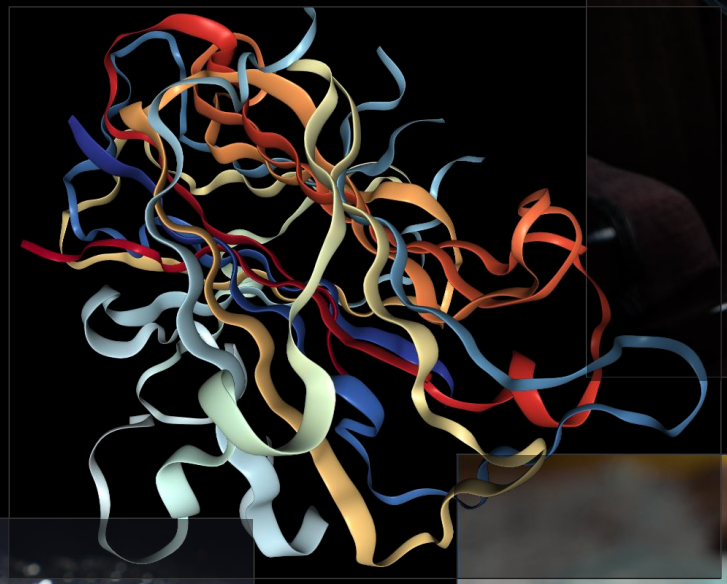
Human Forensic Identification from Fingerprint Touch Samples Using Proteogenomics



Myles W. Gardner, Ph.D.
Senior Staff Scientist, S&T Advisor
Signature Science, LLC

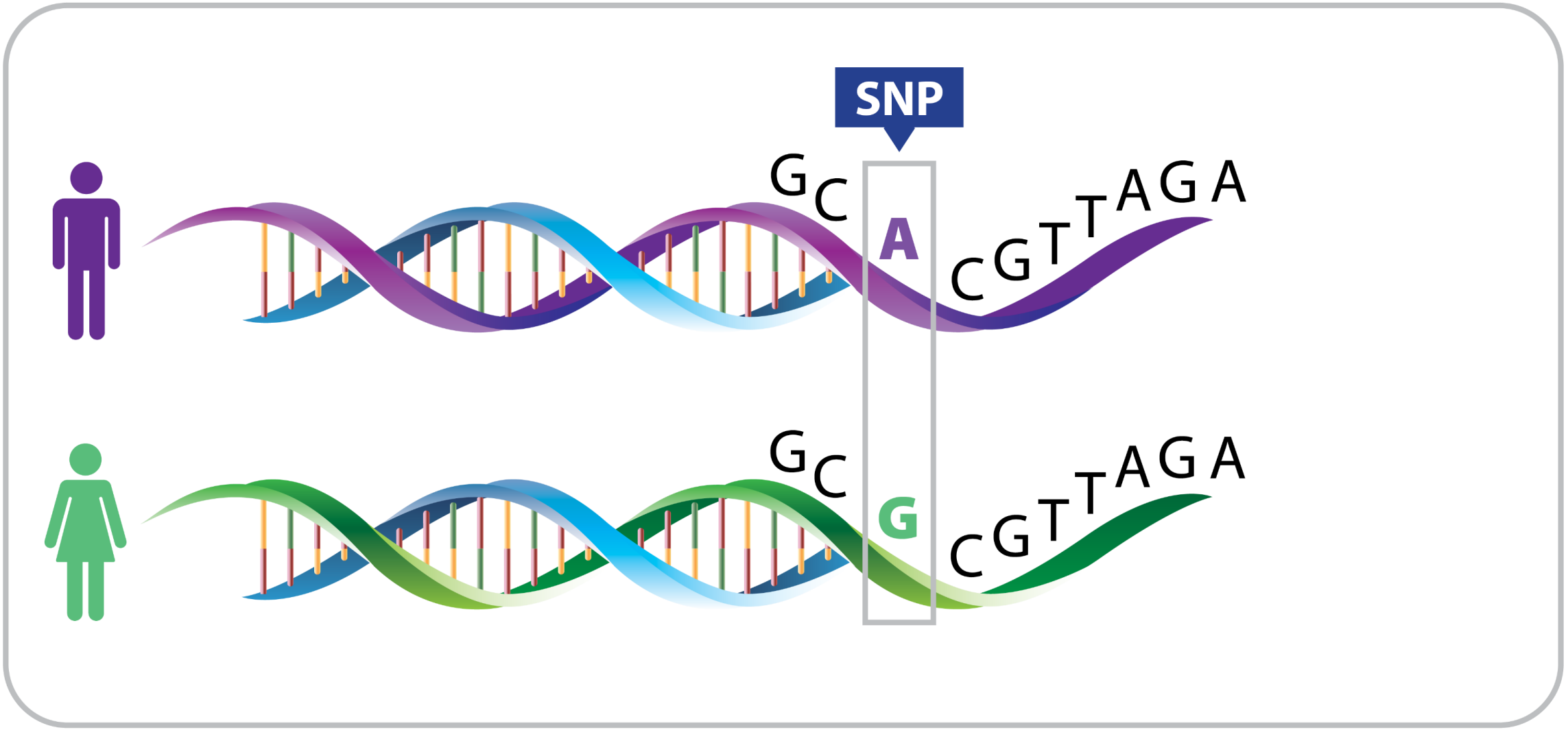
signature science LLC

69th Annual ASMS Conference on Mass Spectrometry and Allied Topics
Forensics: Innovations and Applications, TOH am 08:30
Tuesday, November 2, 2021

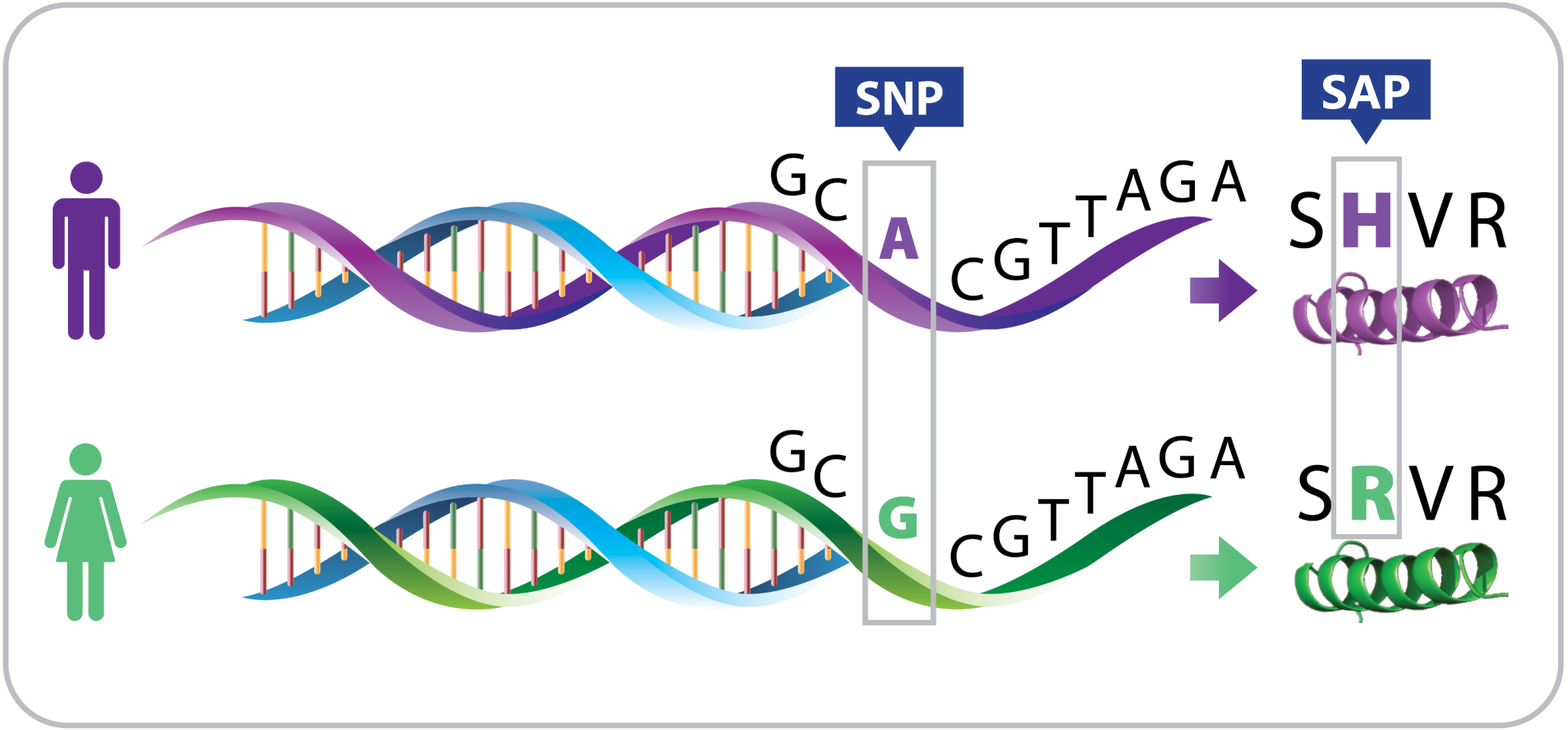
Outline

- Background on Human Forensics
- IARPA Proteos Program
- General Workflow for Human ID from Fingerprints
- Sample Preparation and Chemical Analysis Methods
- Proteogenomics Data Analysis
- Results and Program Highlights
- Limitations
- Future Research Considerations

Current State of the Art – DNA Forensics

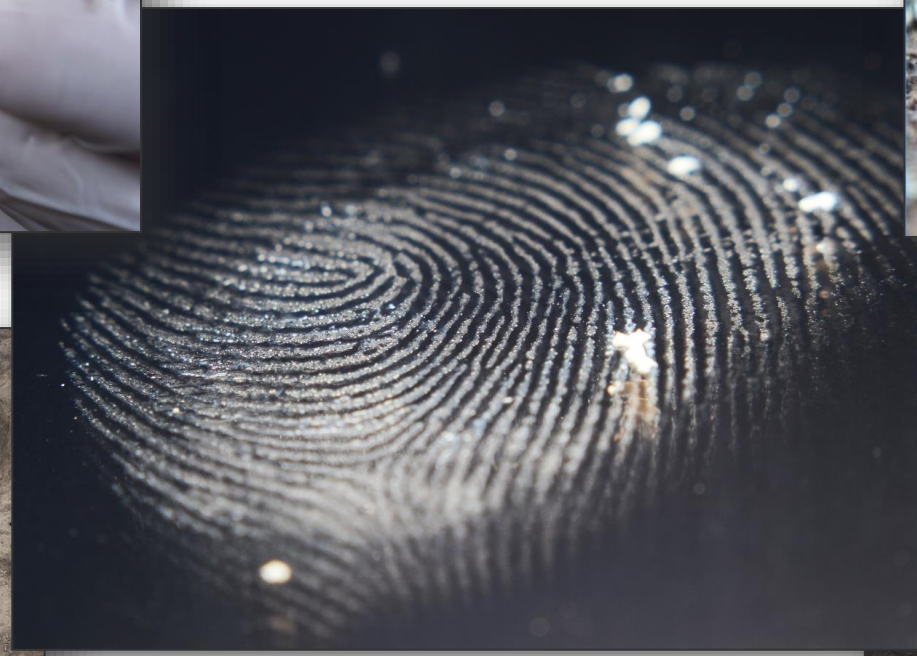


Current State of the Art – DNA Forensics



SAP = Single amino acid polymorphism
GVP = Genetically variable peptide

Challenges in Human Forensic Identification



IARPA Proteos Program

- Intelligence Advanced Research Projects Activity (IARPA) Proteos program aimed to:
  - Develop novel methods for human identification by sequencing of protein variants in touch samples
  - Develop and optimize sample preparation methods to efficiently co-extract protein and DNA in parallel
  - Augment DNA forensic identification methods when there is limited or low-quality DNA available
- 3-year program with multiple R&D and T&E phases focused on challenging developed methods against difficult sample types
  - Brass shell casings; post-IED blast material; elevator buttons; keyboard keys; wood; metal; currency; mixed contributor samples

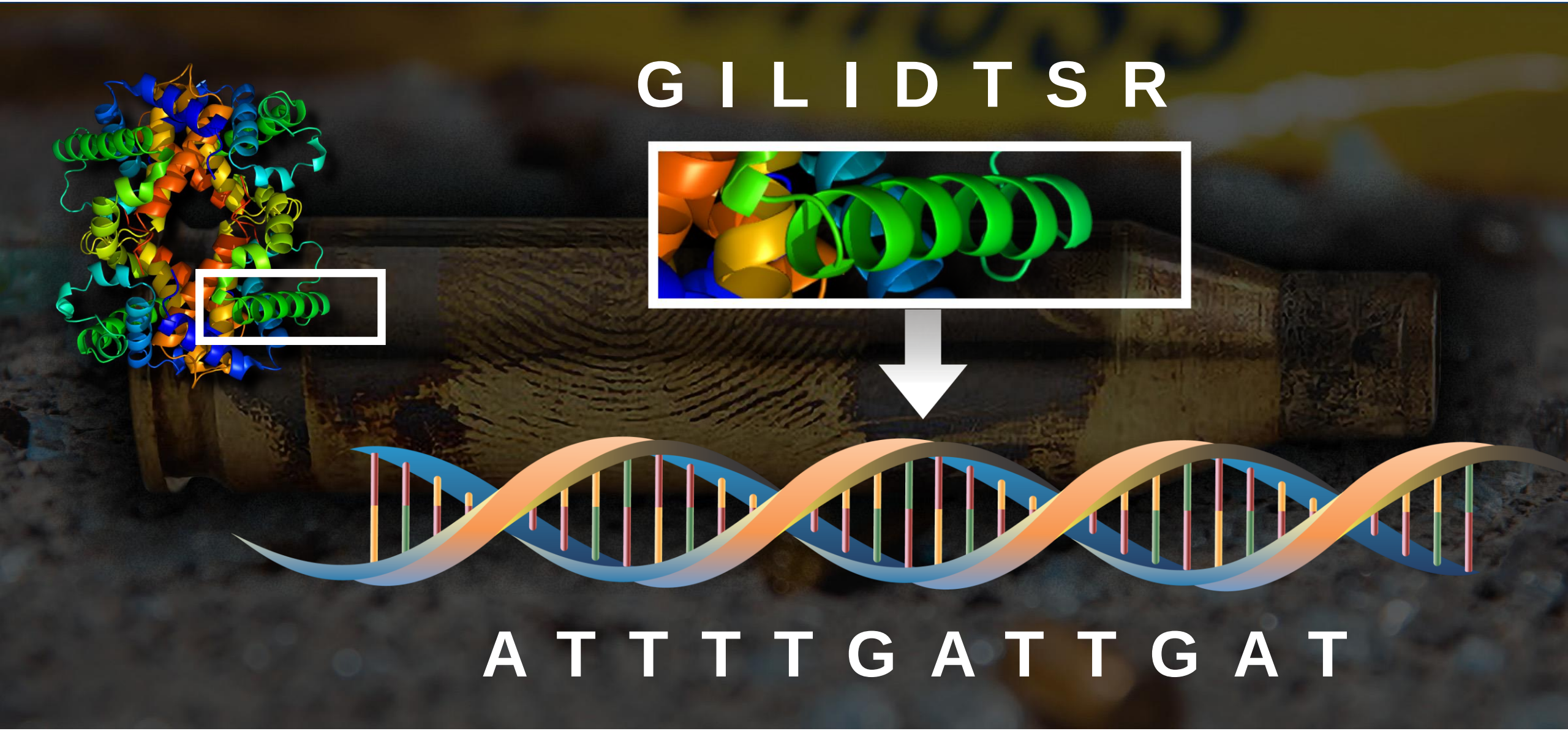
Office of the Director of National Intelligence

IARPA

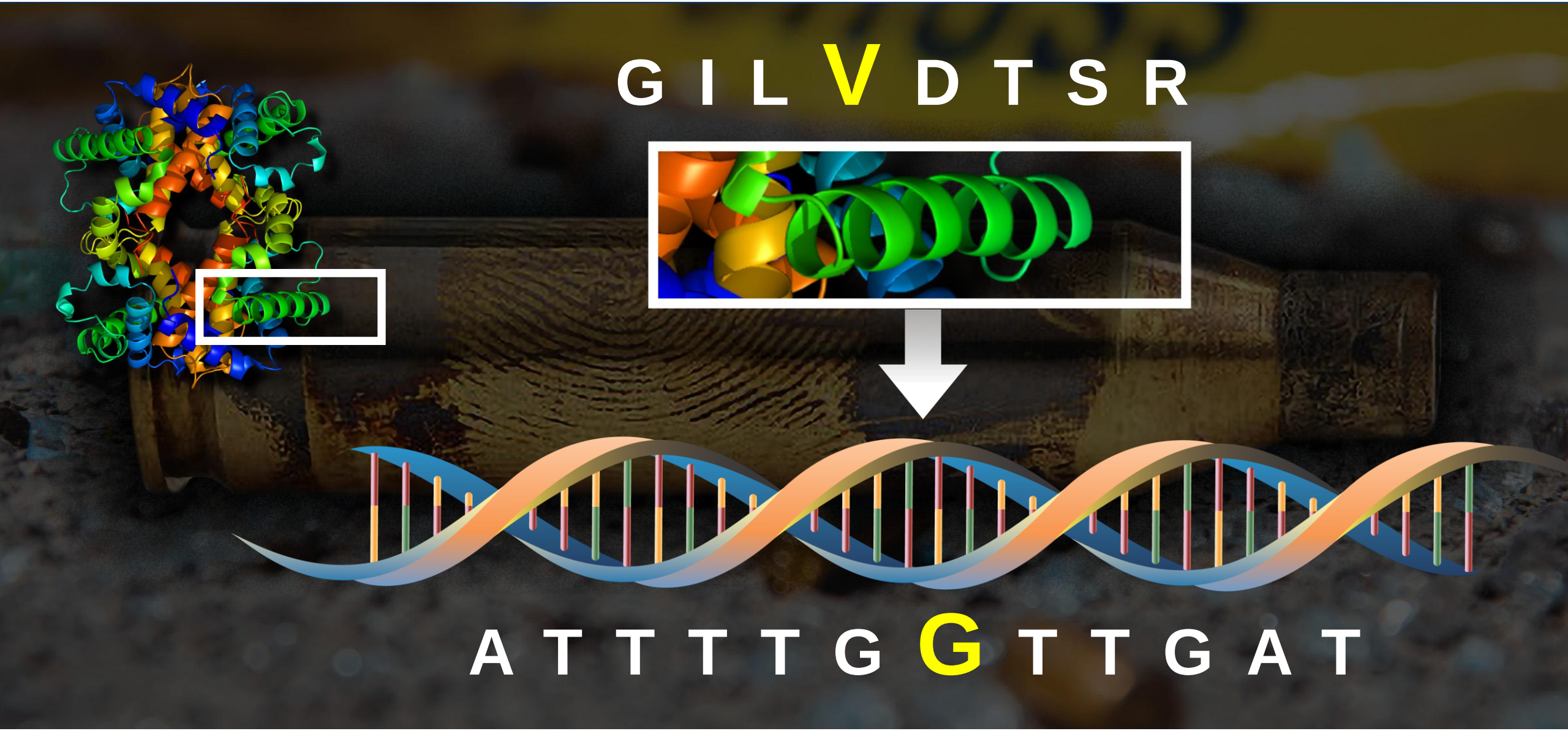
BE THE FUTURE

https://www.iarpa.gov/index.php/research-programs/proteos

Human Identification from Fingerprints

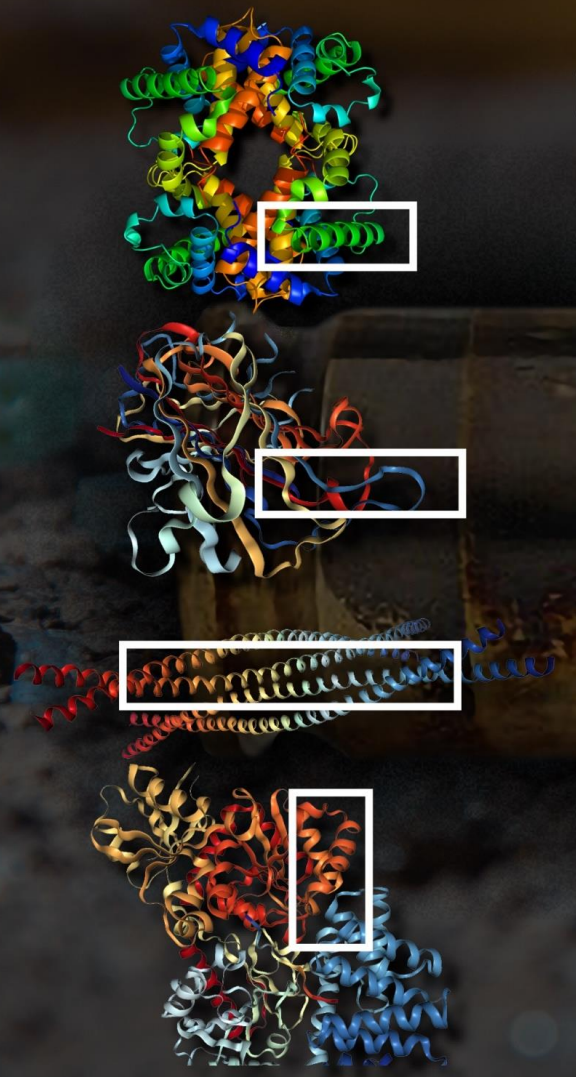


Human Identification from Fingerprints



Human Identification from Fingerprints

Human Identification from Fingerprints



Protein	Mutation	Mutation Frequency	Combined Probability
1	GILVDTSR	15.9%	1 in 6
+			
2	ATSYAELCNEYR	51.2%	1 in 12
+			
3	LNDLEDALQQSK	2.3%	1 in 534
+			
4	MGAQELLR	4.0%	1 in 13,000

Co-Extraction of Protein and DNA from Fingerprints

1

Visualize Fingerprint

2

Solubilize and Collect

Qiagen Buffer ATL (SDS-based tissue lysis buffer)

3

Bind DNA and Protein with Qiagen DNA Investigator Kit

No ProK!

4

Isolate DNA

DNA Analysis

(Standard Protocol)

5

Isolate Protein

Protein Analysis

Elute with S-Trap lysis buffer (contains TEAB, SDS) and DTT

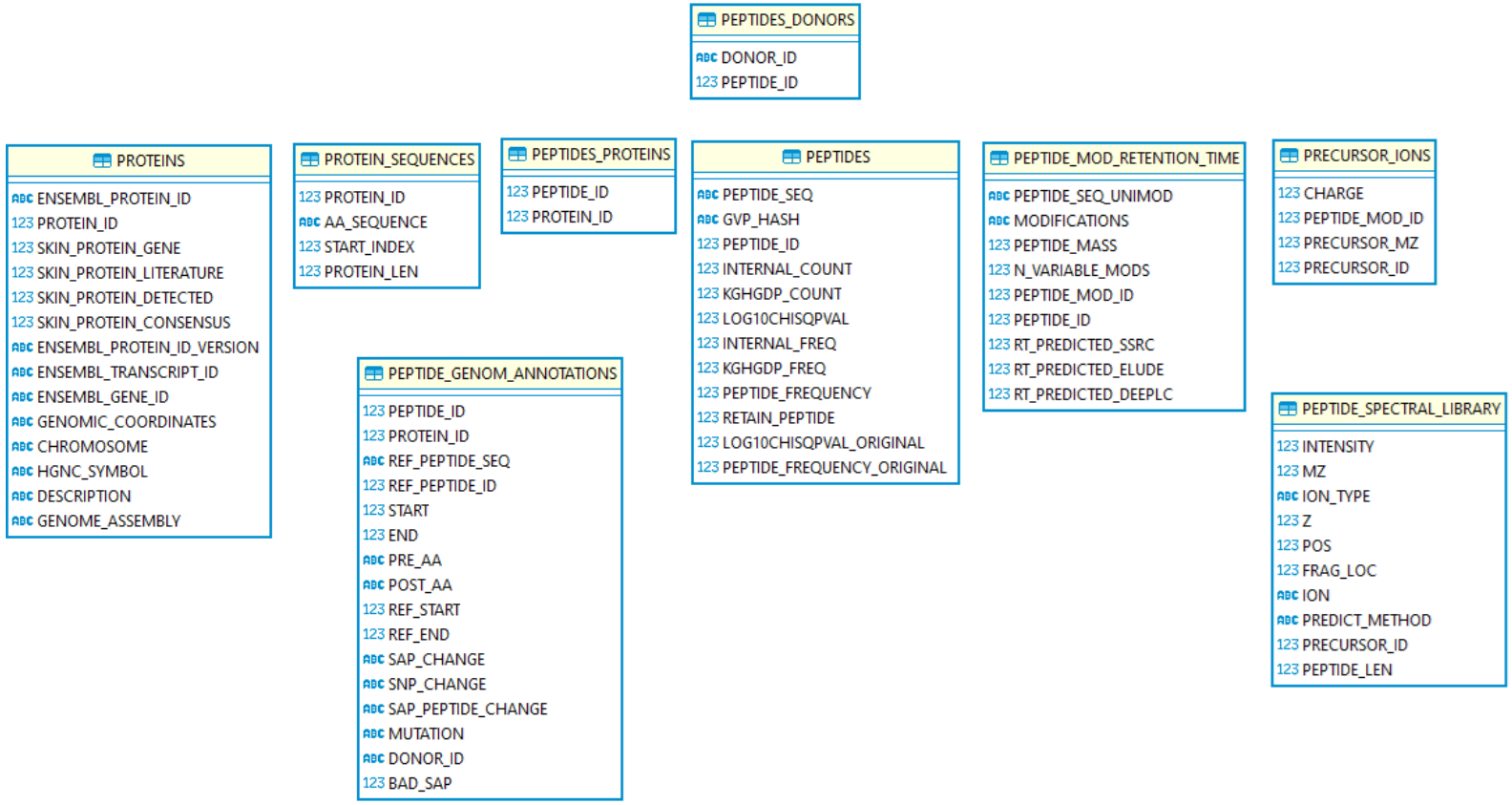
- Protein Preparation
  - DTT reduction
  - IAA alkylation
  - Transfer to S-Trap to remove SDS
  - Trypsin digestion overnight
  - Elute peptides from S-Trap
  - Speed-Vac to dryness
  - Reconstitute and analyze by LC-HRAM-MS2

Proteomic Analysis Method | LC-HRAM-MS2

- Nano-LC
  - **Thermo Scientific Ultimate 3000 RSLCNano**
  - Loading Column: Acclaim PepMap 100 C18, 75 µm × 20 mm
  - Nano Column: EasySpray C18, 2.0 µm, 75 µm × 250 mm
  - Mobile Phase A: 98% H_2O, 2% ACN, 0.5% FA
  - Mobile Phase B: 2% H_2O, 98% ACN, 0.5% FA
  - Flow Rate: 300 nL/min
  - **LC Run Time: 320 min (5 h gradient)**
- High Resolution, Accurate MS2
  - **Thermo Scientific Q Exactive Plus**
  - Full MS Resolution = 70,000
  - Full MS Scan Range = *m/z* 375–1,575
  - Data Dependent MS2 or PRM Resolution = 17,500
  - MS2 Isolation Width = 1.6 *m/z*
  - PRM Inclusion List Precursor Ions = 629 (including internal standards)

Peptide-Protein Database

- Exomes translated *in silico* to proteomes for 50+ donors
- Compiled database of reference and variant protein sequences
- Performed *in silico* tryptic digest
- Calculated peptide frequencies
- Predicted peptide RT and MS2
- SQL database contains
  - Protein sequences
  - Peptides (frequencies)
  - Protein-peptide associations
  - Peptide-donor associations (ground truth)
  - Modified peptides (predicted RT)
  - Predicted peptide MS2
  - Genomic annotations
- Output FASTA files from database

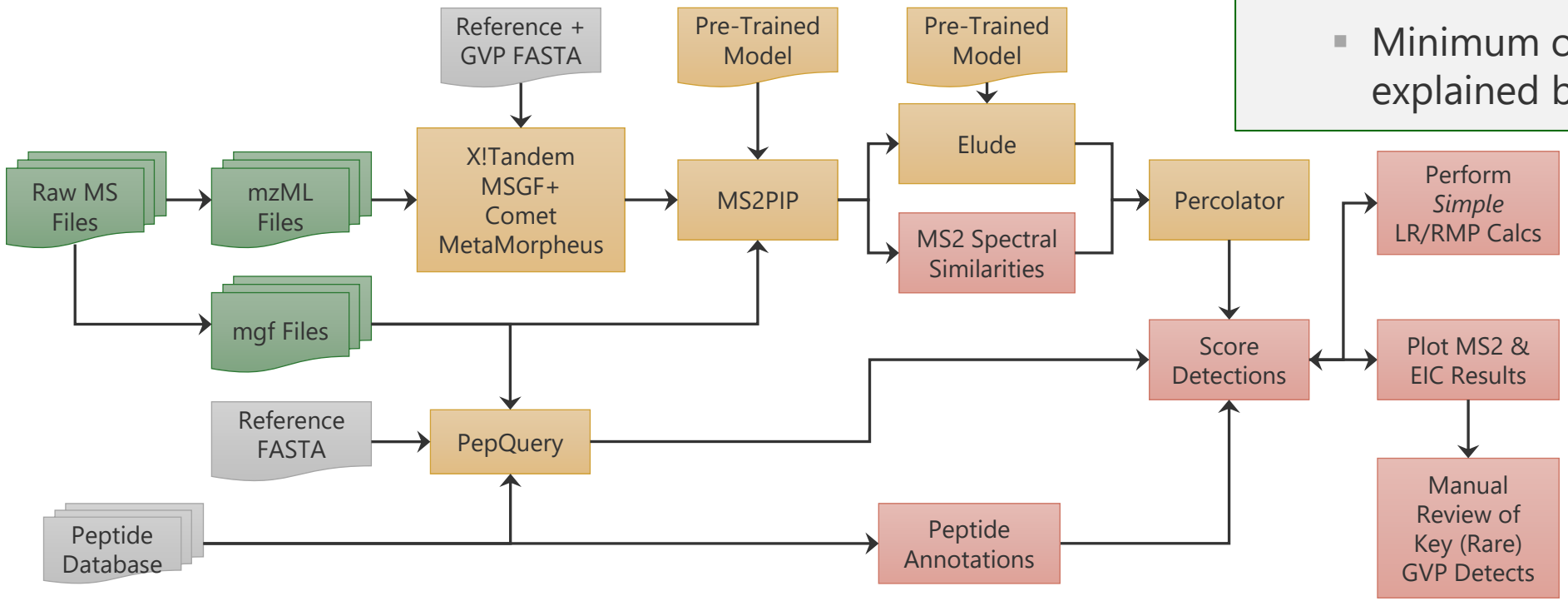


LC-HRAM-MS2 DDA Data Analysis Pipeline for Discovery

MetaMorpheus, MSGF+, X!Tandem, Comet + MS2PIP + Percolator

- In-house scripts written in R
- External Software Requirements (all open source)
  - X!Tandem
  - MS-GF+
  - MetaMorpheus
  - Comet
  - PepQuery
  - MS2PIP
  - Percolator
  - ProteoWizard
  - Elude

- Detection Criteria
  - Good MS2 spectral similarity to predicted MS2 (> 0.75)
  - Detection by minimum of 2 search engines
  - Minimum of 5 b- & y-ions
  - Minimum of 10% of ion current explained by b- & y-ions

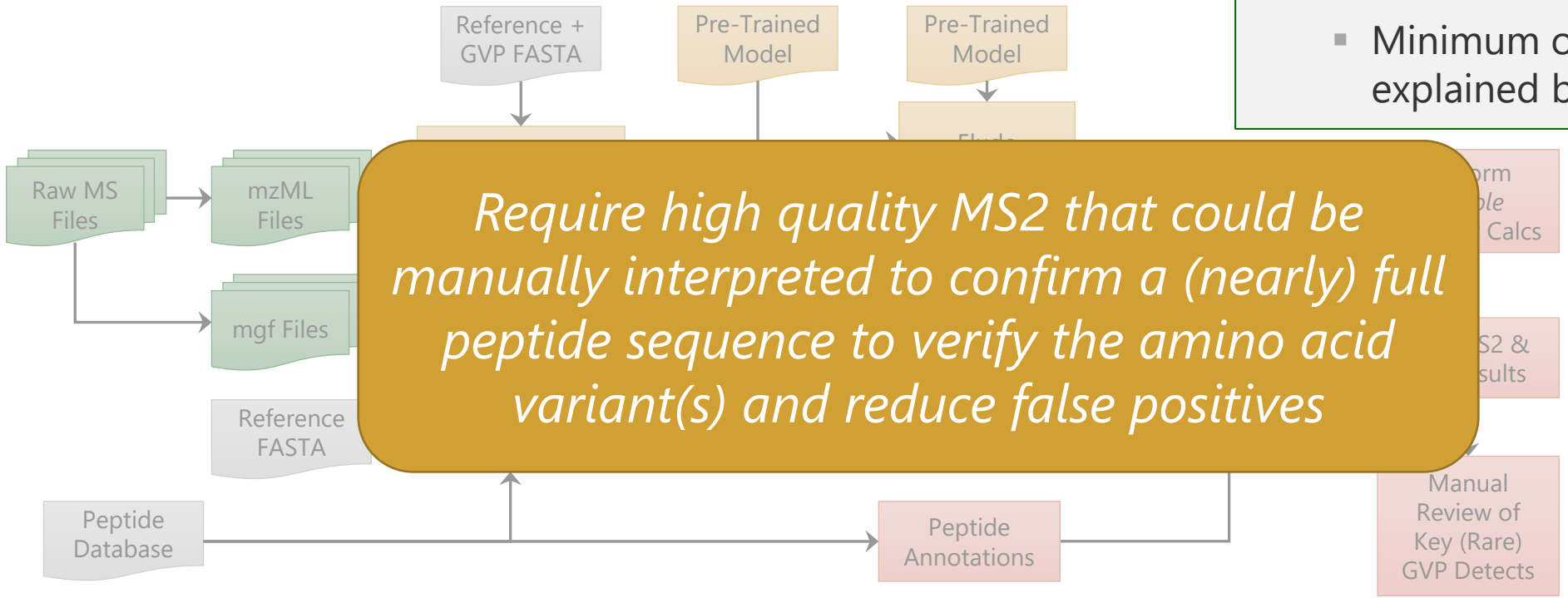


LC-HRAM-MS2 DDA Data Analysis Pipeline

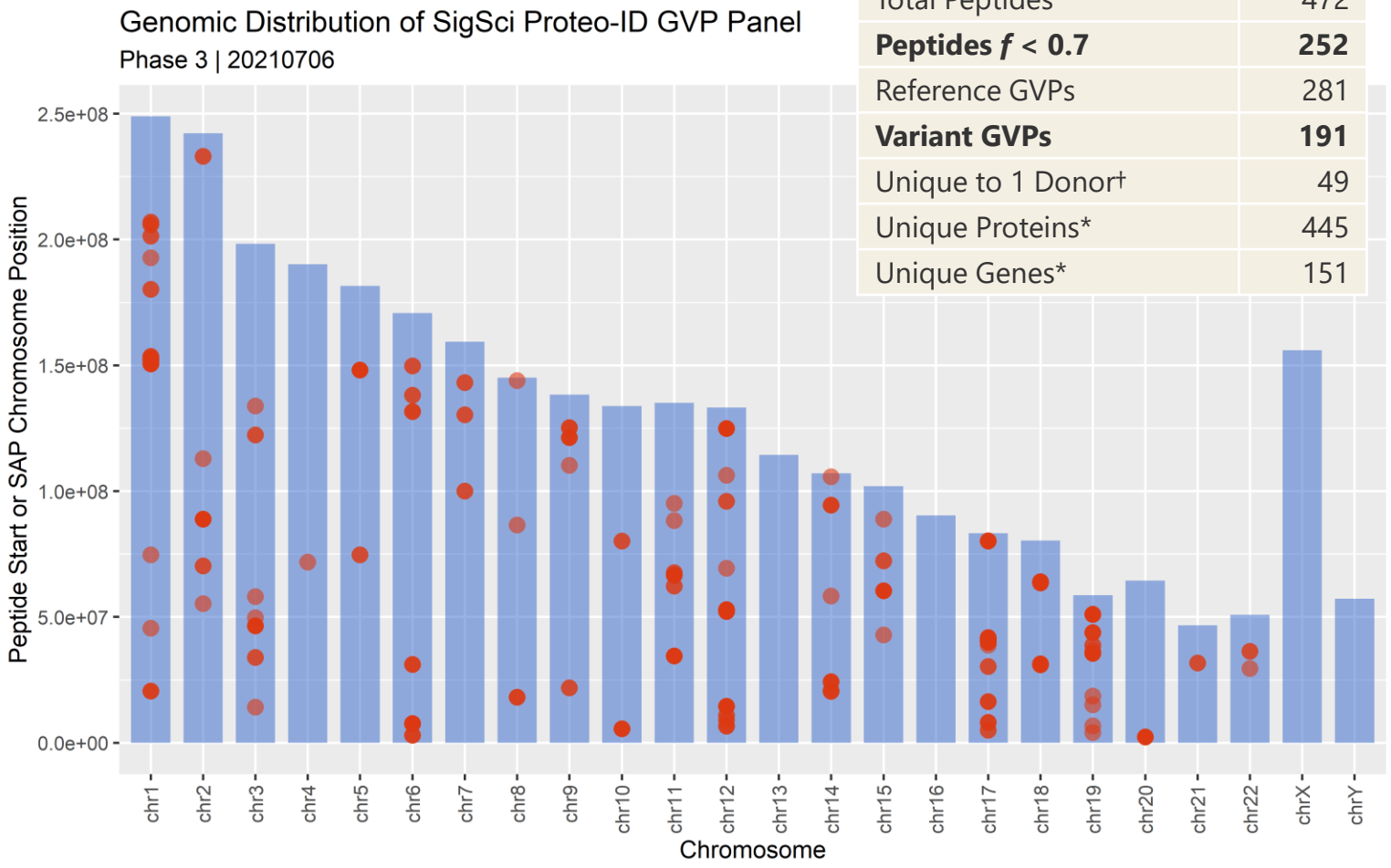
MetaMorpheus, MSGF+, X!Tandem, Comet + MS2PIP + Percolator

- In-house scripts written in R
- External Software Requirements (all open source)
  - X!Tandem
  - MS-GF+
  - MetaMorpheus
  - Comet
  - PepQuery
  - MS2PIP
  - Percolator
  - ProteoWizard
  - Elude

- Detection Criteria
  - Good MS2 spectral similarity to predicted MS2 (> 0.75)
  - Detection by minimum of 2 search engines
  - Minimum of 5 b- & y-ions
  - Minimum of 10% of ion current explained by b- & y-ions



Genetically Variable Peptide (GVP) Panel



Statistic	Value
Total Peptides	472
Peptides *f* < 0.7	**252**
Reference GVPs	281
Variant GVPs	**191**
Unique to 1 Donor†	49
Unique Proteins*	445
Unique Genes*	151

UniProt	Count*	Protein Name
Q86YZ3	18	Hornerine (expressed in epidermis)
P35908	16	Keratin, type II cytoskeletal 2 epidermal
Q8N1N4	15	Keratin, type II cytoskeletal 78
P15924	15	Desmoplakin (DP)
P02533	13	Keratin, type I cytoskeletal 14
Q5D862	13	Filaggrin-2 (FLG-2)
P02538	12	Keratin, type II cytoskeletal 6A
P20930	12	Filaggrin
Q5T749	12	Keratinocyte proline-rich protein
P14923	12	Junction plakoglobin (Desmoplakin-3)
P04264	11	Keratin, type II cytoskeletal 1

Common SAPs	Count
I > V	14
A > T	13
E > K	13
G > S	12
R > Q	11
T > M	10
V > I	10
N > S	10

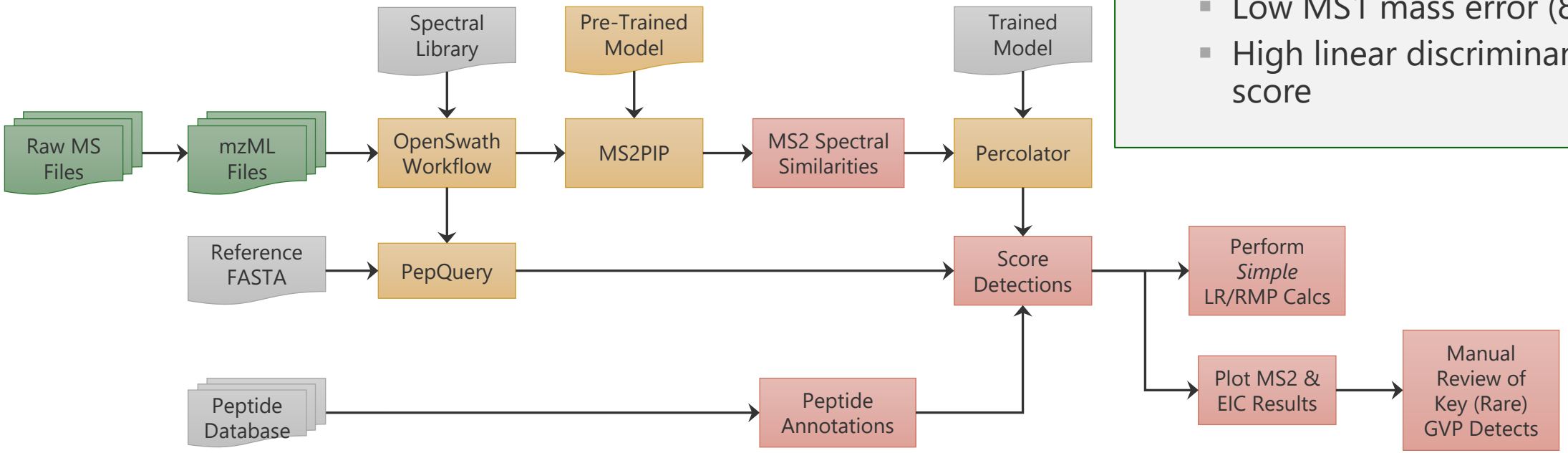
* Includes any skin protein or gene associated with the peptide (i.e., a peptide may be associated with multiple proteins and/or genes), or total peptides that are associated with the protein
† Within a set of 52 total donors

LC-HRAM-MS2 PRM Data Analysis Pipeline for Targeting

OpenSwathWorkflow + MS2PIP + Percolator

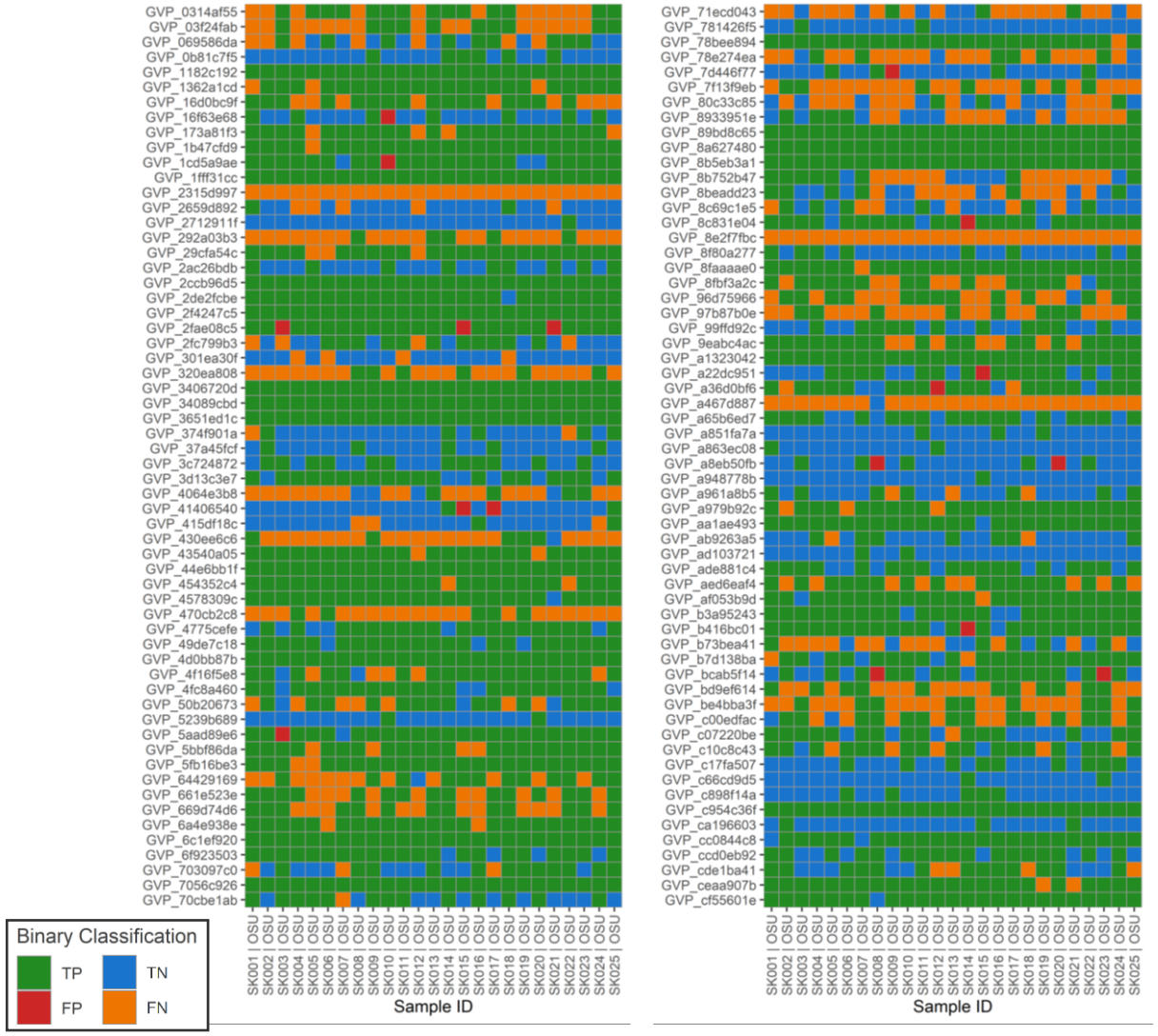
- In-house scripts written in R
- External Software Requirements (all open source)
  - OpenMS
  - PyProphet
  - MS2PIP
  - PepQuery
  - Percolator
  - ProteoWizard

- Detection Criteria
  - Good MS2 spectral similarity to predicted MS2
  - Detect at least 4 (of 6 selected) MS2 transitions
  - High MS1 isotope correlation
  - Low MS1 mass error (8 ppm)
  - High linear discriminant model score

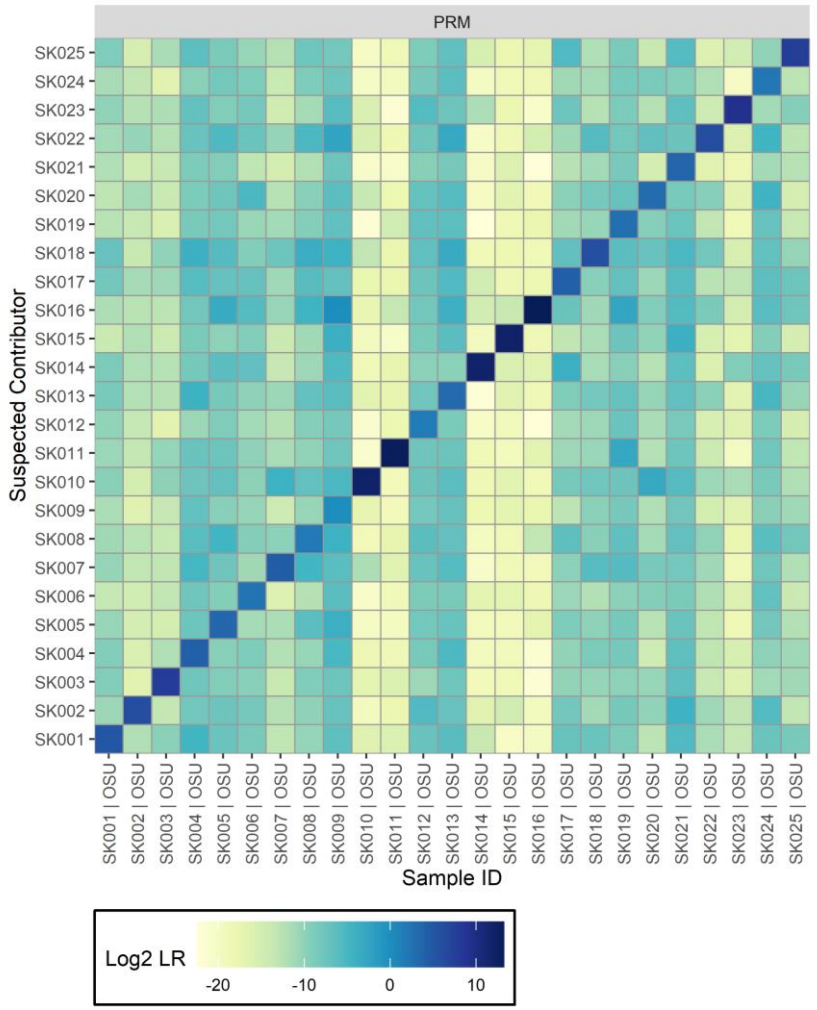


Results | PRM | Bulk Skin Samples Using Common GVPs

Common GVP Detection Results

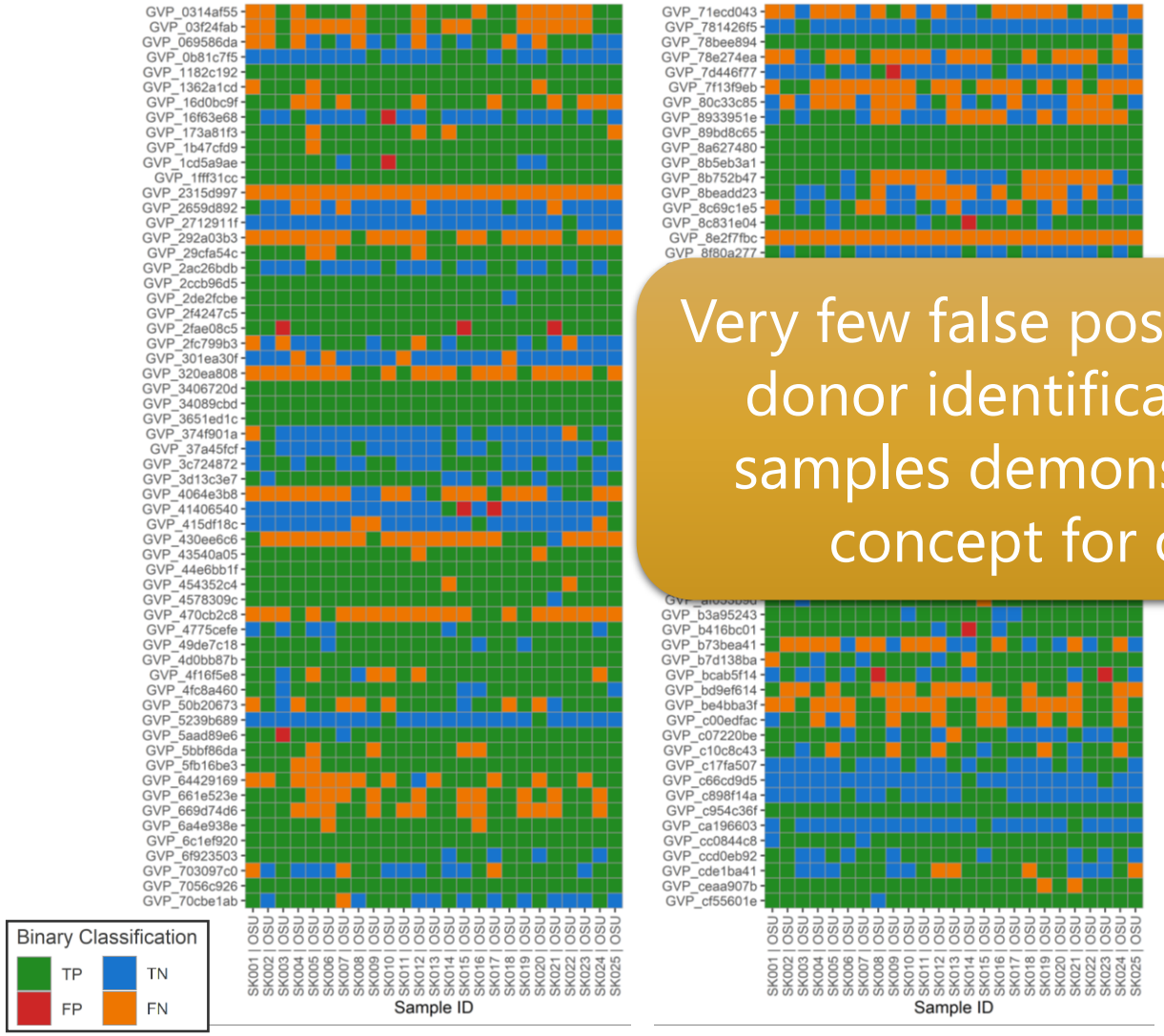


Likelihood Ratio | Suspect vs. Known Donor



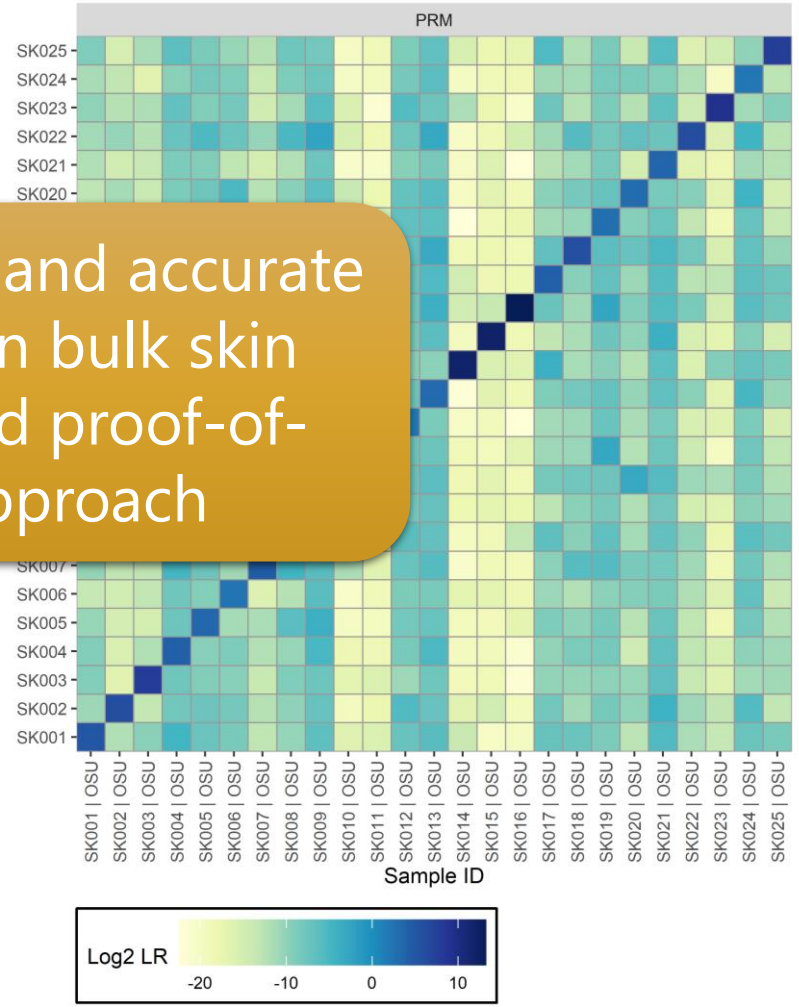
Results | PRM | Bulk Skin Samples Using Common GVPs

Common GVP Detection Results



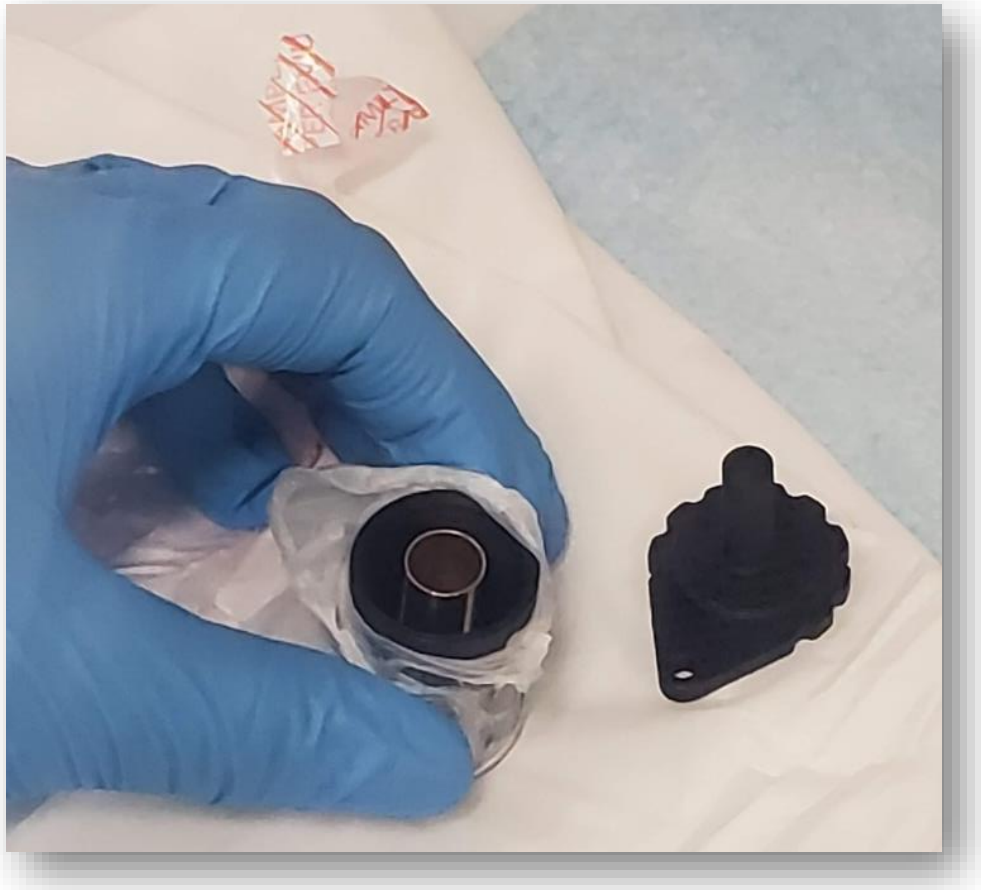
Very few false positives and accurate donor identification in bulk skin samples demonstrated proof-of-concept for our approach

Likelihood Ratio | Suspect vs. Known Donor



Unfired Brass Shell Casing | Case Study

- 9mm Brass Shell Casings (Blinded):
  - Three replicates
  - Question: Who is/are the contributor(s)?



Unfired Brass Shell Casing | 3 Contributor Mixture | PRM

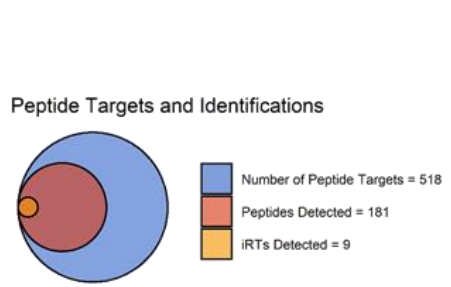
20210618_Proteos_QE_028 | PR01,PR02,PR14 | commonRareAllObservedLowMz

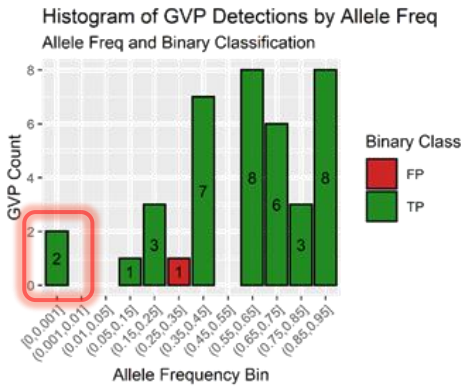
Total Number of Peptides Detected: 181 / 518
Total Number of iRT Peptides Detected: 9 / 11
RMP Estimate (All Peptides): 3.082e-11 (173)

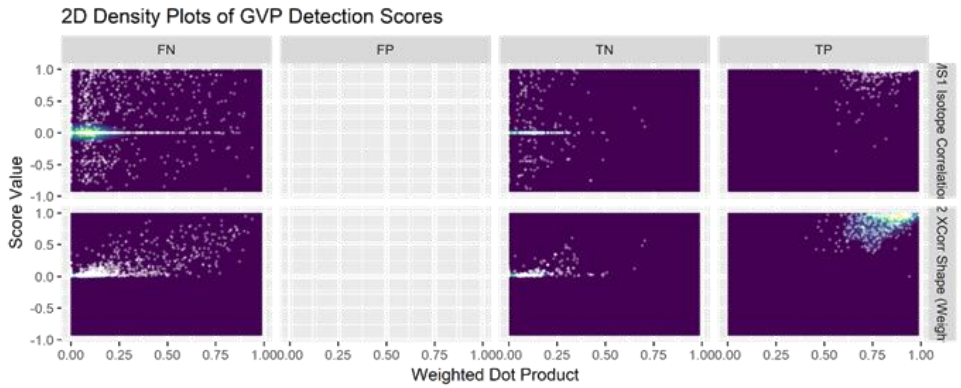
Mean Abs Ret. Time Diff. (sec): 117.9
RMSE Ret. Time Diff (sec): 244.9

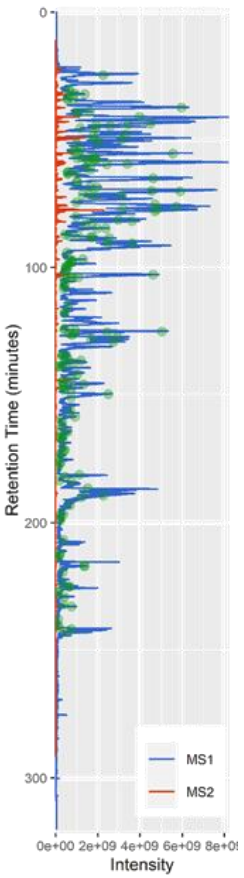
AlleleFreq SS LR | Most Likely Donor(s): PR01
AlleleFreq SS LR | Log2 LR Most Likely Donor(s): 3.672e+00
AlleleFreq SS LR | Log2 LR PR01,PR02,PR14: 6.420e+00

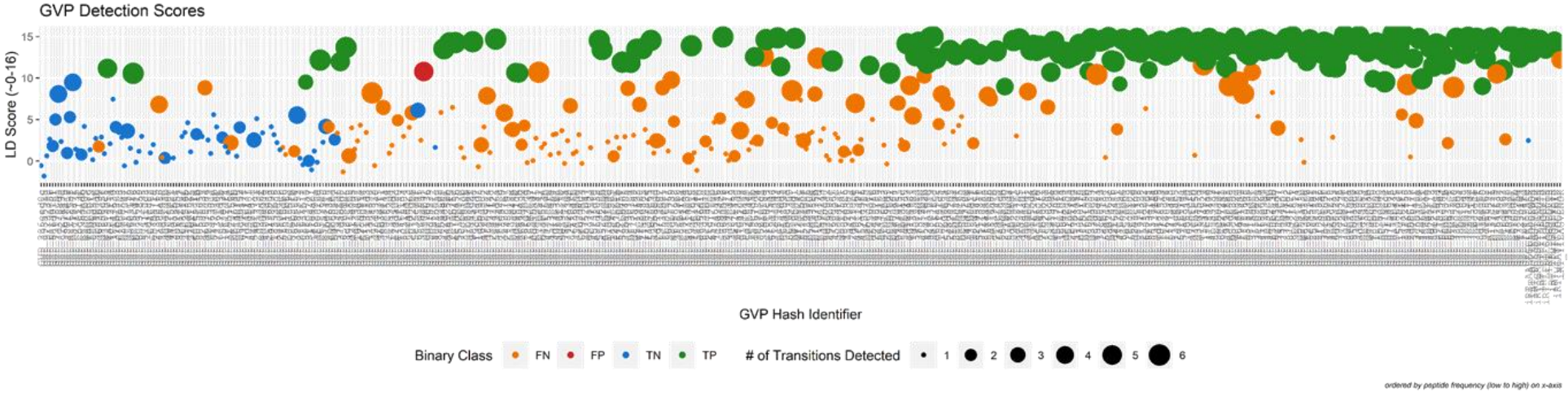
AlleleFreq Mixture LR | Most Likely Donor(s): PR01;PR14
AlleleFreq Mixture LR | Log2 LR Most Likely Donor(s): 6.420e+00











Sample A205 = 0.200
Sample A280 = 0.122

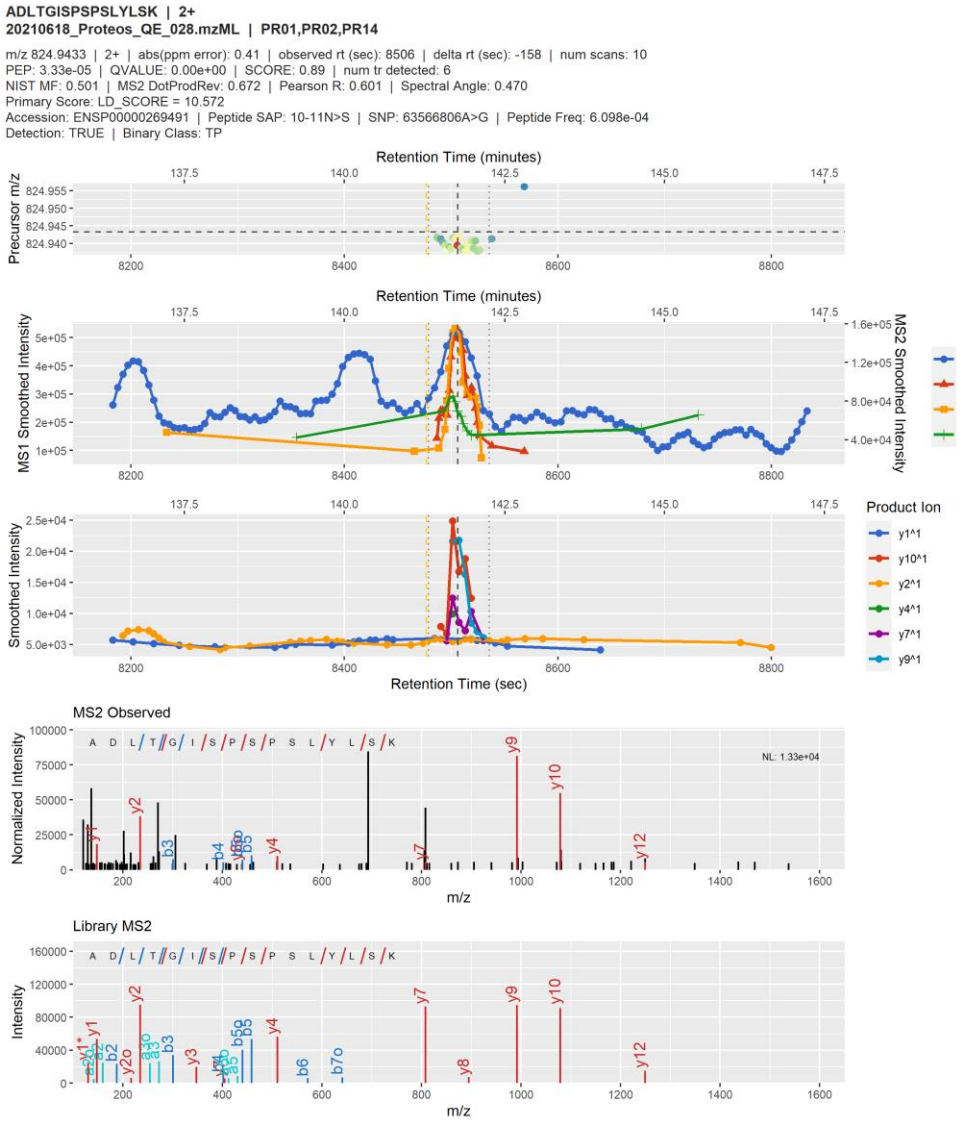
MS1 TIC (RT50) = 3.73e+12
MS2 TIC (RT50) = 3.70e+10

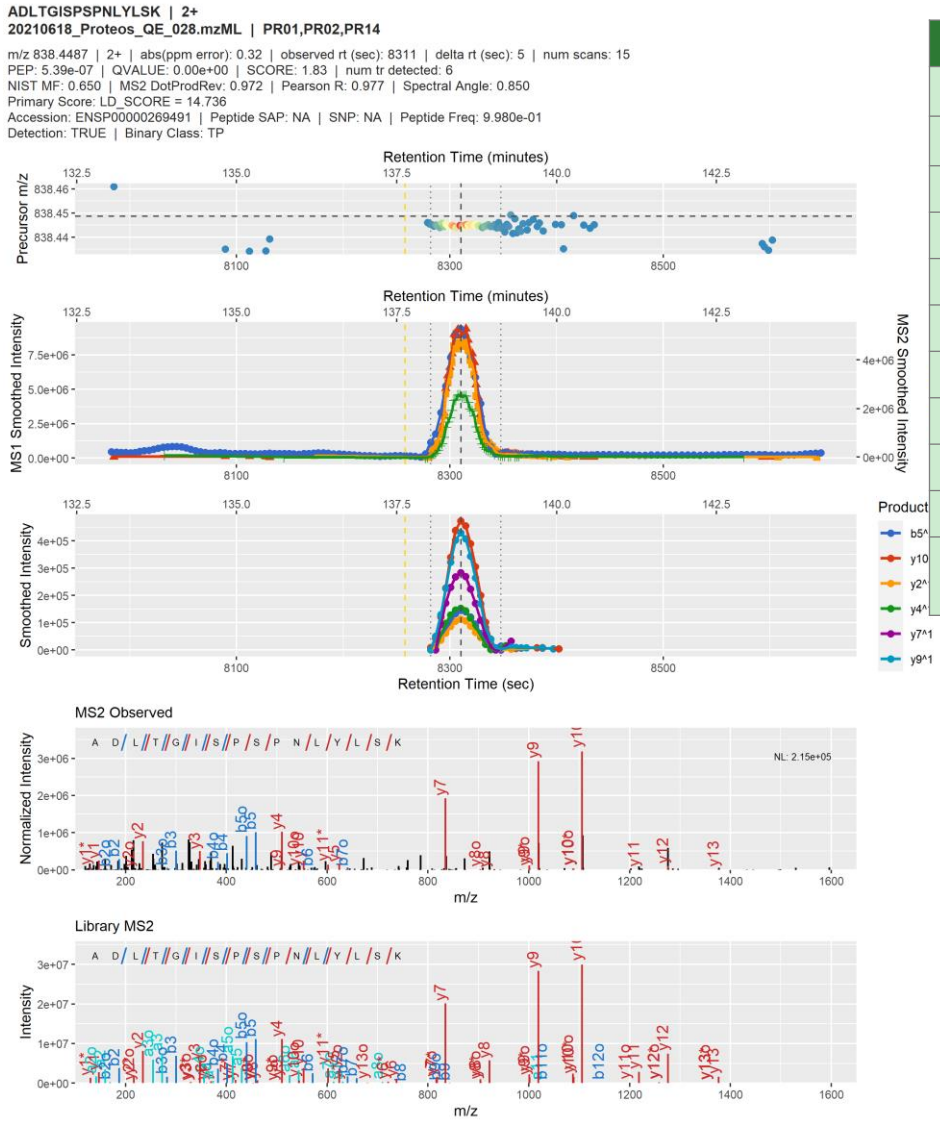
MS1 Scans (RT50) = 5,048
MS2 Scans (RT50) = 100,976

MS1 Median S/N = 504.41
MS2 Median S/N = 1.54
MS2 Median Peak Count = 10

MS1 Median IT = 9.79 ms
MS2 Median IT = 50.00 ms

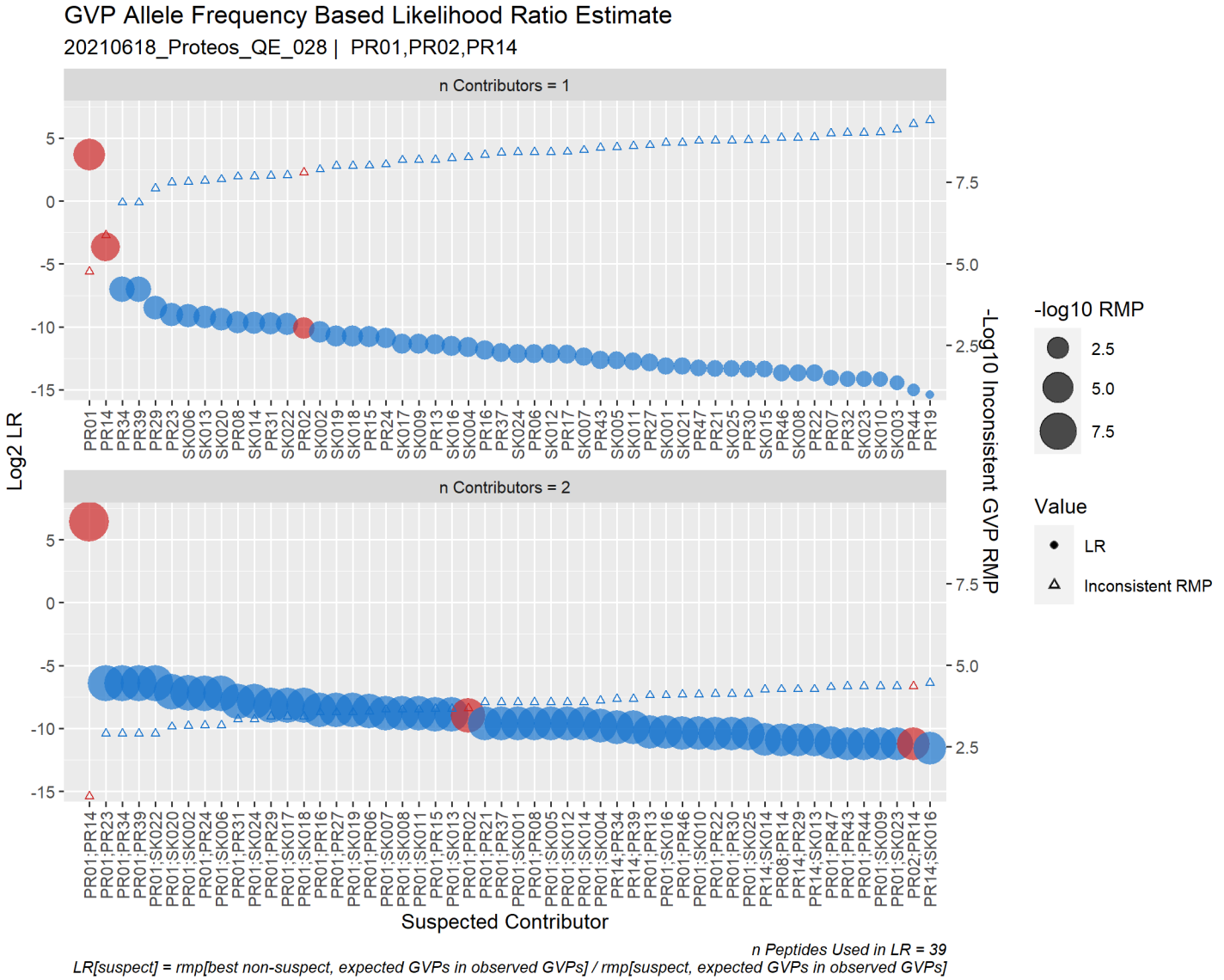
Unfired Brass Shell Casing | Rare/Reference GVPs | PRM





Variable	Value
Peptide Sequence	ADLTGISPSPSLYLSK
Reference Peptide	ADLTGISPSPNLYLSK
SAP (Peptide Coords)	10-11N>S
SNP	63566806A>G
Peptide Frequency	6.098E-04
Internal Count	1 (n = 52)
KGHGDP Count	4 (n = 6,560)
Chromosome	chr18
Protein	ENSP00000269491.1
Transcript	ENST00000269491.5
Gene	ENSG00000166634.6 (SERPINB12)

Unfired Brass Shell Casing | 3 Contributor Mixture | PRM

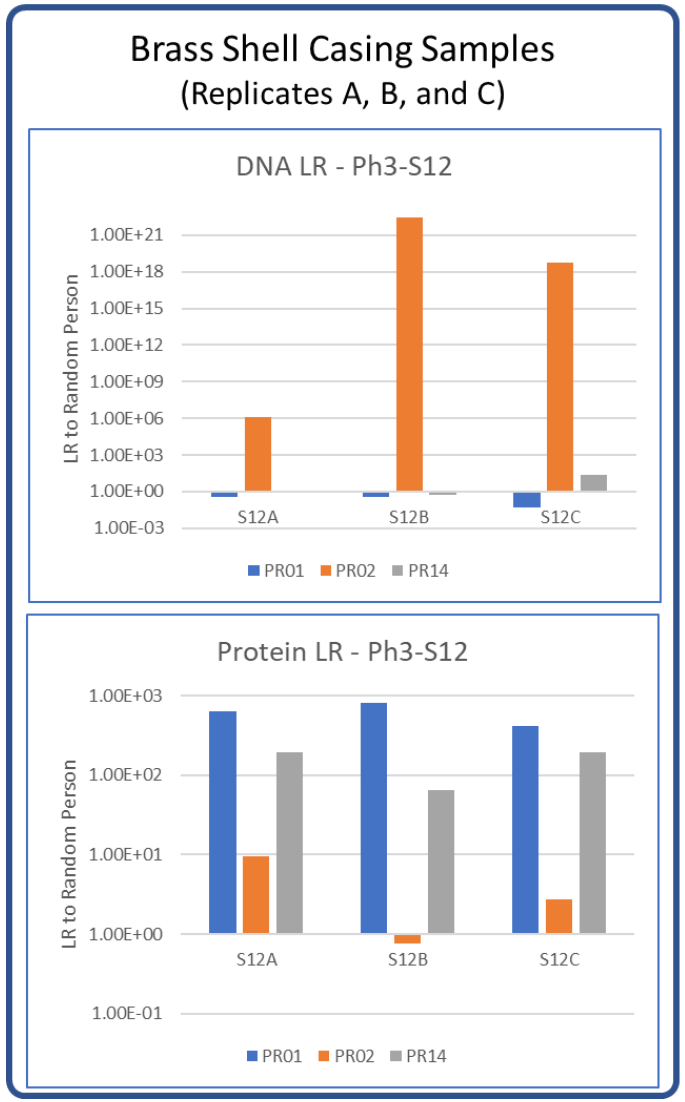


- Top 2 single-source contributors are known contributors
- Only able to identify PR01 and PR14 as all detected GVPs are consistent with this 2-person mixture
- PR01 appears to be the primary contributor (by low frequency GVP detections)

Unfired Brass Shell Casing | Case Study

- 9mm Brass Shell Casings (Blinded)
  - Three replicates
  - Question: Who is/are the contributor(s)?
- Contradictory Results
  - Protein
    - PR01 = major contributor
    - PR14 = minor contributor
  - DNA
    - PR02 = major contributor
    - Probable mixture, but insufficient data for comparison to any reference profiles

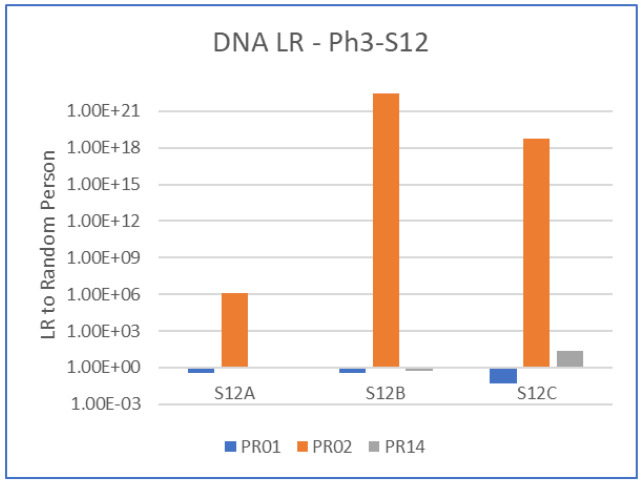
Unfired Brass Shell Casing | DNA vs. Protein

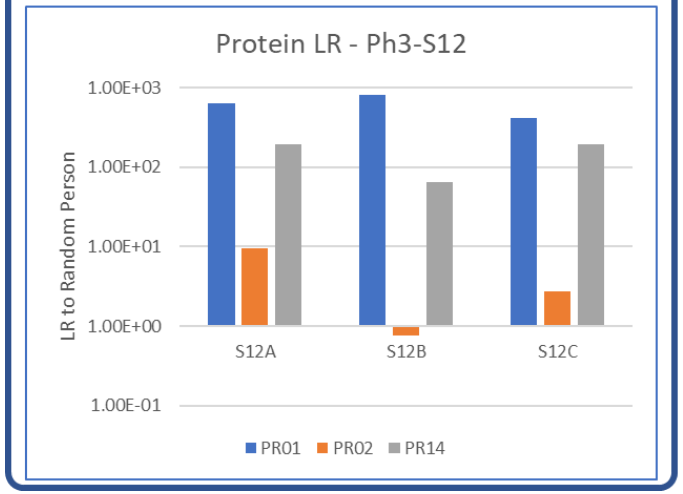


- Ground Truth:
  - Samples provided were mixtures
  - Three contributors: PR01, PR02, and PR14
- Contradictory results were complementary
- Observed similar trends for other 3-contributor mixtures on keyboard keys and elevator buttons

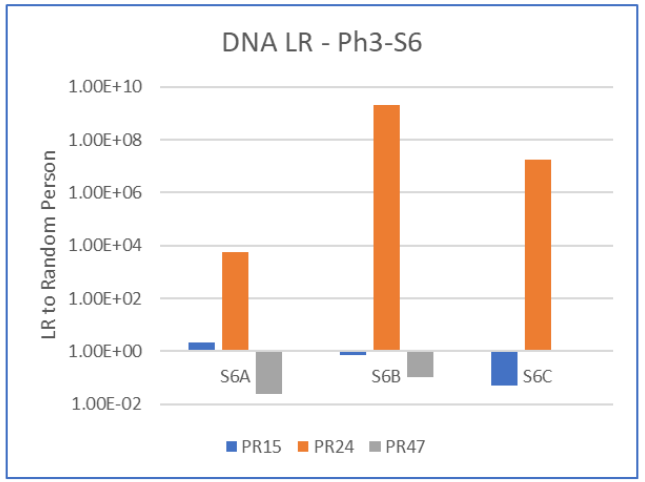
DNA vs. Protein Likelihoods for Mixed Contributor Samples

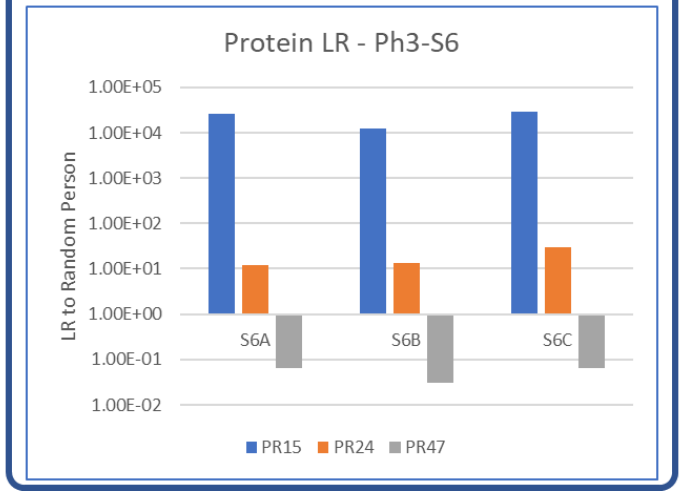
Brass Shell Casing Samples (Replicates A, B, and C)



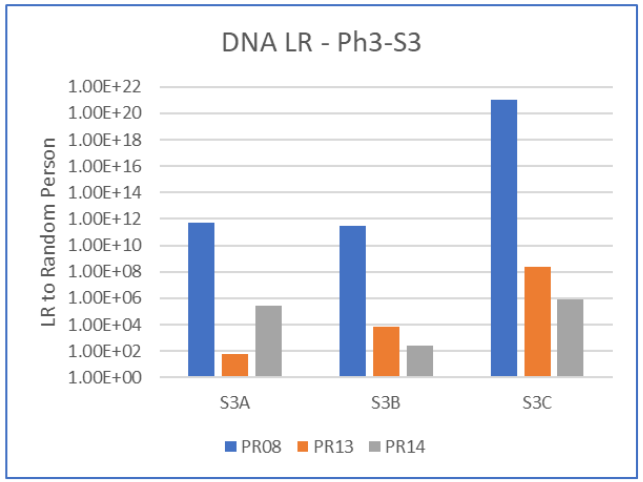


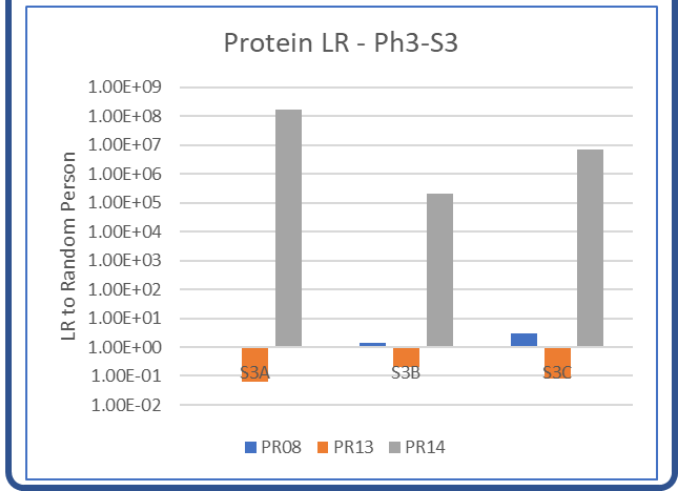
Keyboard Key Samples (Replicates A, B, and C)





Elevator Button Samples (Replicates A, B, and C)



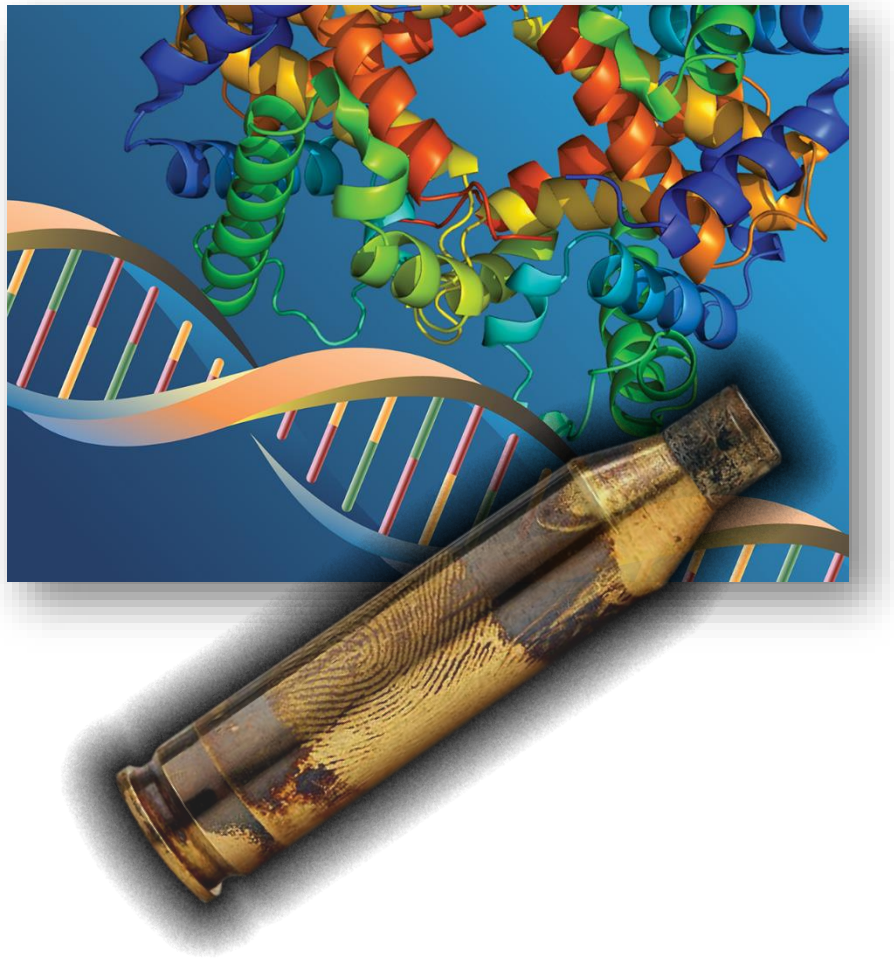


Current Challenges

- Sample collection and extraction methods not in complete alignment with most forensic labs (cannot use ProK!)
- Instrumentation and subject matter experts
- No equivalent to CODIS for protein markers
  - Protein profiles can be compared to each other or a known exome/whole genome sequence
- Minimum protein input requirements (~2 µg)
- Limited GVP panel size and detectability can limit likelihood ratio values
  - Even harder considering that some alleles may not be expressed or detectable
  - May be further complicated by kinship
- Rare GVPs
  - Extremely discriminating or just a false positive?
  - How do you validate a marker you have never encountered before?

Future Directions

- Collection and extraction method optimization
- Method transfer to operational labs and independent laboratory evaluation from beginning to end
- Validation on relevant sample types
- Collection of additional data on sample matrices with known contributor(s)
- Database development (structure, format, hosting, accessibility, content)



- *More Human ID by Proteomics ASMS Presentations*
  - Hair and skin sample prep. optimization – NIST (ThP 291)
  - Proteomics for genotyping to estimate ancestry – G. Parker (ThP 092)
  - Skin variant and reference peptide spectral library from Proteos data – NIST (ThP 075)
  - Human ID from touch samples using DIA LC-MS/MS – Univ. of Washington & Spectragen (FP 226)

Acknowledgements

Project Co-PI
Myles Gardner, Ph.D.

Project Co-PI
Curt Hewitt, Ph.D.

Michael Freitas, Ph.D.

August Woerner, Ph.D.

- Signature Science
  - Danielle LeSassier, Ph.D.
  - Alan Smith
  - Megan Powals
  - Kathleen Schulte, M.S.
  - David Joiner, M.S.
  - Benjamin Ludolph
- The Ohio State University
  - Andrew Reed, Ph.D.
  - Liwen Zhang, Ph.D.
  - Maryam Baniasad
- University of North Texas
  - Ben Crysup, Ph.D.
- Lawrence Livermore National Laboratory (LLNL) – Forensic Science Center (Test & Evaluation Team)
- National Institute of Standards and Technology (NIST) (T&E Team)
- Intelligence Advanced Research Projects Activity (IARPA)

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2018-18041000003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.