Forensic Genetics

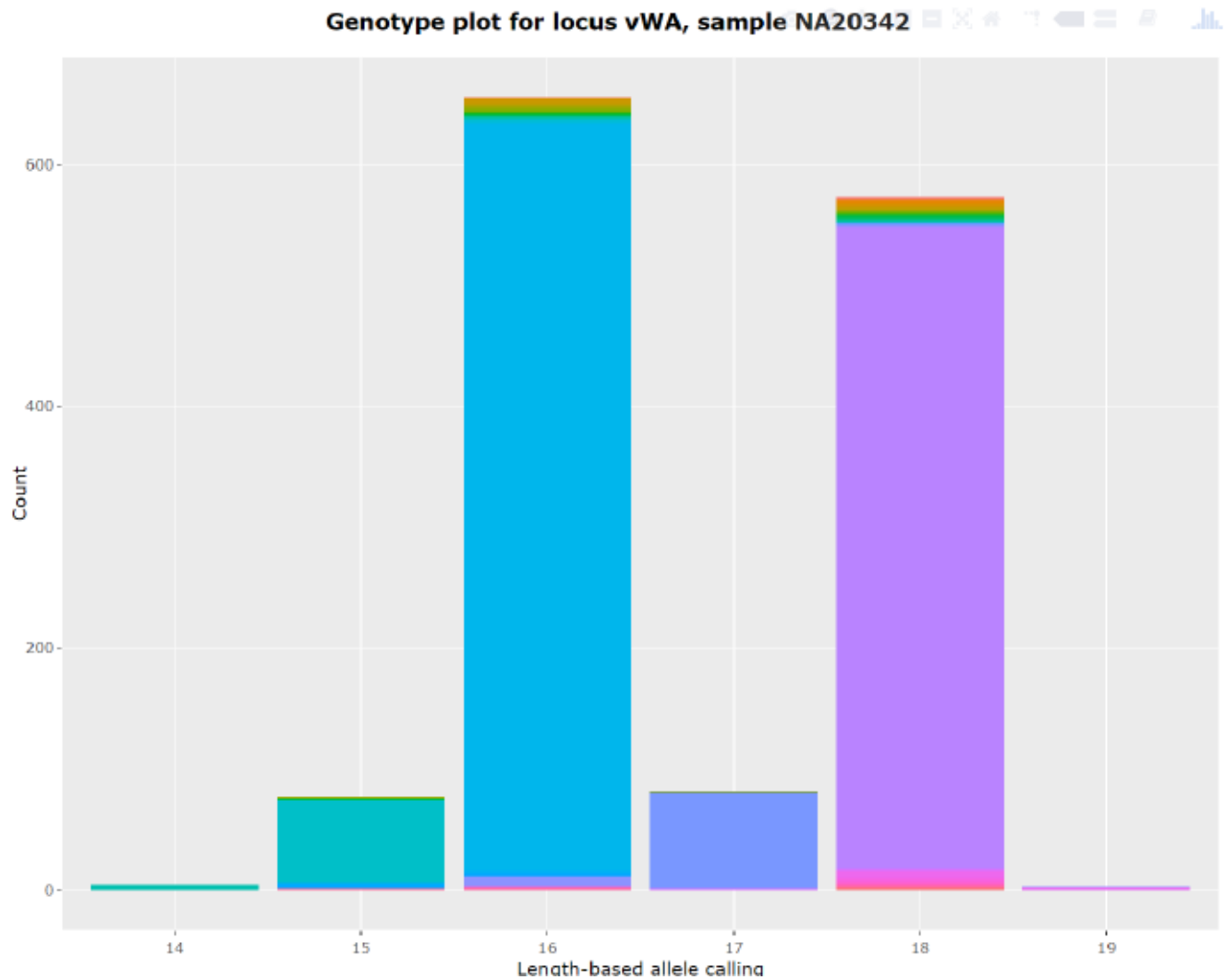
Module 16 – Session 10

Other Techniques

- NGS Data
- NGS Modeling
  - Stutter
  - Population Structure
- Other
  - Duplex Sequencing
  - Microhaplotypes
  - Record Linkage
  - Protein-Based Human Identification
  - Inference of Ancestry
  - Inference of Phenotype
  - Microbial Forensics
  - Rapid DNA

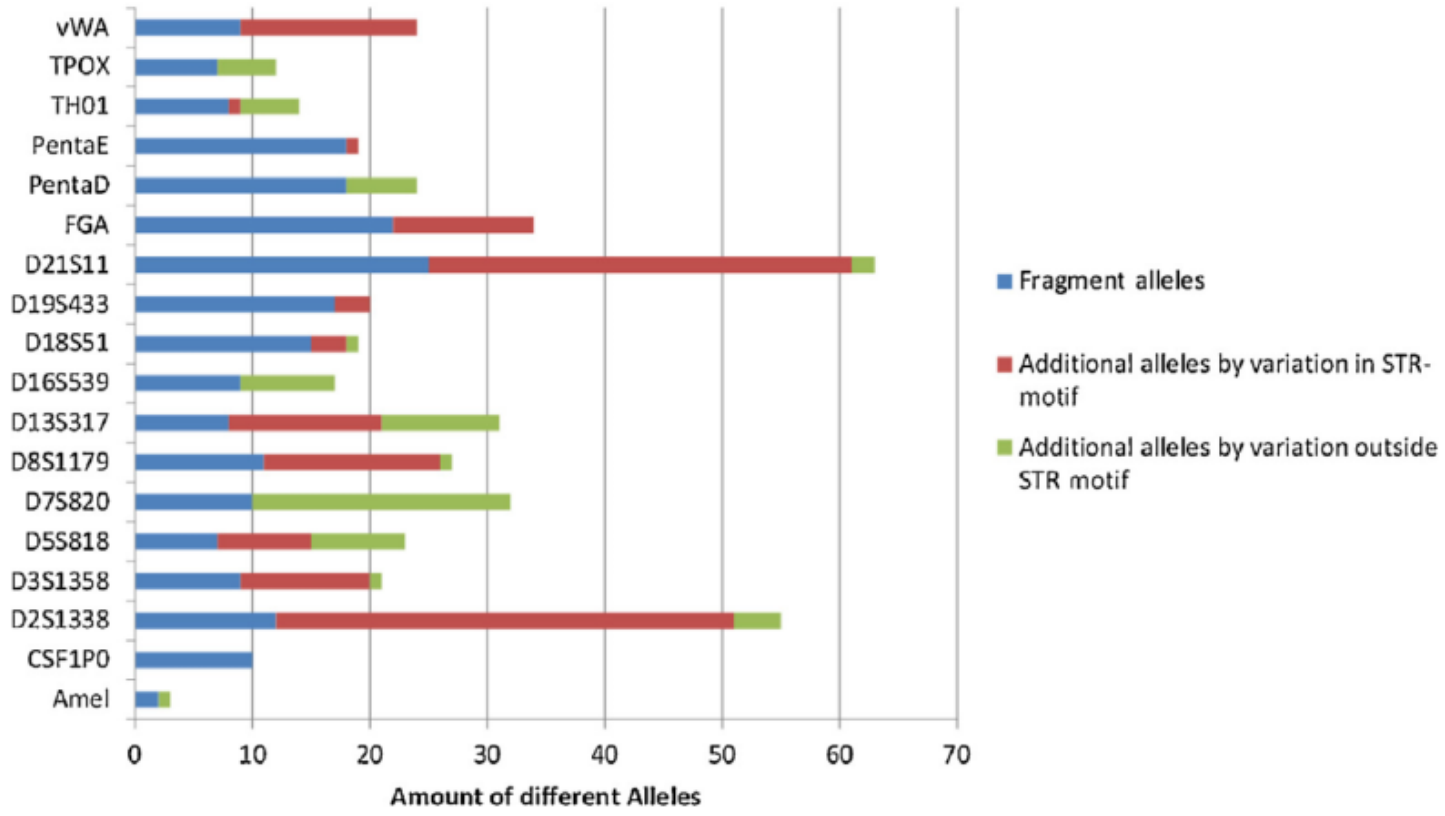
NGS Data

A DNA profile can be visualized similar to an epg:



NGS Data

STR sequence variation divided in length variation, additional sequence variation, and SNP variation:



Source: Massively parallel sequencing of short tandem repeats (van der Gaag et al., 2016).

NGS Stutter Modeling

Recall the definition of the stutter ratio:

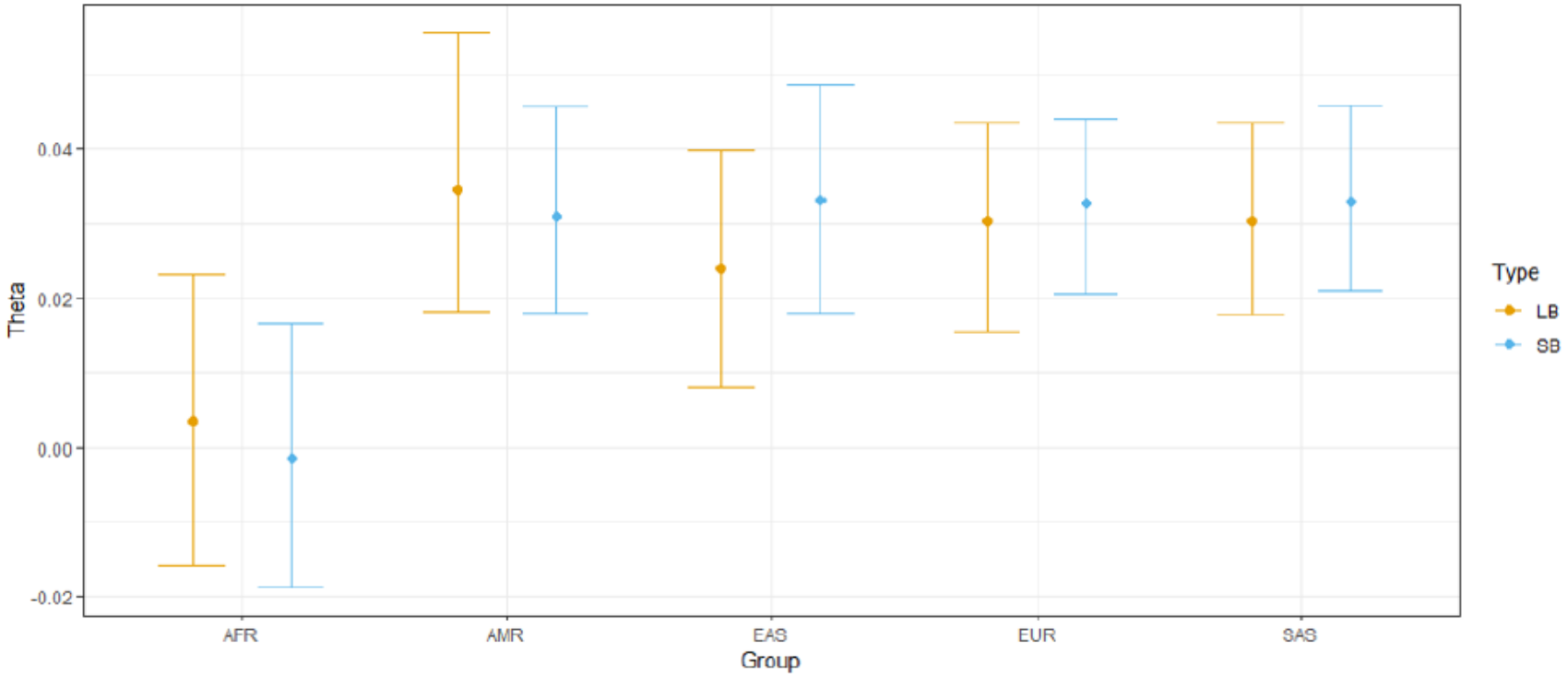
$$SR = \frac{O_{a-1}}{O_a} = \frac{O_A + O_B}{O_a} = \frac{O_A}{O_a} + \frac{O_B}{O_a}$$

Instead of modeling stutter per parental allele, you can also model the ratios per different stutter sequence. This was not possible for CE data.

Category	Allele	Sequence	Count	SR
Allele	9.3	$[AATG]_6ATG[AATG]_3$	100	0.25
Stutter	8.3	$[AATG]_6ATG[AATG]_2$	5	0.05
Stutter	8.3	$[AATG]_5ATG[AATG]_3$	20	0.20

NGS Population Structure

Results show very similar effects of sequencing data on theta estimates as what we have seen for CE-based results.



Confidence intervals per group.

Source: Analyzing population structure for forensic STR markers in next generation sequencing data (Aalbers & Weir, 2020)

NGS Applications

Judge Rules Against Novel DNA Test In One Twin's Rape Case

April 18, 2017 | By WBUR Newsroom

DNA

Case Study: First Criminal Conviction from Next-Gen DNA in Holland

Thu, 06/13/2019 - 1:07pm 1 Comment by Seth Augenstein , Senior Science Writer - @SethAugenstein

Duplex Sequencing

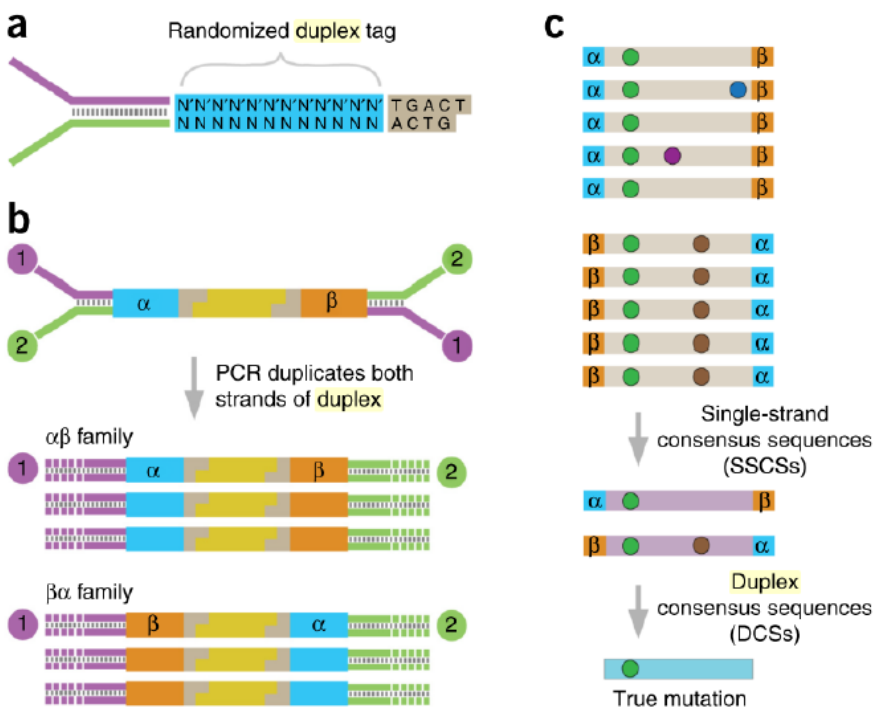
Most NGS approaches have a relatively high error rate and are therefore not suitable for detecting in vivo mutations. To overcome this limitation, a highly sensitive sequencing methodology termed *Duplex Sequencing (DS)* has been developed.

- DNA fragments get labeled with their own unique tag;
- After PCR amplification, each group yields one consensus sequence;
- Two complementary consensus sequences, derived from the same fragment, are then compared to yield a 'duplex consensus sequence'.

Source: Detecting ultralow-frequency mutations by Duplex Sequencing (Kennedy et al., 2014).

Duplex Sequencing

Only true mutations will appear in both duplex sequences, while PCR-related artifacts will be eliminated when establishing the final consensus sequence.



Source: Detecting ultralow-frequency mutations by Duplex Sequencing (Kennedy et al., 2014).

Microhaplotypes

Instead of looking at individual SNPs, it has been suggested that combining multiple SNPs into a microhap that renders highly informative for forensic purposes.

Although microhaps are more sensitive, the absence of stutter yields an increase in potential for mixture deconvolution. SNPs are also shown to be correlated with physical phenotypic traits, information the STRs cannot provide.

To make the use of microhaps feasible for forensic purposes, however, backward compatibility is required with CE data. This might be established through record linkage, based on STR inference from SNP data.

Source: Criteria for selecting microhaplotypes: mixture detection and deconvolution (Kidd & Speed, 2015).

Record Linkage

Instead of looking for a (partial) match in one database, it is also possible to combine different databases, even with no overlapping genetic markers. Provided that sufficiently strong LD exists, SNP and STR profiles can be associated with the same individual or distinct but closely related individuals.

Software can be used to infer STR genotypes from a SNP dataset, making it possible to compute match scores for pairs of individuals between databases. This means that CODIS profiles can possibly be connected to a SNP profile, collected for e.g. biomedical or genealogical research, and this cross-database record matching extends to relatives.

Linkage disequilibrium connects genetic records of relatives typed with disjoint genomic marker sets (Rosenberg et al., 2018).

Protein-Based Human Identification

Whereas DNA is prone to degradation, protein is chemically more robust and can persist for longer periods.

Protein contains genetic variation in the form of single amino acid polymorphisms (SAPs), resulting in a genetically variant peptide (GVP), which can be used to infer SNP profiles, regardless of the presence of DNA template in the sample.

Protein-based methodologies therefore have the potential to provide a complementary and, if necessary, alternative method for use in forensic practice in cases where DNA is absent or not sufficiently informative.

Source: Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome (Parker et al., 2016).

Protein-Based Human Identification

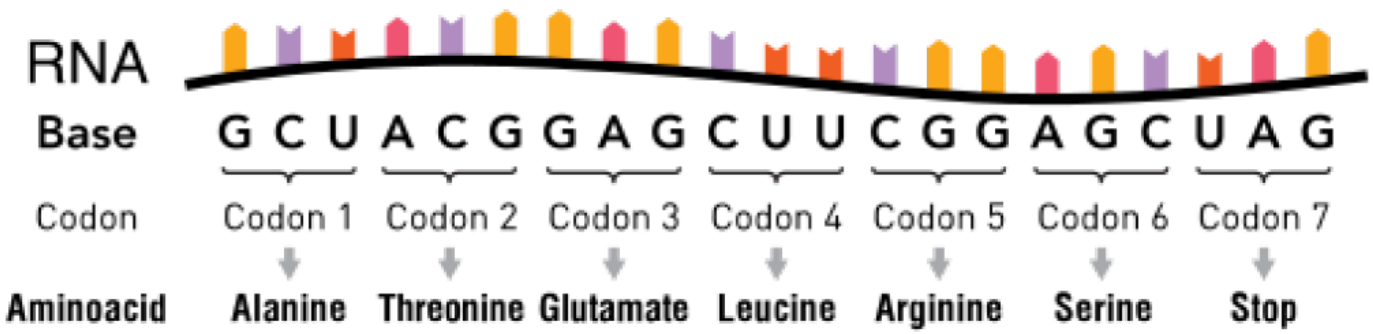
Certain sections of DNA, called *exons*, are coded for a *protein*, i.e. a macro-molecule consisting of one or more long chains of amino acid residues performing a vast array of functions within organisms. Two steps are required to read the information encoded in a gene's DNA and produce the protein it specifies:

- **Transcription**: produces nucleotide sequences complementary to the DNA from which it is transcribed, known as *messenger RNA* (mRNA);

- **Translation**: is the process by which a mRNA molecule is used as a template for synthesizing a new protein.

Protein-Based Human Identification

During translation, the genetic code is read three nucleotides at a time, in units called *codons*, which correspond to an *amino acid*.



Source: `https://en.wikipedia.org/wiki/Gene`

Since there are 64 possible codons (four possible nucleotides at each of the three positions) and only 20 standard amino acids, multiple codons can specify the same amino acid.

Protein-Based Human Identification

Amino Acid	Codes		Codons
Alanine	Ala	A	GCT, GCC, GCA, GCG
Cysteine	Cys	C	TGT, TGC
Aspartic acid	Asp	D	GAT, GAC
Glutamic acid	Glu	E	GAA, GAG
Phenylalanine	Phe	F	TTT, TTC
Glycine	Gly	G	GGT, GGC, GGA, GGG
Histidine	His	H	CAT, CAC
Isoleucine	Ile	I	ATT, ATC, ATA
Lysine	Lys	K	AAA, AAG
Leucine	Leu	L	CTT, CTC, CTA, CTG, TTA, TTG
Methionine (start)	Met	M	ATG
Asparagine	Asn	N	AAT, AAC
Proline	Pro	P	CCT, CCC, CCA, CCG
Glutamine	Gln	Q	CAA, CAG
Arginine	Arg	R	CGT, CGC, CGA, CGG, AGA, AGG
Serine	Ser	S	TCT, TCC, TCA, TCG, AGT, AGC
Threonine	Thr	T	ACT, ACC, ACA, ACG
Valine	Val	V	GTT, GTC, GTA, GTG
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAT, TAC
Stop codons	–	–	TAA, TAG, TGA

Protein-Based Human Identification

It is well-known that human variation is caused by mutations (during DNA replication), leading to polymorphism, i.e. the presence of multiple different alleles in a gene. Most variants are functionally equivalent, although some can give rise to differences, e.g. in phenotypic traits.

Mutations in coding regions compromise less than 2% of all genetic variation, and can be divided into two types:

- Synonymous mutations
- Nonsynonymous mutations

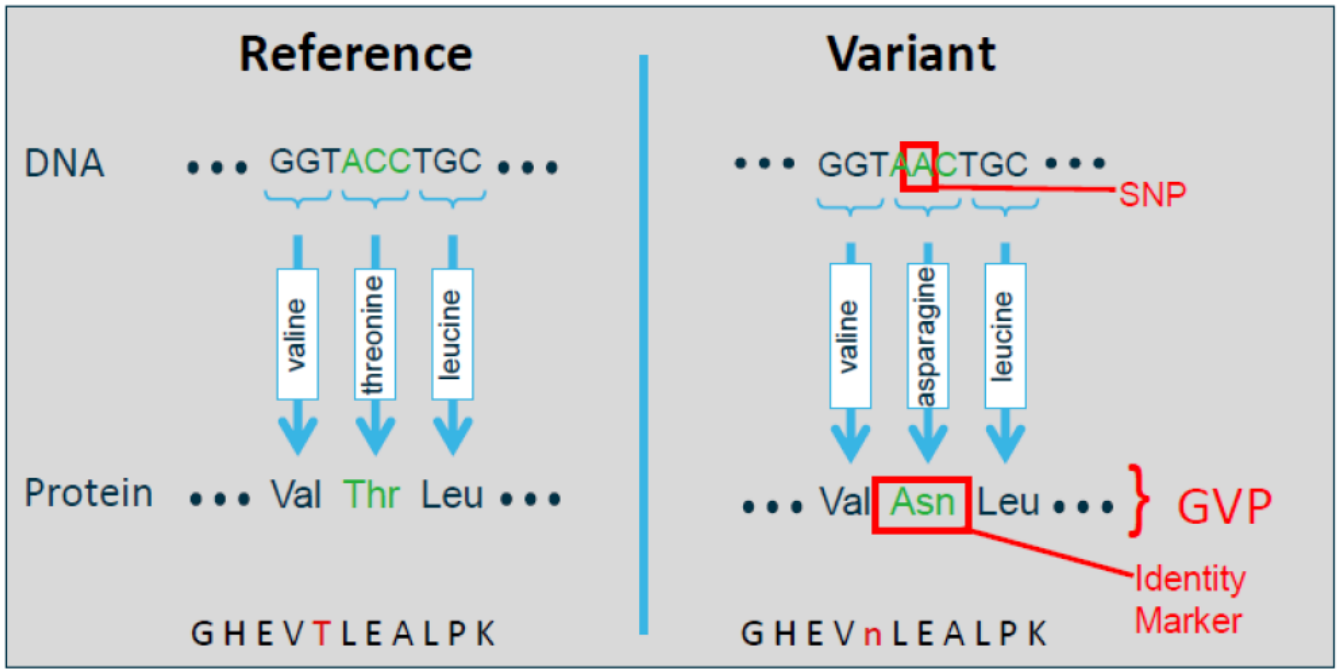
Protein-Based Human Identification

Synonymous mutations: Around 30% of mutations do not change the amino acid sequence, as a result of multiple codons encoding the same amino acid. A *silent* mutation does not affect the individual's fitness, whereas non-neutral changes involve sub-optimal synonyms (i.e. codons that translate less efficiently).

Nonsynonymous mutations: A mutation may also lead to an alteration of the amino acid sequence of the protein, with 10% resulting in *nonsense* mutations (e.g. a premature stop codon and consequently nonfunctional protein product). The remaining 60% are *missense* mutations and are of most relevance to this program.

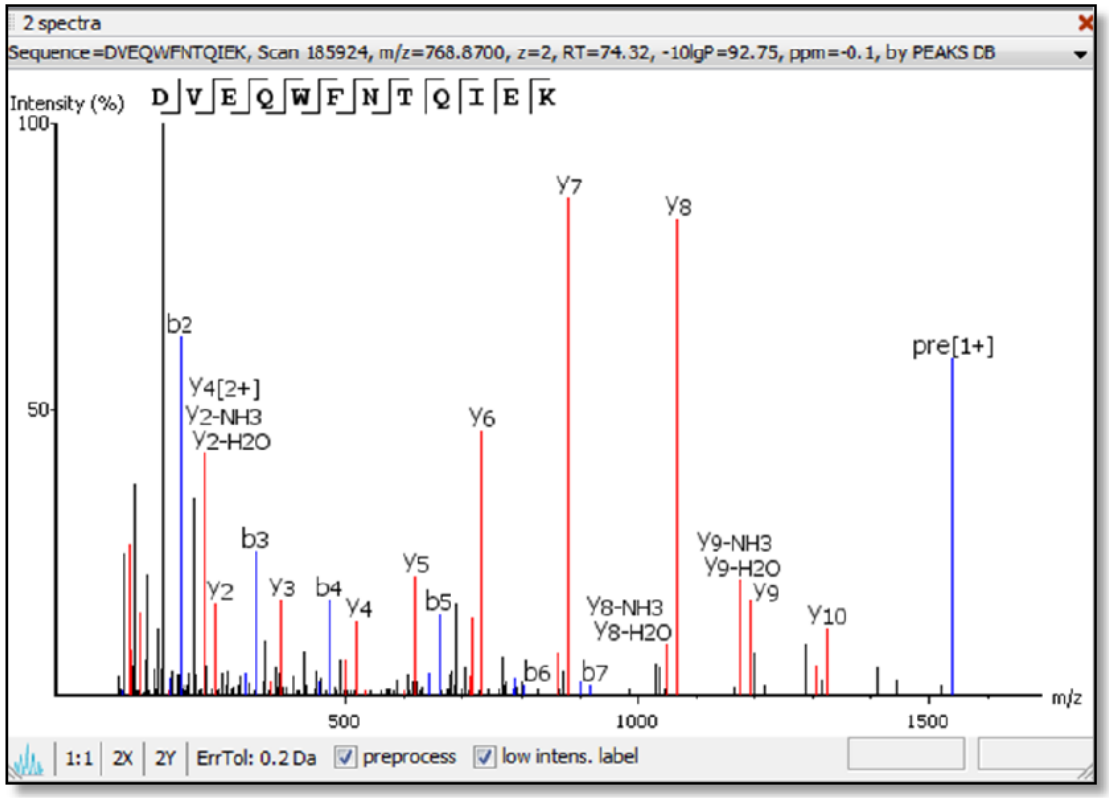
Protein-Based Human Identification

When a mutation involves the substitution of one base for another, it is called a single nucleotide polymorphism (*SNP*). A nonsynonymous SNP (*nsSNP*) leads to an altered amino acid, called a single amino acid polymorphism (*SAP*), which in turn results in a *peptide* (i.e. a relatively short amino acid chain, smaller than proteins) containing a SAP, a so-called genetically variant peptide (*GVP*).



Protein-Based Human Identification

Proteomic data sets can be obtained by analyzing samples via liquid chromatography mass spectrometry (LC/MS), resulting in a peptide fragment spectrum.

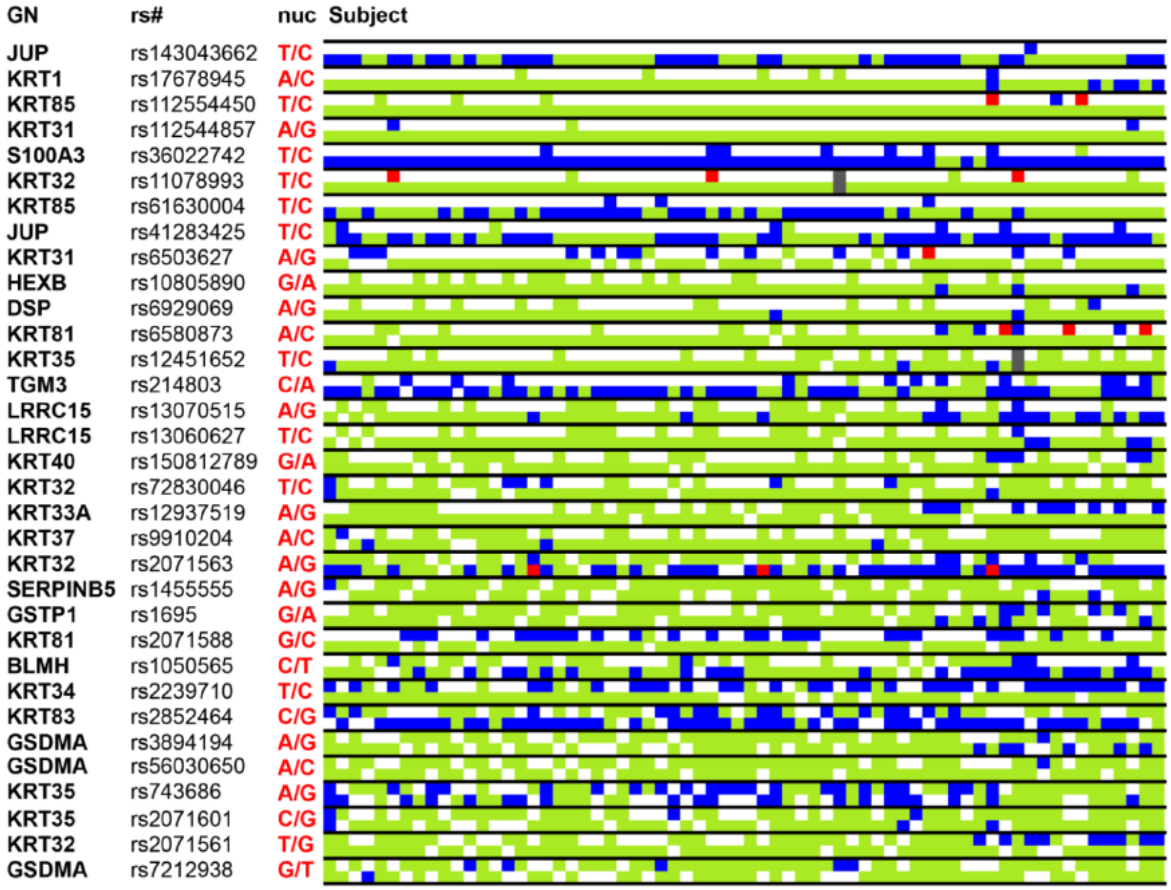


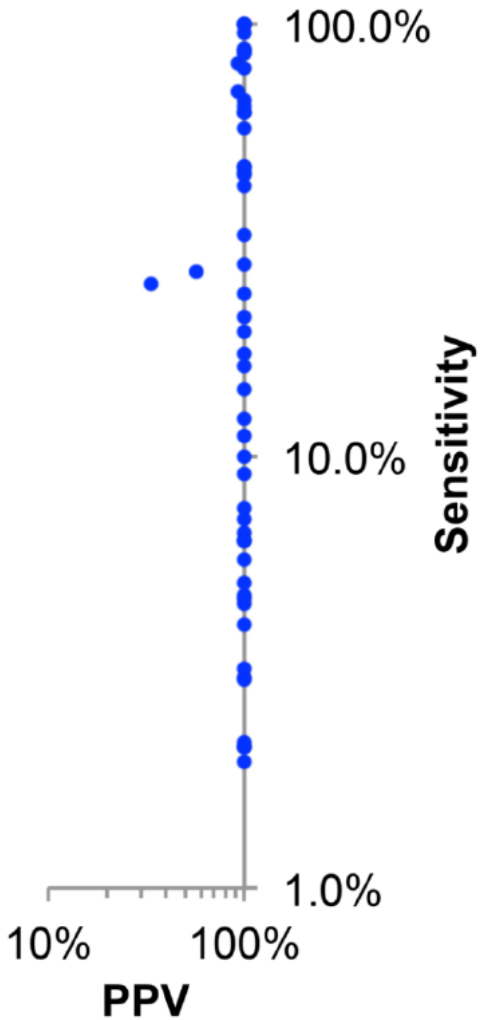
Protein-Based Human Identification

The obtained spectrum can be compared to protein reference databases to identify the protein and underlying peptide sequence. Peptides containing candidate GVPs need to be filtered to reduce false positive assignments. The accepted SAPs can then be used to impute nsSNPs.

Protein	SAP	nsSNP	REF/GVP	Allele
HEXB	I207V	rs10805890	GIL**I**DTSR	A
			GILVDTSR	G
KRT32	T395M	rs2071563	LEGEIN**T**YR	G
			LEGEINMYR	A
KRT32	R280H	rs72830046	CQYEAMVEAN**R**R	C
			CQYEAMVEANHR	T

Protein-Based Human Identification





Source: Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome (Parker et al., 2016).