


Figure 6. Validation of Peptide Ancestry Informative Markers (pAIMs) within Uterine Leiomyoma (ULM) Tissues Collected from Women with Global Ancestry determined by Standard Estimates

(A) Boxplots details the median logFC abundance for pAIMs exhibiting >50% allele frequencies in African (AFR, 8 of 15 pAIMs quantified) reference populations across n10 ULMs sample from women with AFR ancestry as well as by pAIM variant allele genotype observed by companion whole-genome sequence analyses. MWU = Mann-Whitney U test.

(B) Boxplots details the median logFC abundance pAIMs exhibiting >50% allele frequencies for European (EUR, 5 of 15 quantified) reference populations quantified across n = 16 ULM samples from women with EUR ancestry as well as by pAIM variant allele genotype observed by companion whole-genome sequence analyses.

work will support ongoing efforts focused on linking the proteome and genome with patient ancestry to better understand relationships with disease mechanisms driving racial disparities in the pathogenesis of uterine neoplasms.

Limitations of the study

A limitation of this study is that pAIM variant peptide candidates prioritized correspond to peptides predicted to be fully tryptic and the consideration of variant peptides encoding missense substitutions exhibiting altered allele frequencies within reference populations with partial tryptic specificity may result in the identification of additional pAIM candidates. An additional limitation is that several pAIMs quantified also map to largely predicted proteins (Uniprot.org and noted in Tables S2 thru S4) underscoring the need to further refine pAIMs panels with the goal of prioritizing candidates that estimate patient ancestry with the greatest sensitivity and specificity for a given organ site or proximal biofluid.

DISCLAIMER

The views expressed herein are those of the authors and do not reflect the official policy of the Department of Army/Navy/Air Force, Department of Defense, or US Government.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Prioritization and *in silico* analyses of peptide ancestry information markers
  - Sample preparation for proteomics
  - Sample digestion, TMT labeling and offline fractionation

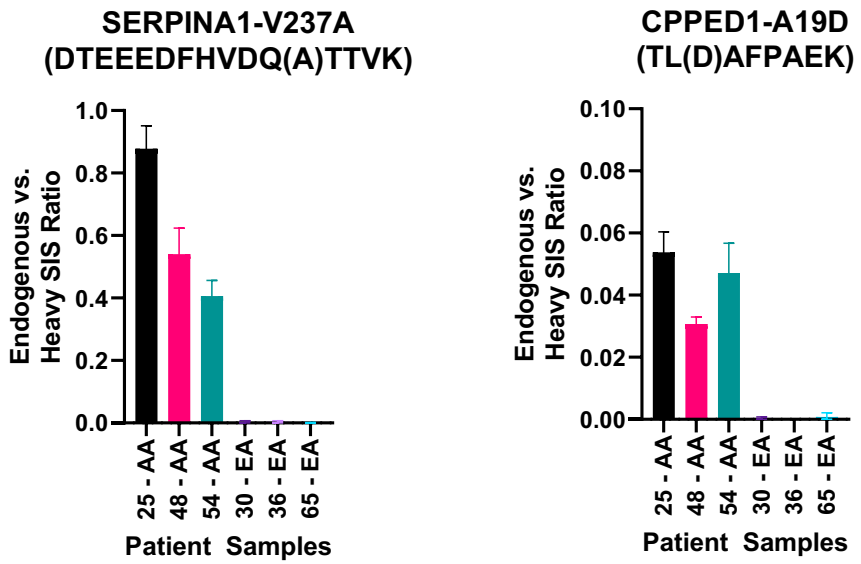


Figure 7. Parallel Reaction Monitoring Analysis of SERPINA1-V237A and CPPED1-A19D pAIM Candidates in Fibroid Tissue Digests from African-American and European-American Women

Stable isotope standard (SIS) peptides were synthesized with heavy-isotope labeled lysine residues corresponding to SERPINA1-V237A pAIM (K.DTEEEDFHVDQ(A)TTVK) and CPPED1-A19D (R.TL(D)AFPAEK) pAIM candidates, and heavy SIS and endogenous peptide abundances were quantified by parallel reaction monitoring assay in fibroid FFPE tissue digests collected from African-American (AA, n = 3) and European-American (EA, n = 3) women. Data reflect the ratio of area under the curve quantified for endogenous versus heavy SIS peptides Relative abundance of SERPINA1-V237A endogenous versus heavy SIS pAIM peptides in AA and EA patient tissues. Relative abundance of CPPED1-A19D endogenous versus heavy SIS pAIM peptides in AA and EA patient tissues.

- Quantitative proteomics - LC-MS/MS analyses
- DNA extraction for endometrial cancer cell lines and uterine fibroid tissue samples
- DNA PCR-free library preparation and whole genome sequencing

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Random sampling analysis
  - Data processing – proteomics
  - Parallel reaction monitoring analysis of SERPINA1-V237A and CPPED1-A19D pAIM candidates
  - DNA WGS processing and variant calling
  - Global ancestry proportion of EC cell lines
  - Global ancestry proportion of uterine fibroid patients

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103665.

ACKNOWLEDGMENTS

This study was supported in part by the U.S. Department of Defense - Uniformed Services University of the Health Sciences (HU0001-16-2-0006 and HU0001-16-2-0014). The authors would like to thank Tracy Litzi, BS; Glenn Gist, BS; Julie Oliver, MS; Dave Mitchell, MS; Domenic Tommarello, MS; and Wei Ao, BS for their technical contributions as well as Albert Dobi, PhD; Gyorgi Petrovics, PhD; and Sara Scannell, BS for critical review of this manuscript.

AUTHOR CONTRIBUTIONS

Contributed to conception: N.W.B., T.D.O., T.P.C., G.L.M. Contributed to experimental design: N.W.B., T.D.O., T.P.C. Contributed to data acquisition, analysis, and/or interpretation of data: N.W.B., T.D.O., C.M.T., T.S.A., B.L.H., K.A.C., M.Z., P.T., A.R.S., A.J., Z.G., J.L., C.T., C.L.D., M.D.W. Drafted and/or revised the article: N.W.B., T.D.O., A.R.S., M.D.K., C.D.S., H.H., M.C., G.J.P., J.S., A.A.-H., J.R.R., N.T.P., Y.C.,

K.M.D., T.P.C., G.L.M. Acquired funding for the research: Y.C., G.L.M. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

G.L.M. is a consultant for Kiyatec, GSK, and Merck. T.P.C. is a ThermoFisher Scientific, Inc SAB member and receives research funding from AbbVie. G.J.P. has received a patent based on concepts presented in this study (US 8,877,455 B2, Australian Patent 2011229918, Canadian Patent CA 2794248, and European Patent EP11759843.3).

Received: May 20, 2021
Revised: October 29, 2021
Accepted: December 17, 2021
Published: January 21, 2022

REFERENCES

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *76*, Chapter 7.

Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinf. *12*, 246.

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Allard, J.E., and Maxwell, G.L. (2009). Race disparities between black and white women in the incidence, treatment, and prognosis of endometrial cancer. Cancer Control *16*, 53–56.

American Cancer Society (2016). Cancer Facts & Figures for African Americans 2016-2018 (American Cancer Society). https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-facts-and-figures-for-african-americans/cancer-facts-and-figures-for-african-americans-2016-2018.pdf.

Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D., and Schectman, J.M. (2003). High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. Am. J. Obstet. Gynecol. *188*, 100–107.

Bateman, N.W., Jaworski, E., Ao, W., Wang, G., Litzi, T., Dubil, E., Marcus, C., Conrads, K.A., Teng, P.N., Hood, B.L., et al. (2015). Elevated AKAP12 in paclitaxel-resistant serous ovarian cancer cells is prognostic and predictive of poor survival in patients. J. Proteome Res. *14*, 1900–1910.

Bateman, N.W., Dubil, E.A., Wang, G., Hood, B.L., Oliver, J.M., Litzi, T.A., Gist, G.D., Mitchell, D.A., Blanton, B., Phippen, N.T., et al. (2017). Race-specific molecular alterations correlate with differential outcomes for black and white endometrioid endometrial cancer patients. Cancer *123*, 4004–4012.

Bateman, N.W., Tarney, C.M., Abulez, T., Soltis, A.R., Zhou, M., Conrads, K., Litzi, T., Oliver, J., Hood, B., Driggers, P., et al. (2021). Proteogenomic landscape of uterine leiomyomas from hereditary leiomyomatosis and renal cell cancer patients. Sci. Rep. *11*, 9371.

Carr, S.A., Abbatiello, S.E., Ackermann, B.L., Borchers, C., Domon, B., Deutsch, E.W., Grant, R.P., Hoofnagle, A.N., Huttenhain, R., Koomen, J.M., et al. (2014). Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. Mol. Cell. Proteomics *13*, 907–917.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS One *7*, e46688.

Deshmukh, S.K., Azim, S., Ahmad, A., Zubair, H., Tyagi, N., Srivastava, S.K., Bhardwaj, A., Singh, S., Rocconi, R.P., and Singh, A.P. (2017). Biological basis of cancer health disparities: resources and challenges for research. Am. J. Cancer Res. *7*, 1–12.

DeVry, C.G., and Clarke, S. (1999). Polymorphic forms of the protein L-isoaspartate (D-aspartate) O-methyltransferase involved in the repair of age-damaged proteins. J. Hum. Genet. *44*, 275–288.

Dubil, E.A., Tian, C., Wang, G., Tarney, C.M., Bateman, N.W., Levine, D.A., Conrads, T.P., Hamilton, C.A., Maxwell, G.L., and Darcy, K.M. (2018). Racial disparities in molecular subtypes of endometrial cancer. Gynecol. Oncol. *149*, 106–116.

Enoch, M.A., Shen, P.H., Xu, K., Hodgkinson, C., and Goldman, D. (2006). Using ancestry-informative markers to define populations and detect population stratification. J. Psychopharmacol. *20*, 19–26.

Farley, J., Risinger, J.I., Rose, G.S., and Maxwell, G.L. (2007). Racial disparities in blacks with gynecologic cancers. Cancer *110*, 234–243.

Felix, A.S., Linkov, F., Maxwell, G.L., Ragin, C., and Taioli, E. (2011). Racial disparities in risk of second primary cancers in endometrial cancer patients: analysis of SEER Data. Int. J. Gynecol. Cancer *21*, 309–315.

Feng Lei, J.L., Shaun-Fei, L., Jian, Z., Hai-Bo, L., An-Quan, J., Jian, Y., Gui-Qiang, W., and Cai-Xia, L. (2019). Development and validation of protein-based forensic ancestry inference method using hair shaft proteome. Prog. Biochem. Biophys. *46*, 81–88.

Fiore, L.D., Rodriguez, H., and Shriver, C.D. (2017). Collaboration to accelerate proteogenomics cancer care: the department of veterans affairs, department of Defense, and the national cancer institute's applied proteogenomics OrganizationaL learning and outcomes (APOLLO) network. Clin. Pharmacol. Ther. *101*, 619–621.

Galanter, J.M., Fernandez-Lopez, J.C., Gignoux, C.R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A.V., Figueroa, L.U., Raska, P., et al. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet. *8*, e1002554.

Ganz, A.B., Park, H., Malysheva, O.V., and Caudill, M.A. (2018). Vitamin D binding protein rs7041 genotype alters vitamin D metabolism in pregnant women. FASEB J. *32*, 2012–2020.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., Mccarthy, S., Mcvean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Gupta, N., Tanner, S., Jaitly, N., Adkins, J.N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R.D., and Pevzner, P.A. (2007). Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res. *17*, 1362–1377.

Han, C., Liao, X., Qin, W., Yu, L., Liu, X., Chen, G., Liu, Z., Lu, S., Chen, Z., Su, H., et al. (2016). EGFR and SYNE2 are associated with p21 expression and SYNE2 variants predict post-operative clinical outcomes in HBV-related hepatocellular carcinoma. Sci. Rep. *6*, 31237.

Ishiguro, J., Ito, H., Tsukamoto, M., Iwata, H., Nakagawa, H., and Matsuo, K. (2019). A functional single nucleotide polymorphism in ABCC11, rs17822931, is associated with the risk of breast cancer in Japanese. Carcinogenesis *40*, 537–543.

Kessler, M.D., Yerges-Armstrong, L., Taub, M.A., Shetty, A.C., Maloney, K., Jeng, L.J.B., Ruczinski, I., Levin, A.M., Williams, L.K., Beaty, T.H., et al. (2016). Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. Nat. Commun. *7*, 12521.

Kessler, M.D., Bateman, N.W., Conrads, T.P., Maxwell, G.L., Dunning Hotopp, J.C., and O'connor, T.D. (2019). Ancestral characterization of 1018 cancer cell lines highlights disparities and reveals gene expression and mutational differences. Cancer *125*, 2076–2088.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

King, J.L., Churchill, J.D., Novroski, N.M.M., Zeng, X., Warshauer, D.H., Seah, L.H., and Budowle, B. (2018). Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit. Forensic Sci. Int. Genet. *36*, 60–76.

Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum. Mutat. *30*, 69–78.

Le Cao, K.A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinf. *12*, 253.

Lee, S., Zhao, L., Rojas, C., Bateman, N.W., Yao, H., Lara, O.D., Celestino, J., Morgan, M.B., Nguyen, T.V., Conrads, K.A., et al. (2020). Molecular analysis of clinically defined subsets of high-grade serous ovarian cancer. Cell Rep. *31*, 107502.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and Maccoss, M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics *26*, 966–968.

Mahdi, H., Schlick, C.J., Kowk, L.L., Moslemi-Kebria, M., and Michener, C. (2014). Endometrial cancer in Asian and American Indian/Alaskan Native women: tumor characteristics, treatment and outcome compared to non-Hispanic white women. Gynecol. Oncol. *132*, 443–449.

Mathias, R.A., Taub, M.A., Gignoux, C.R., Fu, W., Musharoff, S., O'connor, T.D., Vergara, C., Torgerson, D.G., Pino-Yanes, M., Shringarpure, S.S., et al. (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat. Commun. *7*, 12522.

Maxwell, G.L., Risinger, J.I., Hayes, K.A., Alvarez, A.A., Dodge, R.K., Barrett, J.C., and Berchuck, A. (2000). Racial disparity in the frequency of PTEN mutations, but not microsatellite instability, in advanced endometrial cancers. Clin. Cancer Res. *6*, 2999–3005.

Maxwell, G.L., Tian, C., Risinger, J., Brown, C.L., Rose, G.S., Thigpen, J.T., Fleming, G.F., Gallion, H.H., Brewster, W.R., and Gynecologic Oncology Group, S. (2006). Racial disparity in survival among patients with advanced/recurrent endometrial adenocarcinoma: a Gynecologic Oncology Group study. Cancer *107*, 2197–2205.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics *32*, 1749–1751.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Oliver, K.E., Enewold, L.R., Zhu, K., Conrads, T.P., Rose, G.S., Maxwell, G.L., and Farley, J.H. (2011). Racial disparities in histopathologic characteristics of uterine cancer are present in older, not younger blacks in an equal-access environment. Gynecol. Oncol. *123*, 76–81.

Parker, G.J., Leppert, T., Anex, D.S., Hilmer, J.K., Matsunami, N., Baird, L., Stevens, J., Parsawar, K., Durbin-Johnson, B.P., Rocke, D.M., et al. (2016). Demonstration of protein-based human identification using the hair shaft proteome. PLoS One *11*, e0160653.

Pedersen, B.S., and Quinlan, A.R. (2017). Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with Peddy. Am. J. Hum. Genet. *100*, 406–413.

Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S., and Coon, J.J. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol. Cell. Proteomics *11*, 1475–1488.

Powe, C.E., Evans, M.K., Wenger, J., Zonderman, A.B., Berg, A.H., Nalls, M., Tamez, H., Zhang, D., Bhan, I., Karumanchi, S.A., et al. (2013). Vitamin D-binding protein and vitamin D status of black Americans and white Americans. N. Engl. J. Med. *369*, 1991–2000.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Raczy, C., Petrovski, R., Saunders, C.T., Chorny, I., Kruglyak, S., Margulies, E.H., Chuang, H.Y., Kallberg, M., Kumar, S.A., Liao, A., et al. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics *29*, 2041–2043.

Ren, Y., Liu, W., Chen, J., Wang, J., Wang, K., Zhou, J., Cai, S., Zheng, M., Liu, J., Liu, L., and Xue, D. (2017). A missense variant of the ABCC11 gene is associated with Axillary Osmidrosis susceptibility and clinical phenotypes in the Chinese Han Population. Sci. Rep. *7*, 46335.

Rocconi, R.P., Lankes, H.A., Brady, W.E., Goodfellow, P.J., Ramirez, N.C., Alvarez, R.D., Creasman, W., and Fernandez, J.R. (2016). The role of racial genetic admixture with endometrial cancer outcomes: an NRG Oncology/Gynecologic Oncology Group study. Gynecol. Oncol. *140*, 264–269.

Roller, E., Ivakhno, S., Lee, S., Royce, T., and Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. Bioinformatics *32*, 2375–2377.

Ruiz-Iruela, C., Padro-Miquel, A., Pinto-Sala, X., Baena-Diez, N., Caixas-Pedragos, A., Guell-Miro, R., Navarro-Badal, R., Jusmet-Miguel, X., Calmarza, P., Puzo-Foncilla, J.L., et al. (2018). KIF6 gene as a pharmacogenetic marker for lipid-lowering effect in statin treatment. PLoS One *13*, e0205430.

Schilling, B., Rardin, M.J., MacLean, B.X., Zawadzka, A.M., Frewen, B.E., Cusack, M.P., Sorensen, D.J., Bereman, M.S., Jing, E., Wu, C.C., et al. (2012). Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. Mol. Cell. Proteomics *11*, 202–214.

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and NG, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452–W457.

Sjaarda, J., Gerstein, H.C., Kutalik, Z., Mohammadi-Shemirani, P., Pigeyre, M., Hess, S., and Pare, G. (2020). Influence of genetic ancestry on human serum proteome. Am. J. Hum. Genet. *106*, 303–314.

Spratt, D.E., Chan, T., Waldron, L., Speers, C., Feng, F.Y., Ogunwobi, O.O., and Osborne, J.R. (2016). Racial/ethnic disparities in genomic sequencing. JAMA Oncol. *2*, 1070–1074.

Suhre, K., Mccarthy, M.I., and Schwenk, J.M. (2020). Genetics meets proteomics: perspectives for large population-based studies. Nat. Rev. Genet. *22*, 19–37.

Tarney, C.M., Tian, C., Wang, G., Dubil, E.A., Bateman, N.W., Chan, J.K., Elshaikh, M.A., Cote, M.L., Schildkraut, J.M., Shriver, C.D., et al. (2018). Impact of age at diagnosis on racial disparities in endometrial cancer patients. Gynecol. Oncol. *149*, 12–21.

Tennessen, J.A., O'connor, T.D., Bamshad, M.J., and Akey, J.M. (2011). The promise and limitations of population exomics for human evolution studies. Genome Biol. *12*, 127.

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

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.

Xu, L., Zhou, X., Jiang, F., Xu, L., and Yin, R. (2014). STK15 rs2273535 polymorphism and cancer risk: a meta-analysis of 74,896 subjects. Cancer Epidemiol. *38*, 111–117.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523.

Zhu, Y., Orre, L.M., Johansson, H.J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R.M.M., and Lehtio, J. (2018). Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat. Commun. *9*, 903.

Zhuo, D.X., Zhang, X.W., Jin, B., Zhang, Z., Xie, B.S., Wu, C.L., Gong, K., and Mao, Z.B. (2013). CSTP1, a novel protein phosphatase, blocks cell cycle, promotes cell apoptosis, and suppresses tumor growth of bladder cancer by directly dephosphorylating Akt at Ser473 site. PLoS One *8*, e65679.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Uterine Leiomyoma Tissue	Human	Inova Fairfax Hospital
Uterine Leiomyoma Tissue	Human	Johns Hopkins Medical Institute
Chemicals, peptides, and recombinant proteins		
Synthetic Peptide	Thermo Fisher Scientific	TLDAFPAEK
Synthetic Peptide	Thermo Fisher Scientific	DTEEEDFHVDQATTVK
Deposited data		
LC-MS PRM Results	Panorama	https://panoramaweb.org/Gy37K1.url
LC-MS Data Files	ProteomeXchange	PXD029323
Experimental models: Cell lines		
Cell Line	ATCC	AN3CA
Cell Line	ATCC	KLE
Cell Line	ATCC	RL-95-2
Cell Line	DSMZ GmbH	MFE-296
Cell Line	National Cancer Institute	NCI-EC1
Cell Line	Memorial Health University Medical Center	ACI-80
Cell Line	Memorial Health University Medical Center	ACI-181
Cell Line	Memorial Health University Medical Center	ACI-52
Cell Line	Memorial Health University Medical Center	ACI-61
Cell Line	Memorial Health University Medical Center	ACI-68
Cell Line	JCRB	SNG-M
Cell Line	ATCC	HEC1A
Cell Line	ATCC	Ishikawa
Software and algorithms		
Software	https://metascape.org/gp/index.html#/main/step1	MetaScape
Software	https://dalexander.github.io/admixture/	ADMIXTURE
Software	https://www.cog-genomics.org/plink/2.0/	PLINK v1.90b3.32
Software	http://db.systemsbiology.net/kaviar/cgi-pub/Kaviar.pl	Kaviar
Software	https://www.matrixscience.com/	MASCOT v. 2.6.0
Software	Thermo Fisher Scientific	Proteome Discoverer v2.2.0.388
Software	https://proteowizard.sourceforge.io/download.html	MSConvert v. 3.0.19106
Software	https://github.com/yafeng/SpectrumAI	SpectrumAI
Software	https://www.medcalc.org/	MedCalc v. 20.014
Software	https://biit.cs.ut.ee/clustvis/	ClustVis (BETA)
Software	https://bioconductor.org/packages/release/bioc/html/mixOmics.html	mixOmics v.6.8.5
Software	https://www.rstudio.com/	RStudio v. 3.6.0

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software	https://skyline.ms/project/home/begin.view?	Skyline-daily 64-bit, v. 21.0.9.118
Software	https://support.illumina.com/sequencing/sequencing_software/hiseq-analysis-software-v2-1.html	Illumina's HiSeq Analysis Software v. 2.5.55.1311
Software	https://samtools.github.io/bcftools/	bcftools
Software	https://github.com/brentp/peddy	Peddy
Software	https://useast.ensembl.org/Homo_sapiens/Tools/VEP	Variant Effect Predictor

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to the Lead Contact, Dr. Nicholas W. Bateman (batemann@whirc.org).

Materials availability

This study did not generate any unique reagents.

Data and code availability

Supplemental data tables include chromosomal loci, missense substitutions and amino acid positions, reference SNP Ids (rsIDs) corresponding to peptide ancestry informative markers (pAIMs), as well as pAIM abundances quantified in endometrial cancer cell lines and uterine leiomyoma tissues. TMT-10/11 LC-MS data files have been deposited in ProteomeXchange under dataset identifier PXD029323. Skyline files for PRM analysis have been deposited on Panorama and can be accessed here: https://panoramaweb.org/Gy37K1.url.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Commercial endometrial cancer cell lines, i.e. MFE-296 cells were purchased from DSMZ GmbH (Berlin, Germany) and cultured in MEM with 10% fetal bovine serum (FBS) and 1X penicillin/streptomycin (Pen/Strep). SNG-M cells were purchased from the JCRB cell bank (Osaka, Japan) and cultured in Ham's F12 median with 10% FBS and 1X Pen/Strep. HEC1A (McCoy's 5a with 10% FBS and 1X Pen/Strep), RL95-2 (DMEM:F12, 0.005 mg/mL insulin, 10% FBS and 1X Pen-Strep), Ishikawa (MEM with 2 mM Glutamine, 5% Fetal Bovine Serum (FBS) and 1X Pen/Strep), AN3CA (EMEM with 10% FBS and 1X Pen/Strep) and KLE (DMEM:F12 with 10% FBS and 1X Pen/Strep) cells were from ATCC (Manassas, VA). NCI-EC1, ACI-80, ACI-181, ACI-52, ACI-61 and ACI-68 were obtained from John Risinger at Michigan State University (East Lansing, MI) and maintained in DMEM:F12 with 10% FBS and 1X Pen/Strep (ATCC). Uterine leiomyoma tissues (ULM, discovery cohort), Age-matched, formalin-fixed, paraffin embedded uterine leiomyomas from self-described white and black women were obtained under an institutional-review board approved protocol from Inova Fairfax Hospital (Falls Church, VA). Uterine leiomyoma tissues (ULM, validation cohort), flash-frozen uterine leiomyomas were obtained under an institutional-review board approved protocol from Johns Hopkins Memorial Institute (Baltimore, MD).

METHOD DETAILS

Prioritization and *in silico* analyses of peptide ancestry information markers

80,855,907 SNPs for European (i.e. FIN, GBR, IBS, and TSI; N=404), African (i.e. ESN, GWD, LWK, MSL, and YRI; N=504), and East Asian (i.e. CDX, CHB, CHS, and KHV; N=400) populations largely balanced for males and females were downloaded from the 1000 Genomes Project (Genomes Project et al., 2015). SNPs were further mapped to refSeq version 85 to identify candidate missense variants (646,768 nsSNPs) occurring within putative tryptic peptides $\geq$6 or <40 amino acids in length that did not encode isoleucine to leucine substitutions and were further filtered for those exhibiting $\geq$ 50% difference in allele frequency between European, African, and East Asian populations. These substitutions were then mapped to several RefSeq proteome databases to confirm amino acid substitution positions and tryptic specificity resulting in a final

subset of 1,037 candidates. nsSNPs encoding pAIMs were mapped to rsIDs using Kaviar (http://db.systemsbiology.net/kaviar/cgi-pub/Kaviar.pl) against hg19 (GRCh37). Candidates were further assessed using the Variant Effect Predictor tool (https://useast.ensembl.org/Homo_sapiens/Tools/VEP) against hg19 (GRCh37) and extracted metadata detailing putative clinical significance, i.e. ClinVar traits and clinical significance, as well as predicted impact on protein function, i.e. SIFT (Sim et al., 2012) and Poly-Phen-2 (Adzhubei et al., 2013) and PROVEAN (Choi et al., 2012) outputs. Candidates were further mapped to the Uniprot resource and amino acid position of pAIM substitutions were correlated with putative functional domains within parent proteins and pAIM substitutions positions mapping with putative domain regions were prioritized for downstream analyses. Functional enrichment analyses was performed using the "express analysis" settings in MetaScape (Zhou et al., 2019). Determination of fixation index for pAIMs. Fixation index (FST) values were calculated using reference and alternative allele frequency data from the 1000 genomes project. The following equation was used; FST = (HT-HS)/HT, where HT is the expected heterozygosity in the total population and HS is the expected heterozygosity in the subpopulation (Weir and Cockerham, 1984).

Sample preparation for proteomics

Cell lines were plated at equivalent densities and lysed at ~80% confluency in 1% SDS and 10 mM Tris-HCl, pH 7.4. Cell lysates (30 μg) were run approximately one inch into a 4–15% bis-acrylamide gel (Bio-Rad) and processed for in-gel digestion as previously described (Bateman et al., 2015). Briefly, gel spots were excised, dehydrated with 100% acetonitrile (ACN), digested with trypsin and peptides were extracted using were extracted with 70% ACN, 5% formic acid, dried by vacuum centrifugation and resuspended in 100 mM TEAB. Peptide extracts were labelled with TMT-11 reagents (Thermo Fisher Scientific) as described below. Discovery uterine leiomyoma cohort - thin (10 μm) tissue sections were cut using a cryostat and placed on polyethylene naphthalate membrane slides. After staining with aqueous H&E, laser microdissection (LMD) was used to harvest tumor cells from thin sections. Validation uterine leiomyoma cohort - tissues were cryopulverized, resuspended in 100 mM triethylammonium bicarbonate (TEAB), and sonicated. Protein was quantified by BCA Protein Assay (Thermo Scientific) and 50 μg of total protein in 100 mM TEAB and 10% acetonitrile was incubated at 99°C for 1 h.

Sample digestion, TMT labeling and offline fractionation

Sample digestion, TMT labeling and offline fractionation of tissue samples was performed as recently described (Lee et al., 2020). Briefly, tissues were collected in 20 μL of 100 mM TEAB in MicroTubes (Pressure Biosciences, Inc). Following the addition of 1 μg of SMART Digest Trypsin (Thermo Scientific) to each sample, MicroTubes were capped with MicroPestles. Pressure-assisted lysis and digestion was performed in a barocycler (2320 EXT, Pressure BioSciences, Inc) by sequentially cycling between 45 kpsi and atmospheric pressure for 60 cycles at 50°C. The peptide digests were transferred to 0.5 mL microcentrifuge tubes, vacuum-dried, resuspended in 100 mM TEAB, pH 8.0 and the peptide concentration of each digest was determined using the bicinchoninic acid assay (BCA assay). Forty-fifty micrograms of peptide from each sample, along with a pooled reference sample assembled from equivalent amounts of peptide digests pooled from individual patient samples for individual sample sets, were aliquoted into a final volume of 100 μL of 100 mM TEAB and labeled with tandem-mass tag (TMT) isobaric labels (TMT10 or 11plex™ Isobaric Label Reagent Set, Thermo Fisher Scientific) according to the manufacturer's protocol. Each TMT sample plex was fractionated by basic reversed-phase liquid chromatography (bRPLC) into 96 fractions through development of a linear gradient of acetonitrile (0.69%/min). Concatenated fractions (36 total pooled samples for the endometrial cancer cell lines, FFPE ULMs, 24 total for the frozen ULMs) were generated in a serpentine fashion for global LC-MS/MS analysis.

Quantitative proteomics - LC-MS/MS analyses

The TMT sample multiplex bRPLC fractions were resuspended in 100 mM NH_4HCO_3 (pH 7.0) and analyzed by LC-MS/MS employing a nanoflow LC system (EASY-nLC 1200, Thermo Fisher Scientific) coupled online with an Orbitrap Fusion Lumos MS or Q-Exactive HF-X (Thermo Fisher Scientific). In brief, each fraction (~500 ng total peptide) was loaded on a nanoflow HPLC system fitted with a reversed-phase trap column (Acclaim PepMap100 C18, 20 mm, nanoViper, Thermo Scientific) and a heated (50°C) reversed-phase analytical column (Acclaim PepMap RSLC C18, 2 μm, 100 Å, 75 μm × 500 mm, nanoViper, Thermo Fisher Scientific) coupled online with the MS. Peptides were eluted by developing a linear gradient of 2% mobile phase B (95% acetonitrile, 0.1% formic acid) with 98% mobile phase A (0.1% formic acid) to 32% mobile phase B over 120 min at a constant flow rate of 250 nL/min. For both instrument platforms, the electrospray

source capillary voltage and temperature were set at 2.0 kV and 275°C, respectively. High resolution (R=60,000 at m/z 200) broadband (m/z 400–1600) mass spectra (MS) were acquired, followed by selection of the top 12 most intense molecular ions in each MS scan for high-energy collisional dissociation (HCD). Instrument specific parameters were set as follows for each instrument platform. Orbitrap Fusion Lumos - Full MS: AGC, 5×10^5; RF Lens, 30%; Max IT, 50 ms; Charge State, 2–4; Dynamic Exclusion, 10ppm/20 sec; MS2: AGC, 1×10^5; Max IT, 120 ms; Resolution, 50k; Quadrupole Isolation, 0.8 m/z; Isolation Offset, 0.2 m/z; HCD, 38; First Mass, 100. Q Exactive HF-X - Full MS: AGC, 3×10^6; RF Lens, 40%; Max IT, 45 ms; Charge State, 2–4; Dynamic Exclusion, 10ppm/20 sec; MS2: AGC, 1×10^5; Max IT, 95 ms; Resolution, 45k; Quadrupole Isolation, 1.0 m/z; Isolation Offset, 0.2 m/z; NCE, 34; First Mass, 100; Intensity Threshold, 2×10^5; TMT Optimization, On.

DNA extraction for endometrial cancer cell lines and uterine fibroid tissue samples

Cells were lysed in SNET Digestion buffer (20 mM Tris, 5 mM EDTA, 400 mM NaCl, 1% w/v SDS) with Proteinase K (Thermo Fisher Scientific, Pittsburgh, PA) overnight at 55°C with intermittent shaking. An equal volume of 25:24:1 phenol:chloroform:isoamyl alcohol (Thermo Fisher Scientific, Pittsburgh, PA) was added, mixed by inversion and centrifuged at 21,000 x *g* for 5 min. The supernatant was removed to a new tube, half sample volume of 7.5 M ammonium acetate and one and a half sample volumes of 100% ethanol were added, mixed by inversion and centrifuged at 14,000 x *g* for 10 min. The DNA pellet was washed with 70% ethanol and centrifuged at 14,000 x *g* for 5 min. The supernatant was removed and the pellet was air-dried and resuspended in 50 μL of 10 mM Tris buffer. Tissue scrolls were generated from OCT-embedded fresh-frozen tumors and collected in ATL buffer (Qiagen Sciences LLC, Germantown, MD) followed by storage at −80°C until isolation. Samples were normalized to 360 μL ATL buffer, 40 μL of Proteinase K was added for lysis and incubated at 56°C for 4 h with intermittent shaking. Isolation was performed according to the manufacturer's protocol (DNA Purification from Tissues) using the QiAamp DNA Mini Kit (Qiagen Sciences LLC, Germantown, MD). DNA was eluted after a 10 min incubation with 40 μL of Buffer AE, followed by another 10 min incubation with 160 μL of nuclease-free water (Thermo Fisher Scientific) and reduced to 50 μL using a CentriVap Concentrator (Labconco, Kansas City, MO). Quantity and 260/280 purity reading was established using the Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific) and Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific) according to manufacturer's protocols using eight standards from 0 to 50 ng/uL. Samples were run in triplicate and measurements were taken on a SpectraMax M4 microplate reader (Molecular Devices, San Jose, CA).

DNA PCR-free library preparation and whole genome sequencing

TruSeq DNA PCR-free Library Preparation Kit (Illumina, San Diego, CA) was performed following manufacturer's instructions. Briefly, genomic DNA (gDNA) was diluted to 20 ng/μL using Resuspension Buffer (RSB, Illumina) and 55 μL were transferred to Covaris microTubes (Covaris, Woburn, MA). The normalized gDNA was then sheared on an LE220 focused-ultrasonication system (Covaris) to achieve target peak of 450 bp with an Average Power of 81.0 W (SonoLab settings: duty factor, 18.0%; peak incident power, 45.0 watts; 200 cycles per burst; treatment duration, 60 s; water bath temperature, 5°C–8.5°C). The quality of the final DNA libraries was assessed with the High Sensitivity dsDNA (AATI). Per manufacturer's protocol, library peak size was in the range of 550 to 620 bp. The DNA libraries were quantified by real- time quantitative PCR, using the KAPA SYBR FAST Library Quantification Kit (KAPA Biosystems, Boston, MA) optimized for the Roche LightCycler 480 instrument (Roche). DNA libraries were then normalized to 2 nM and clustered on the Illumina cBot 2 at 200pM using a HiSeq X Flowcell v2 and the HiSeq X HD Paired-End Cluster Generation Kit v2. Paired-end sequencing was performed with the HiSeq X HD SBS Kit (300 cycles) on the Illumina HiSeq X.

QUANTIFICATION AND STATISTICAL ANALYSIS

Random sampling analysis

After the 1037 pAIMs were selected from the 1000 Genomes Project training data, we then performed ADMIXTURE (Alexander et al., 2009) randomly sampling 20 pAIMs without replacement, using the supervised learning version (Alexander and Lange, 2011) with the training data as fixed for their respective ancestries (K=3 for African, European, and East Asian). Test data also came from the following 1000 Genomes Project populations (Genomes Project et al., 2015), which were not included in the training: JPT to represent East Asian, CEU to represent European, and both ACB and ASW as admixed African ancestry. This procedure was repeated 500 times with the mean and standard deviation calculated and plotted against

the estimates for these same samples from 266,403 SNPs. This larger set of sites was selected in the standard way (Kessler et al., 2016), where they were LD pruned using the PLINK(Purcell et al., 2007) command –indep-pairwise 50 5 0.1), and filtered for allele frequency of 5%.

Data processing – proteomics

Peptide ancestry informative marker identifications are generated by searching .RAW data files with a publicly-available human proteome database (RefSeq human proteome, downloaded 9/22/17 and updated with entries from 8/19/2019 when pAIM amino acid positions did not map to entries in 9/22/17 version) appended with 1,037 full-length proteins encoding predicted pAIM substitutions of interested using Mascot (v2.6.0, Matrix Science) and Proteome Discoverer (v2.2.0.388, Thermo Fisher Scientific) as previously described (Lee et al., 2020). Briefly, samples are searched using the following parameters: precursor mass tolerance of 10 ppm, fragment ion tolerance of 0.05 Da, a maximum of two tryptic miscleavages, dynamic modifications for oxidation (15.9949 Da) on methionine residues and TMT reporter ion tags (229.1629 Da) on peptide N-termini and lysine residues. Peptide spectral matches (PSMs) are filtered using a false-discovery rate (FDR) < 1.0% (Percolator q-value < 0.01). TMT reporter ion intensities are extracted at a mass tolerance of 20 ppm and PSMs lacking TMT reporter ion signal in the pooled reference channel, in all patient TMT channels or exhibiting an isolation interference of $\geq$50% were excluded from downstream analyses. Normalized peptide abundance was determined by calculating TMT reporter ion ratios relative to a pooled reference channel and applying a mode centered, z-score transformation of each sample channel per TMT multiplex, i.e. normalized (PSM (Log_2Ratio) = (PSM (Log_2Ratio) – ModeCenter (PSM (Log_2Ratio) / σ (PSM (Log_2Ratio). Peptide fragment ion spectra for pAIM peptide variants are confirmed to encode the substitution of interest by automated inspection of MS2 fragment ion spectra in .mzML converted .RAW files (MSConvert, Proteowizard) using SpectrumAI (Zhu et al., 2018) and the following settings, fragment ion tolerance was set to 50 ppm, relative set to True and a candidate was considered high-confidence if diagnostic ions flanking the substitution of interest were identified in at least one sample multiplex in endometrial cancer cell line and uterine leiomyoma cohorts. Estimates of ancestry proportions using pAIM abundance reflect Pearson correlations of pAIM abundances with allele frequencies for African, European or East Asian populations reported by 1000 genomes for pAIMs quantified, where +1 is added to each of these values followed by division by +2. Resulting values are then normalized to 100% to identify the relative proportions European, East Asian and African ancestry. Differential analyses of pAIM abundance was performed using Mann-Whitney U rank sum testing in MedCalc (version 19.0.3). pAIMs variants were visualized in heatmaps and by principle component analysis (PCA) using default settings in the ClustVis web tool (https://biit.cs.ut.ee/clustvis/). Sparse Partial Least Squares Discriminant analysis (sPLS-DA) (Le Cao et al., 2011) was performed using mixOmics (ver 6.8.5) in RStudio (ver 3.6.0). The sPLS-DA model was run in regression mode for an optimized 18 feature set for two principal components. Plot loadings were calculated using the median to assess contribution onto the first principal component. The AUC and ROC curve were generated from the sPLS-DA model.

Parallel reaction monitoring analysis of SERPINA1-V237A and CPPED1-A19D pAIM candidates

Stable isotope standard versions of SERPINA1-V237A (DTEEEDFHVDQ(A)TTVK) and CPPED1-A19D (TL(D)AFPAEK) pAIM peptide candidates were synthesized with heavy isotope labelled lysine (K, i.e. C13(6), 15N(2) residues (Heavy SIS peptides, Thermo Fisher Scientific). A total of 10fmol final of heavy SIS peptides was spiked into ~1 ug total fibroid tissue digest for each patient sample and samples were analyzed in triplicate by LC-MS/MS employing a nanoflow LC system (EASY-nLC 1200, Thermo Fisher Scientific) coupled online with a Q-Exactive HF-X (Thermo Fisher Scientific). Briefly, each sample was loaded on a nanoflow HPLC system fitted with a reversed-phase trap column (Acclaim PepMap100 C18, 20 mm, nanoViper, Thermo Scientific) and a heated (50°C) reversed-phase analytical column (Acclaim PepMap RSLC C18, 2 μm, 100 Å, 75 μm × 150 mm, nanoViper, Thermo Scientific) coupled online with the MS. Peptides were eluted by developing a linear gradient of 2% mobile phase B (95% acetonitrile, 0.1% formic acid) to 32% mobile phase B over 60 min at a constant flow rate of 300 nL/min, then to 99% mobile phase B over an additional 15 min, followed by flushing and re-equilibration prior to the next injection. Source capillary voltage and temperature were set at 2.0 kV and 275°C, respectively. High resolution (R=120,000 at m/z 200) broadband (m/z 400–1400) mass spectra (MS) were acquired in profile mode, followed by selection of the top 12 most intense molecular ions in each MS scan for high-energy collisional dissociation (HCD). Instrument specific global parameters were set as follows: Full MS: AGC, 3×10^6; RF Lens, 40%; Max IT, 50 ms; Charge State, 2–3; MS2: AGC, 1×10^5; Max IT, 50 ms; Resolution, 15k; Quadrupole Isolation, 1.0 m/z; Isolation

Offset, 0.2 m/z; NCE, 30; Intensity Threshold, 1.6×10^5; Dynamic Exclusion, Off; TMT Optimization: Off. Inclusion list entries consisted of the endogenous/variant and heavy/variant SIS pAIMs peptides and the top 12 most abundance molecular ions were selected for data-dependent acquisition when not monitoring inclusion list masses. All targeted peptides were fragmented at an NCE of 28. RAW data files were searched using parameters noted above excluding static modifications for TMT reporter ion tags. RAW data files and Proteome Discoverer search results were imported into Skyline-daily (64-bit, v. 21.0.9.118) (MacLean et al., 2010) using the "Import PRM Peptide Search" workflow and the top five y-ions quantified for heavy stable isotope standard (SIS) and endogenous versions of the SERPINA1-V237A (z = 3+) and CPPED1-A19D (z = 2+) pAIM peptide candidates were extracted using default settings for orbitrap instrumentation. Sum y-ion intensities and dot product similarity scores were exported from Skyline and plotted using Graphpad Prism (v 8.4.3).

DNA WGS processing and variant calling

WGS sample raw reads were aligned to the hg19 reference genome and further processed through the Re-sequencing workflow within Illumina's HiSeq Analysis Software (HAS; Isis version 2.5.55.1311; https://support.illumina.com/sequencing/sequencing_software/hiseq-analysis-software-v2-1.html). This workflow utilizes the Isaac read aligner (iSAAC-SAAC00776.15.01.27) and variant caller (starka-2.1.4.2) (Raczy et al., 2013), the Manta structural variant caller (version 0.23.1) (Chen et al., 2016), and the Canvas CNV caller (version 1.1.0.5) (Roller et al., 2016).

Global ancestry proportion of EC cell lines

To estimate admixture proportions, we used reference samples with known ancestry from the 1000 Genome Project (Genomes Project et al., 2015). These reference samples were comprised of 99 European samples from the CEU population, 108 African samples from the YRI population, and 208 Asian samples from the CHB and JPT populations (414 total references samples). Working with genotype data from these reference samples and each of the endometrial cancer cell lines, we first used bcftools(Narasimhan et al., 2016) to remove Indels and non-biallelic variants. We then used Plink v1.90b3.32 (Purcell et al., 2007) to remove singleton sites (i.e. variants with only one alt allele copy in the dataset. After this processing, we merged these datasets with Plink by taking the intersect of autosomal variants and removing any ambiguous A/T and G/C variants from this combined data. We then pruned for linkage using the plink linkage pruning algorithm command of –indep-pairwise 50 5 0.5, which uses a window of 50 with an r^2 greater than 0.5 and a SNP step of 5. Ancestry was estimated for each cell line independently, and so this process was repeated independently for each cell line. At the end of this processing, we were left with ~500,000 LD-pruned variants per reference cell line merged dataset (range 459,592–701,192).

Global ancestry proportion of uterine fibroid patients

We predicted sample super population ancestries using the methods implemented in Peddy (Pedersen and Quinlan, 2017). Briefly, principal component reduction was performed on genotype calls at specific loci from 2,504 samples in the 1000 Genomes project and a support vector machine (SVM) classifier was trained on the resulting first four components, using known ancestries as the training labels. Genotype calls at the same loci from each sample collected in this study were then mapped to principal component space and the trained SVM was used to predict ancestries. All classifier prediction probabilities were >0.89.