Forensic Genetics

Module 15 – Session 9

Schedule – Day 3

Review	Session 7 & 8	8:00-8:50
Session 9	Other Techniques	9:05-9:55
Review	Whole course	10:10-11:00

Other Techniques

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

Next Generation Sequencing

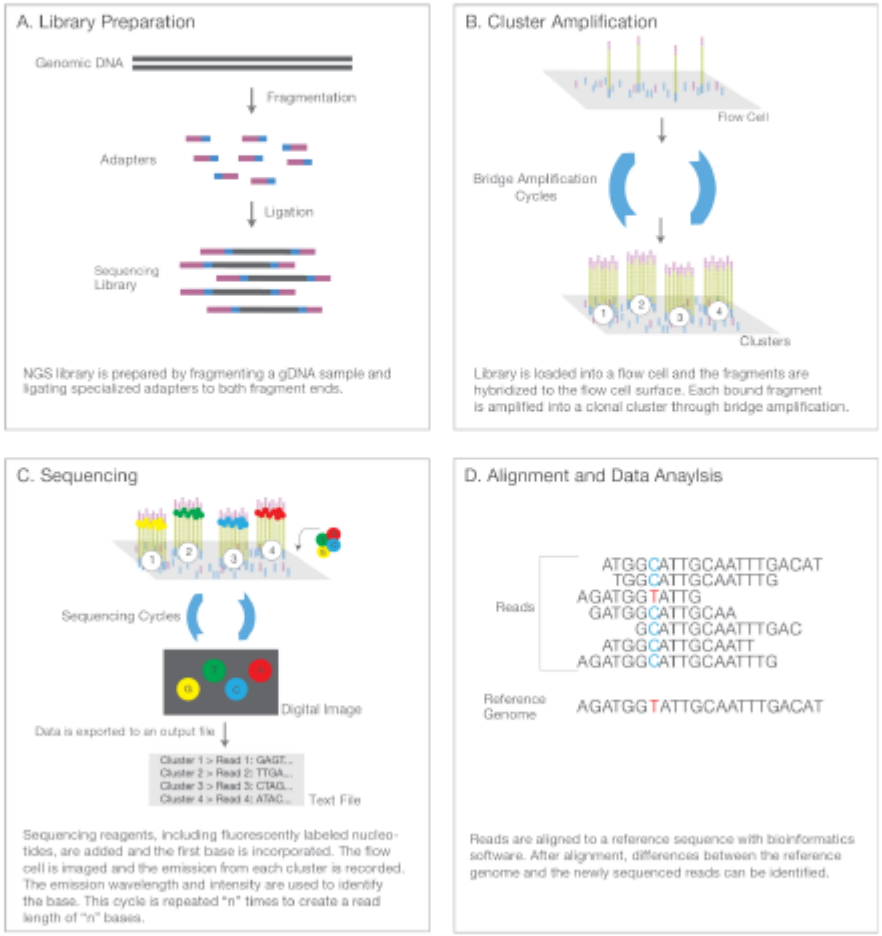
The introduction of *Next Generation Sequencing (NGS)* added a new dimension to the field of forensic genetics, providing distinct advantages over traditional CE systems in terms of captured information.

Locus	Allele number	Allele sequence
D3S1358	15	$[TCTA][TCTG]_3[TCTA]_{11}$
D3S1358	15	$[TCTA][TCTG]_2[TCTA]_{12}$
D18S51	20	$[AGAA]_{20}$
D18S51	20	$[AGAA]_{16}GGAA[AGAA]_3$

NGS is also referred to as Massively Parallel Sequencing (MPS), Second Generation Sequencing (SGS) or High-Throughput (HTP) sequencing.

NGS Workflow

By far the biggest player in the field of sequencing instruments is Illumina. Their workflow includes four basic steps:



Source: An Introduction to NGS Technology (Illumina, 2015).

NGS Workflow

The first three steps of the workflow consist of:

- **Library preparation:** A DNA sample gets fragmented and adapters are added to both fragment ends, after which a library is obtained through PCR amplification.

- **Cluster generation:** Each fragment bounds to the surface of a flow cell and is amplified through bridge amplification, resulting in a cluster that will produce a single sequencing read.

- **Sequencing:** Base calls are made per cluster using fluorescently labeled and reversible terminator-bound nucleotides.

NGS Data Output

The most common format for storing the output of NGS instruments is a text-based FASTQ file. In addition to the observed sequence string, the file also lists its corresponding quality score, representing an estimate by the base calling software of the potential error at each sequence position.

```
@SRR2120054.41 41 length=122
TGGGTTATTAATTGAGAAAACTCCTTACAATTTTCTTTTCTTTTCTTTTCTTTTCTTTTC
+
FBBGHHEGFGFHGGGCGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHH
@SRR2120054.42 42 length=117
CAACATTTGTATCTTTATCTGTATCCTTATTTATACCTCTATCTATCTATCTATCTATCTA
+
HHFCHHGHGHHHGHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHH
@SRR2120054.43 43 length=148
GTTGCTACTATTTCTTTTCTTTTTCTCTTTCTTTCCTCTCTCTTTTTCTTTCTTTCTTTCTTTCT
+
ADBGBFDFFFFFGGGGFFFHGGHHGCCHFHGHFHHHHFHHHHGHHHHHHHHHHGHHFGHHGGGHG
```

NGS Data

Results from sequencing platforms usually entail raw data, and need to be translated into information suitable for further (statistical) analysis.

- Software tools are available that align the reads to a reference sequence (**alignment**);
- Detect variations in the individual's genome (**variant calling**);
- And annotate the data using external information, resulting in a summarized data structure (**annotation**).

Instead of aligning to a reference sequence, sequence-searching techniques can be used that will use flanking sequences to detect STRs.

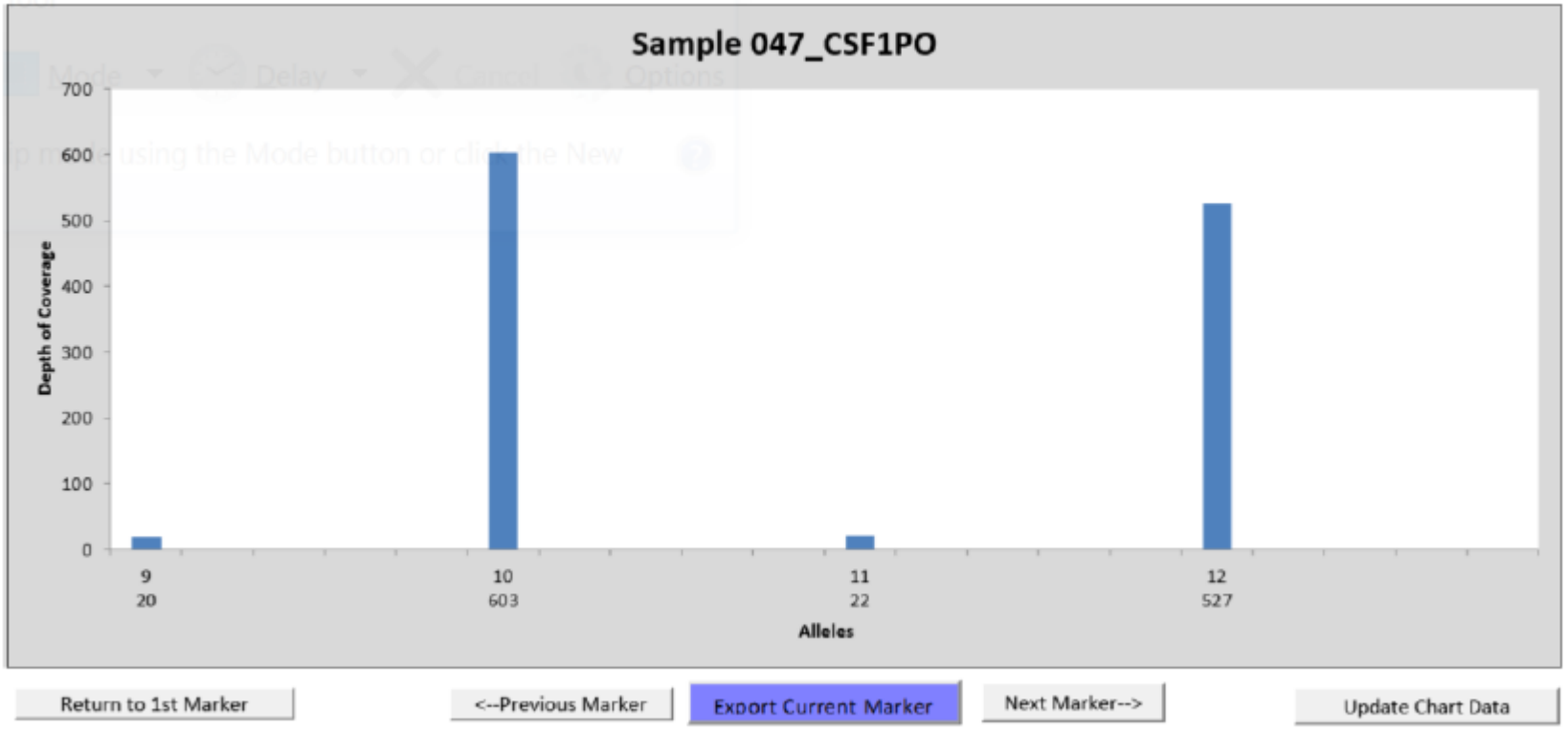
NGS Data Output

STRait Razor is an example of a sequence-searching technique, and produces output that looks as follows:

```
Amelogenin:0    63 bases    TAGTGTGTTGATTCTTTATCCCAGATGTATCTCAAGTGGTCCTGATTTTACAGTTCCTACCAC 1    0
Amelogenin:0    63 bases    TAGTGTGTTGATTCTTTATCCCAGACGTTTCTCAAGTGGTCCTGATTTTACAGTTCCTACCAC 1    0
Amelogenin:0    63 bases    TAGTGTGTTGATTCTTTACCCCAGATGTTTCTCAAGTGGTCCTGATTTTACAGTTCCTACCAC 1    0
Amelogenin:0    63 bases    TAGTGTGTTGATTCTTTATCCCAGATGTTTCTCAAGTGGTCCTGATTTTACAGTTCCTACCAT 1    0
CSF1PO:11       64 bases    CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT       0    2040
CSF1PO:12       68 bases    CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT   0    1810
CSF1PO:10       60 bases    CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT    0    70
CSF1PO:13       72 bases    CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT    0    14
CSF1PO:9        56 bases    CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT       0    3
CSF1PO:11       64 bases    CTCCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT       0    3
CSF1PO:11       64 bases    CTTACTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT       0    3
CSF1PO:11       64 bases    CTTCCTACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT       0    3
CSF1PO:11       64 bases    CTTCCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCTATCTAATCTATCTATCTT       0    2
```

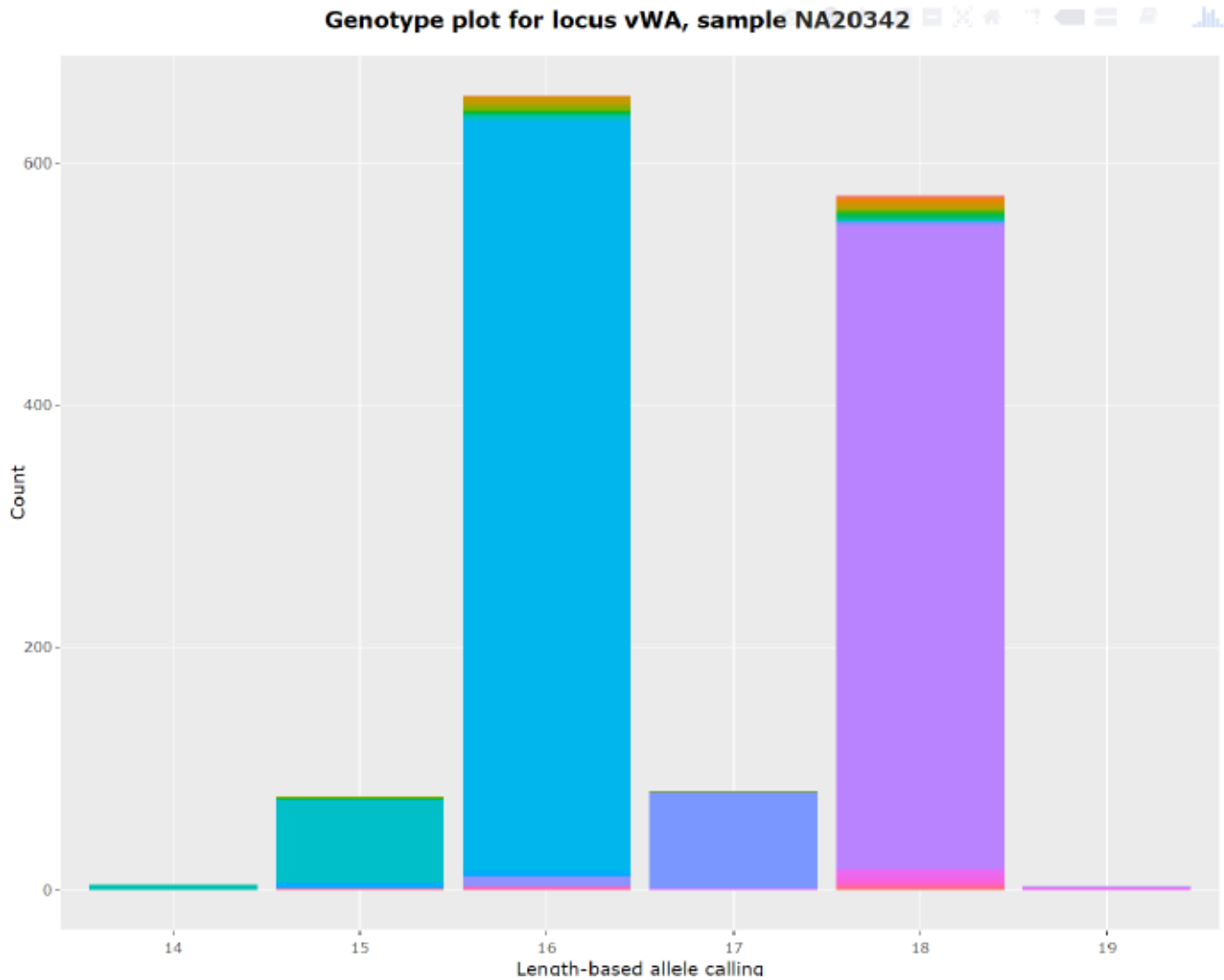
NGS Data Output

A DNA profile can be visualized similar to an epg:



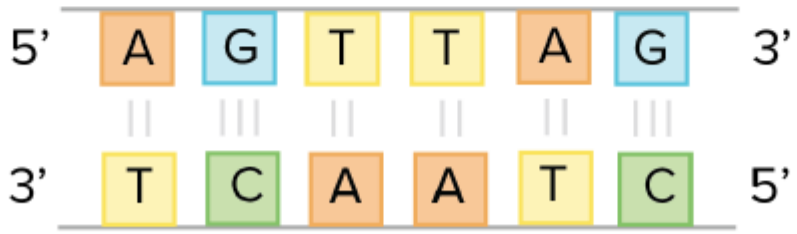
NGS Data Output

A DNA profile can be visualized similar to an epg:



NGS Considerations

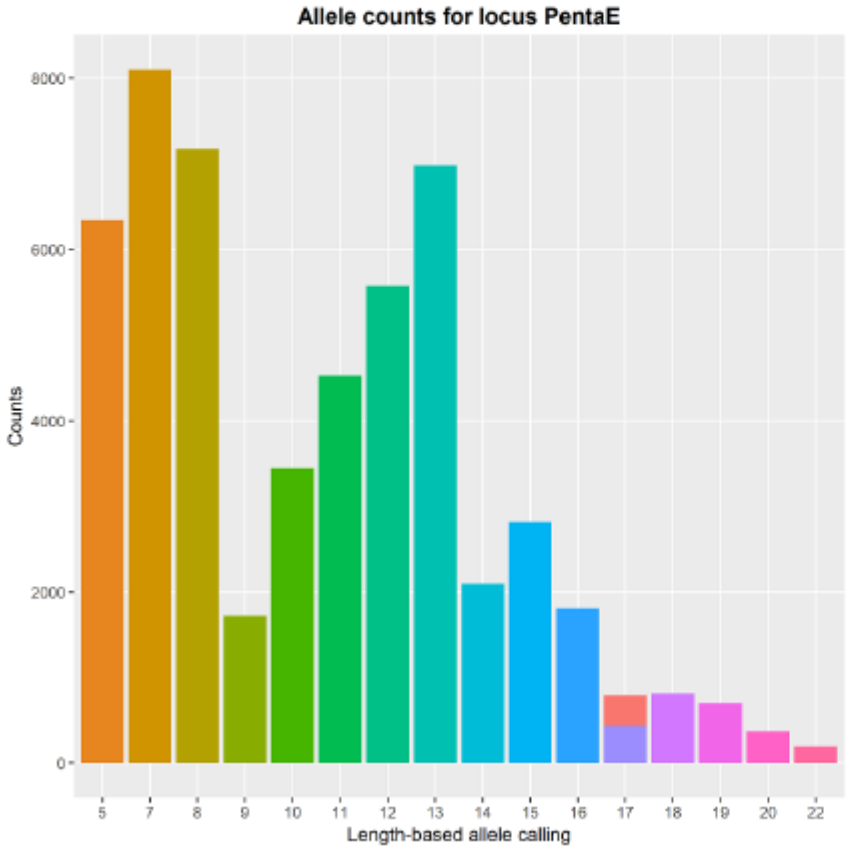
- Reads vs. peaks (discrete vs. continuous data)
- Discovery of previously unknown alleles and more variability
- New system of nomenclature needed
- Direction of strand reporting

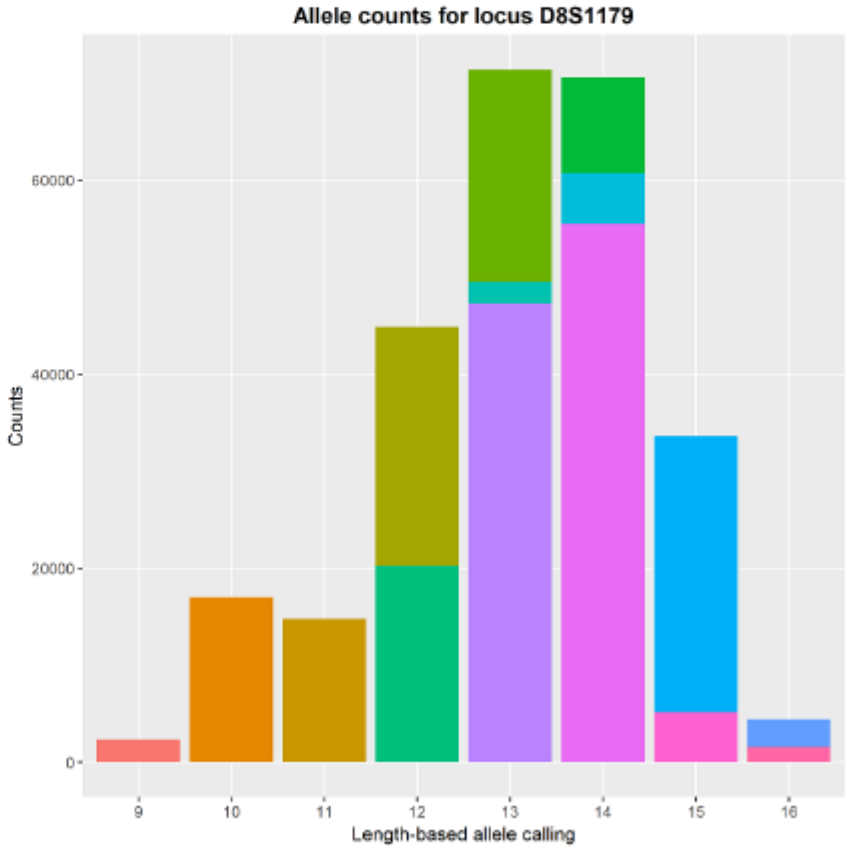


Source: https://www.khanacademy.org/science/biology/dna-as-the-genetic-material/dna-replication/a/molecular-mechanism-of-dna-replication.

LB vs SB Allele Callings

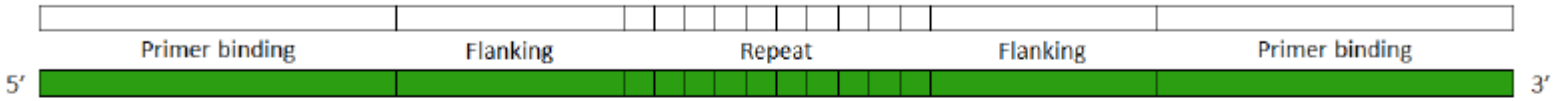
Locus Penta E is already quite polymorphic, so NGS data does not lead to significant improvements. For locus D8S1179, sequencing leads to a substantial increase in variability.





Flanking Region SNPs

Additional variation has been found in the flanking regions adjacent to repeat motifs.



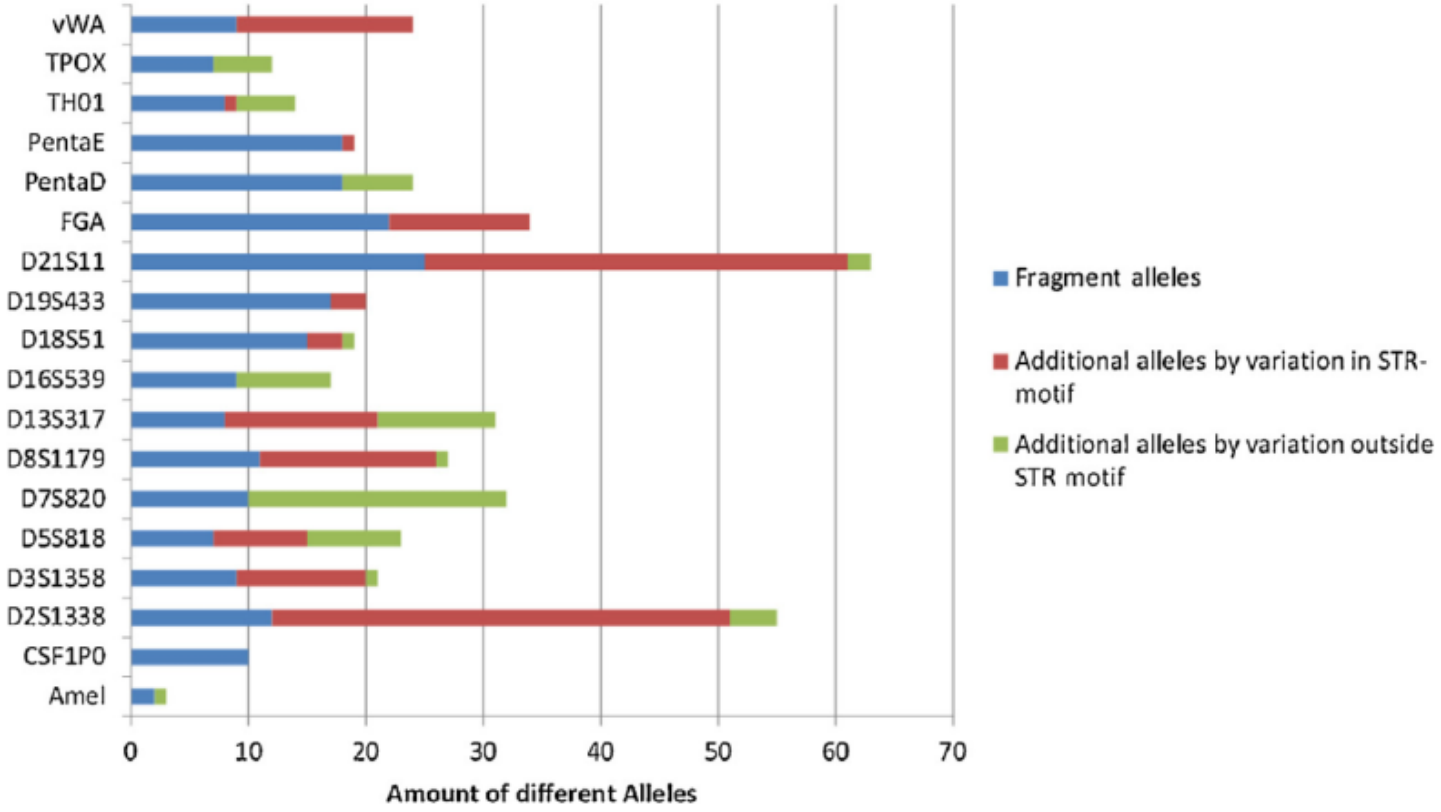
Source: Forensic DNA Evidence Interpretation (Buckleton et al., 2016).

For STR loci in which repeat regions do not display sequence differences, flanking region SNPs may still add substantial variability. Knowledge of these variants can be utilized in primer design to ensure optimal positioning during the PCR process.

Locus	LB Allele	SB Allele	SB Allele with SNPs
D16S539	11	$[GATA]_{11}$	$[GATA]_{11}$rs11642858[A]
D16S539	11	$[GATA]_{11}$	$[GATA]_{11}$rs11642858[C]

Observed Sequence Variation

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 Modeling

New models need to be developed and implemented to accommodate NGS data, with the ultimate goal of developing a probabilistic approach for NGS mixture interpretation.

CE-based models can be used as a basis for NGS modeling. Both methods make use of the PCR process, so it is expected that artifacts such as stutter are similar.

However, peak heights need to be substituted with read counts and the remaining biological processes differ. This will materially affect the modeling parameters.

NGS Stutter Modeling

NGS data generally show higher stutter percentages than CE data. Illumina's ForenSeq uses the following thresholds (compared with Thermo Fisher's NGM Select Kit for CE data):

	Stutter Filter (%)	
Locus	**CE**	**NGS**
TH01	5	10
D2S441	9	7.5
vWA	11	22
FGA	11.5	25
D12S391	15	33
D22S1045	17	20

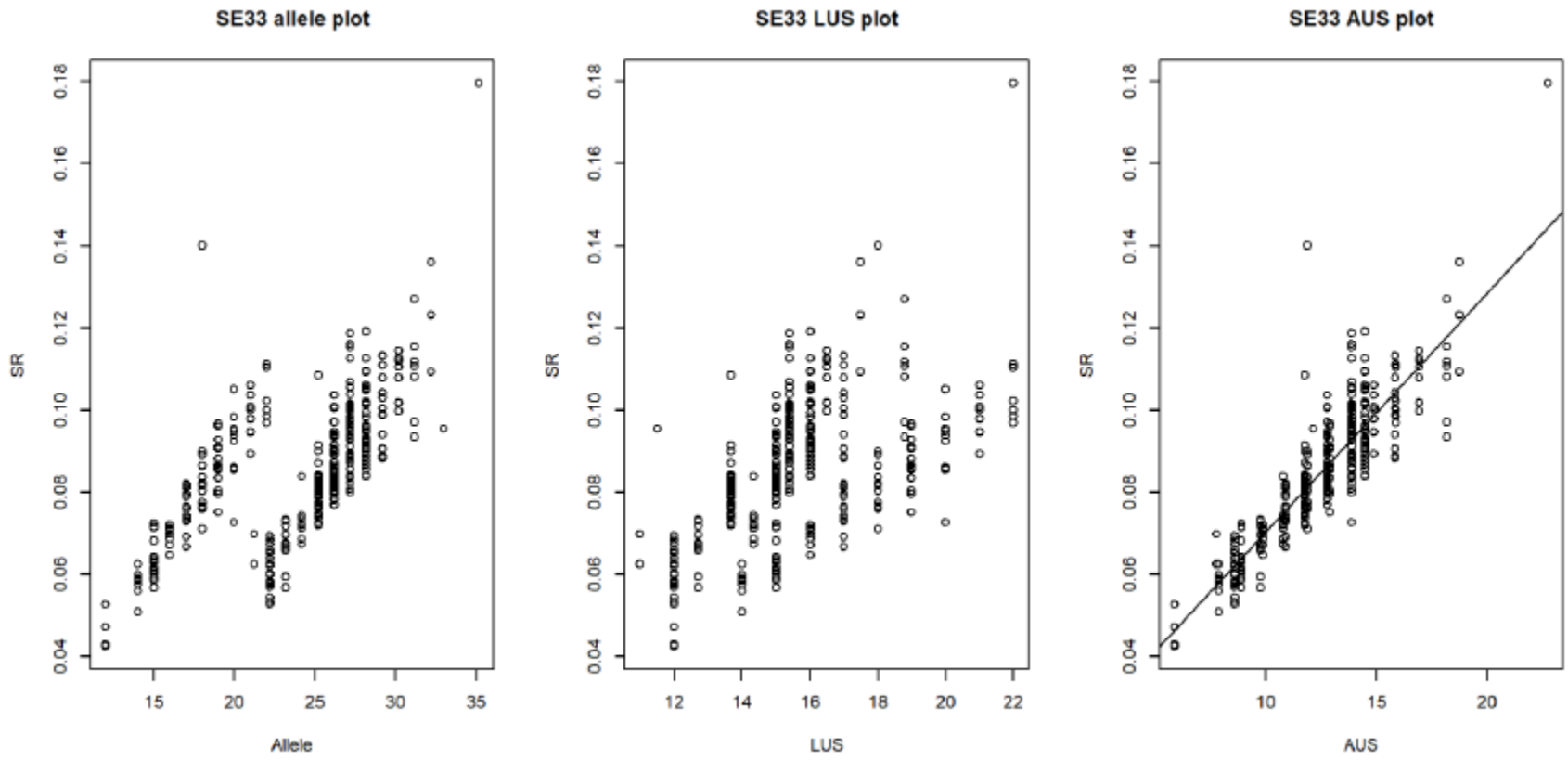
Multi-Sequence Stutter Model

A multi-sequence model takes into account all uninterrupted stretches (AUS) as potentially contributing to stuttering.

Allele	Repeat motif
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{11}G\ AAGG[AAAG]_2AG$
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{11}AA\ AAAG[AAAG]_9G\ AAGG[AAAG]_2AG$
22	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{22}G[AAAG]_3AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_7AA\ AAAG[AAAG]_{14}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_8[AG]_5[AAAG]_{12}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{12}GAAGG[AAAG]_2AG$

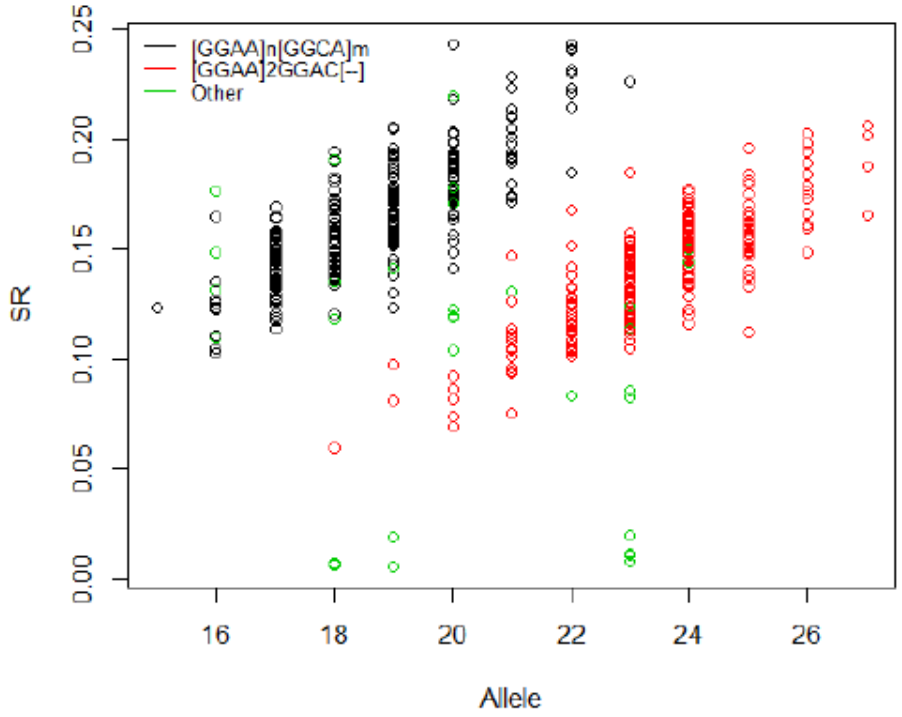
Examples of locus SE33 sequences.

Multi-Sequence Stutter Model for SE33



Stutter Modeling and Sequence Variation

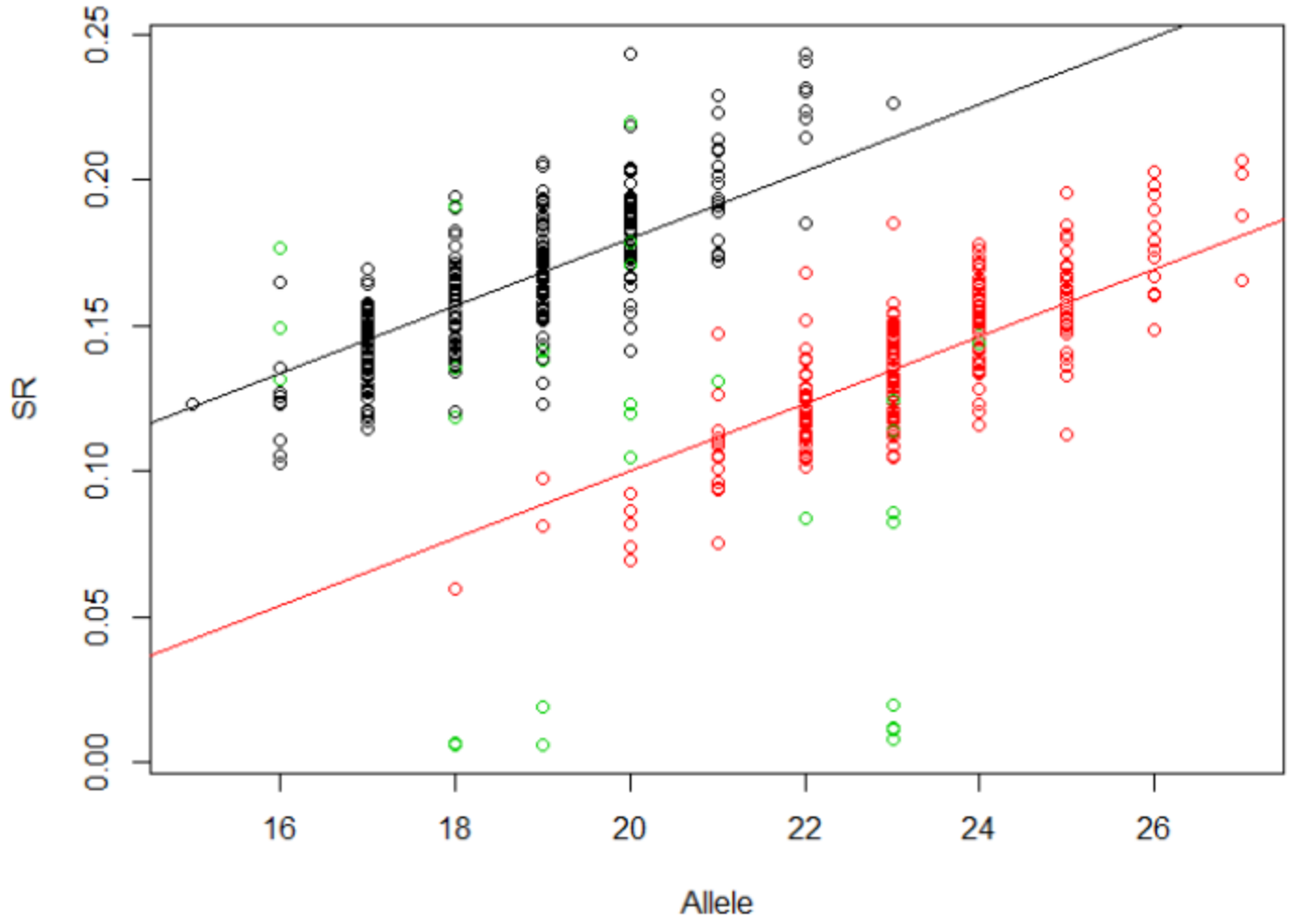
What about variation that is suggested to be attributable to sequence motif?



Stutter ratios for locus D2S1338.

Models fitted based on AUS still left some variability unexplained for some loci.

NGS Stutter Modeling – Sequence Variation



Stutter ratio model for locus D2S1338.

NGS Stutter Modeling

With the sequence variations now in hand, it is possible to decompose certain stutter affected heterozygotes, composite stutter and regular stutter products.

For locus TH01, for example, there are two possible (back) stutter products:

Product	LB Allele	SB Allele
A	8.3	$[AATG]_6ATG[AATG]_2$
B	8.3	$[AATG]_5ATG[AATG]_3$

NGS Stutter Modeling

The total expected stutter count is now the sum of the two stutter products:

Product	LB Allele	SB Allele
A	8.3	$[AATG]_6ATG[AATG]_2$
B	8.3	$[AATG]_5ATG[AATG]_3$

$$E_{(a-1)} = \phi_A E_A + \phi_B E_B,$$

with ϕ_A and ϕ_B the proportion of stutter product A and B, respectively.

These proportions will likely reflect previous observations (e.g. longer sequences stutter more, but not all stutter come from the LUS).

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 Stutter Modeling – BLMM Model

Model stutter per stutter sequence instead of parental allele, based on the block length of the missing motif (BLMM).

Category	Allele	Sequence	BLMM
Allele	9.3	$[AATG]_6ATG[AATG]_3$	–
Stutter	8.3	$[AATG]_6ATG[AATG]_2$	3
Stutter	8.3	$[AATG]_5ATG[AATG]_3$	6

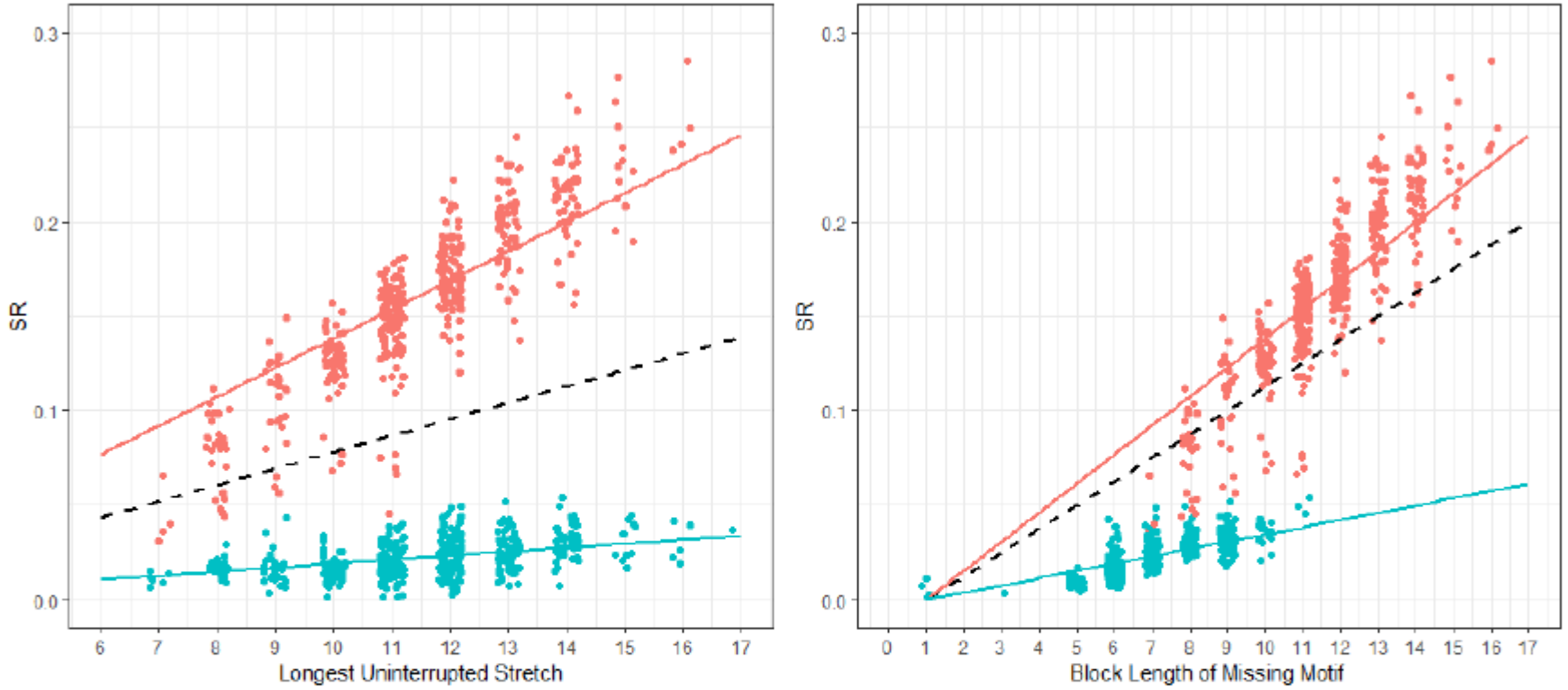
A linear model can be fitted with intercept through $(1, 0)$, based on the idea that stutter can occur only after the first repeat.

$$SR' = \beta(\text{BLMM} - 1)$$

This also avoids the problem of predicting negative stutter ratios.

Source: Stutter analysis of complex STR MPS data (Vilsen et al., 2018).

NGS Stutter Modeling



Stutter ratio model for locus D12S391.

The larger stutter ratios result from stutter from the LUS of the parental allele.

NGS Stutter Modeling - Discussion

- How to determine variation?
- What about micro-variants?
- What about the possible influence from flanking variation?
- What about dependencies between stretches?

NGS Population Structure

Likelihood ratios use match probabilities, which rely on appropriate estimation of the population structure parameter θ. Values of $1\% - 3\%$ are common in forensic DNA evidence evaluations.

When implementing NGS-based methods, the effect of sequence data on θ estimates needs to be analyzed.

Allele and/or genotype matching between individuals within and between populations can help us assess relative relatedness[1].

[1] Population-specific F_{ST} values for forensic STR markers: A worldwide survey (Buckleton et al., 2016).

NGS Population Structure

Our data consist of DNA samples from 350 individuals over 5 different continental groups sequenced and annotated with Illumina instrumentation.

- Using length-based allele callings, within-population matching was 0.2165 and between-population matching was 0.1968.

- Using sequence-based allele callings, within-population matching was 0.1878 and between-population matching was 0.1664.

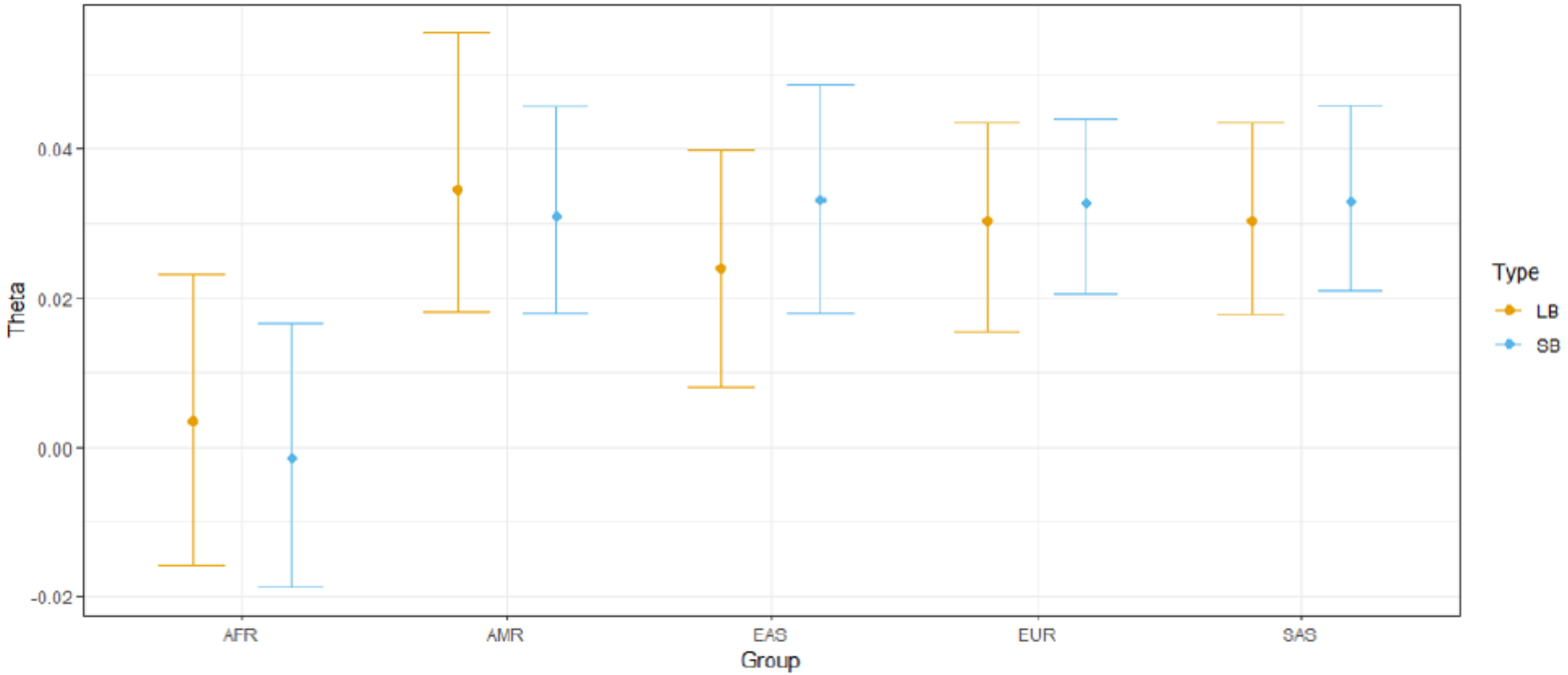
NGS Population Structure

Locus-specific θ estimates may decrease, increase, or stay the same.

	# Alleles			θ Estimate	
Locus	**LB**	**SB**	**Diff**	**LB**	**SB**
D21S11	17	65	48	0.0259	0.0383
D1S1656	18	34	16	0.0174	0.0146
TPOX	8	8	0	0.0402	0.0402

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 et al., 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

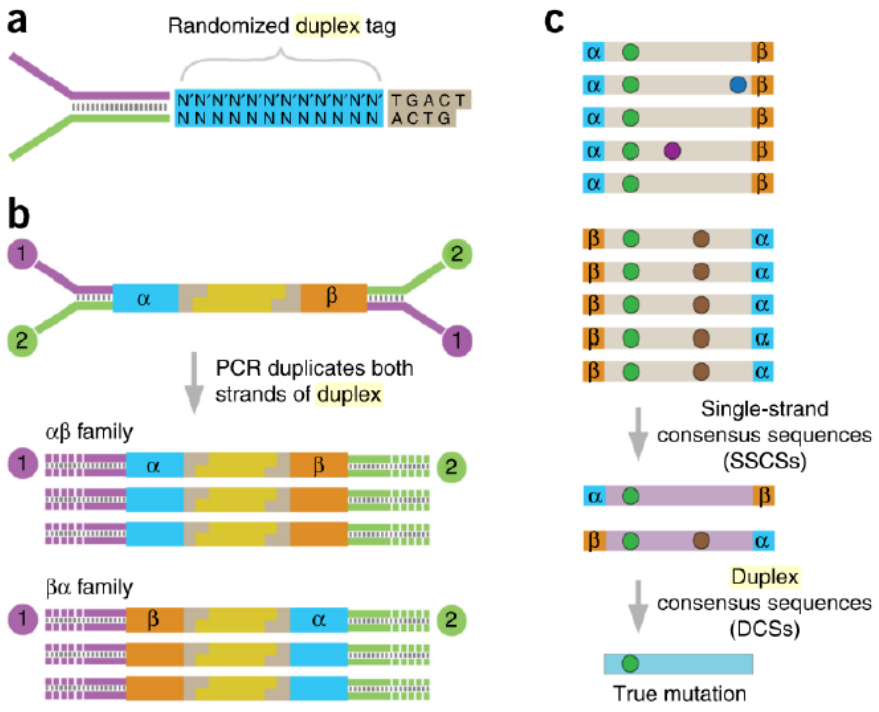
Most NGS approaches have a relatively high error rate and are therefore not suitable for detecting in vivo mutations. To overcome this limitation, a highly sensitive sequencing methodology termed *Duplex Sequencing (DS)* has been developed.

- 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).

Inference of Ancestry

Suppose that a population can be classified into K groups. The probability of a DNA sample with profile D coming from group k, can be written as:

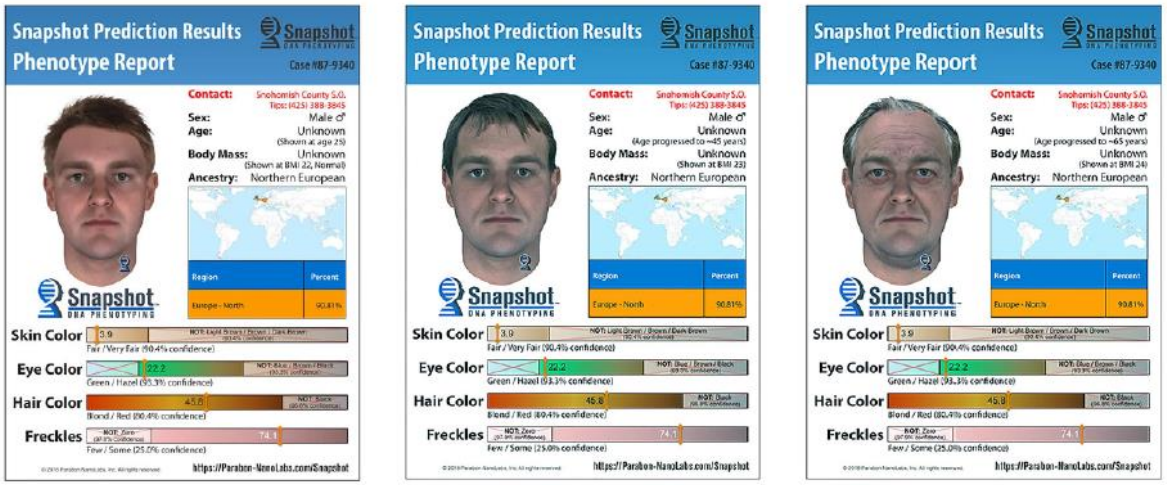
$$\Pr(\text{group } k|D) = \frac{\Pr(D|\text{group } k)\Pr(\text{group } k)}{\sum_{j=1}^{K}\Pr(D|\text{group } j)\Pr(\text{group } j)}.$$

STR profiles can give some information, although they provide limited discriminatory power in this context. Instead, SNP sets (so-called ancestry informative markers) have been demonstrated to be useful for distinguishing individuals from certain (sub-)populations.

Inference of Phenotype

SNPs may be linked to some visual phenotypes, including hair color and eye color. Other facial characteristics can now also be predicted from genotypes with some accuracy.

These SNP associations can potentially be used in forensic settings, e.g. in combination with a description of an eyewitness of a target individual.



Picture rendered by Parabon Nanolabs.

Source: Technique Used to Find Golden State Killer Leads to a Suspect in 1987 Murders (Murphy, 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

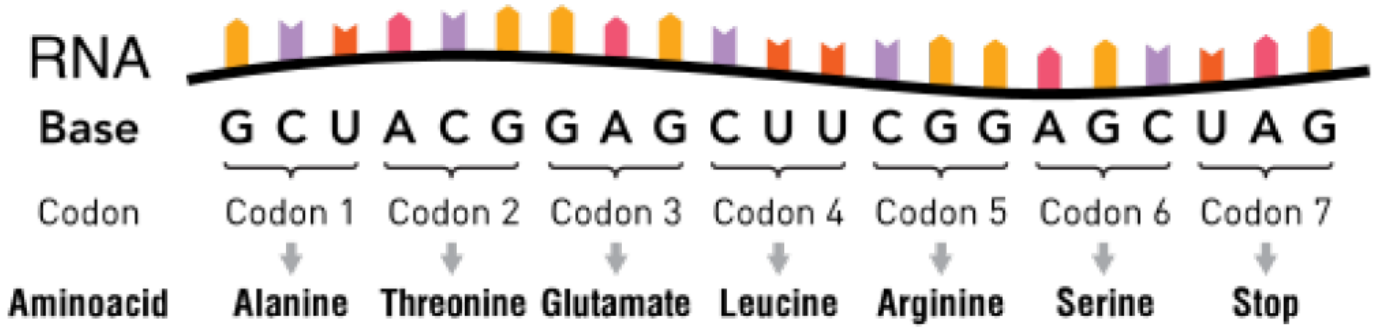
Certain sections of DNA, called *exons*, are coded for a *protein*, i.e. a macro-molecule consisting of one or more long chains of amino acid residues performing a vast array of functions within organisms. Two steps are required to read the information encoded in a gene's DNA and produce the protein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

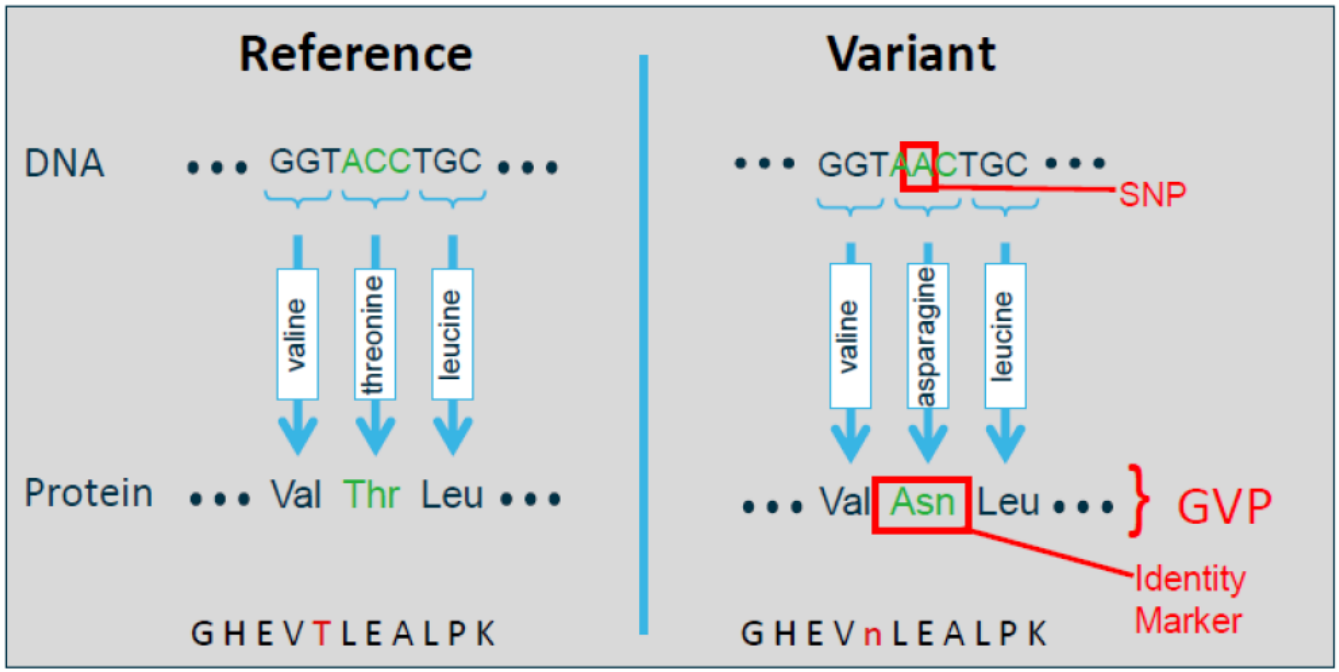
Protein-Based Human Identification

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.

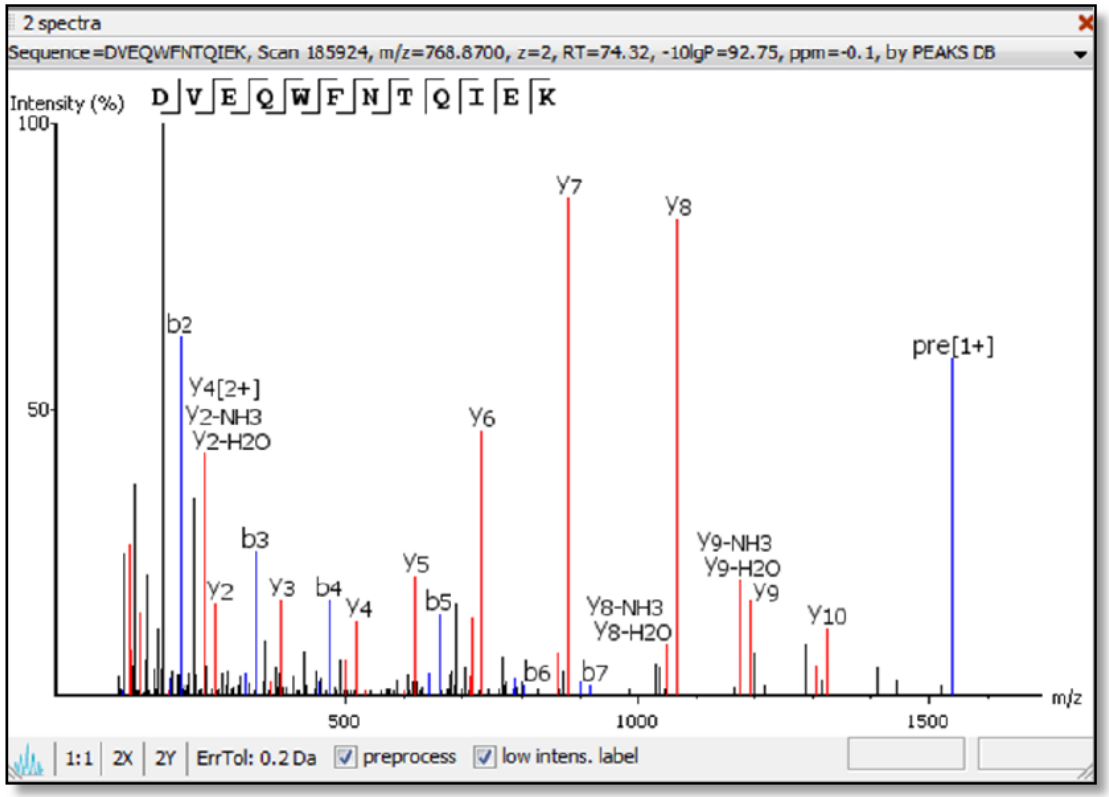
Protein-Based Human Identification

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.

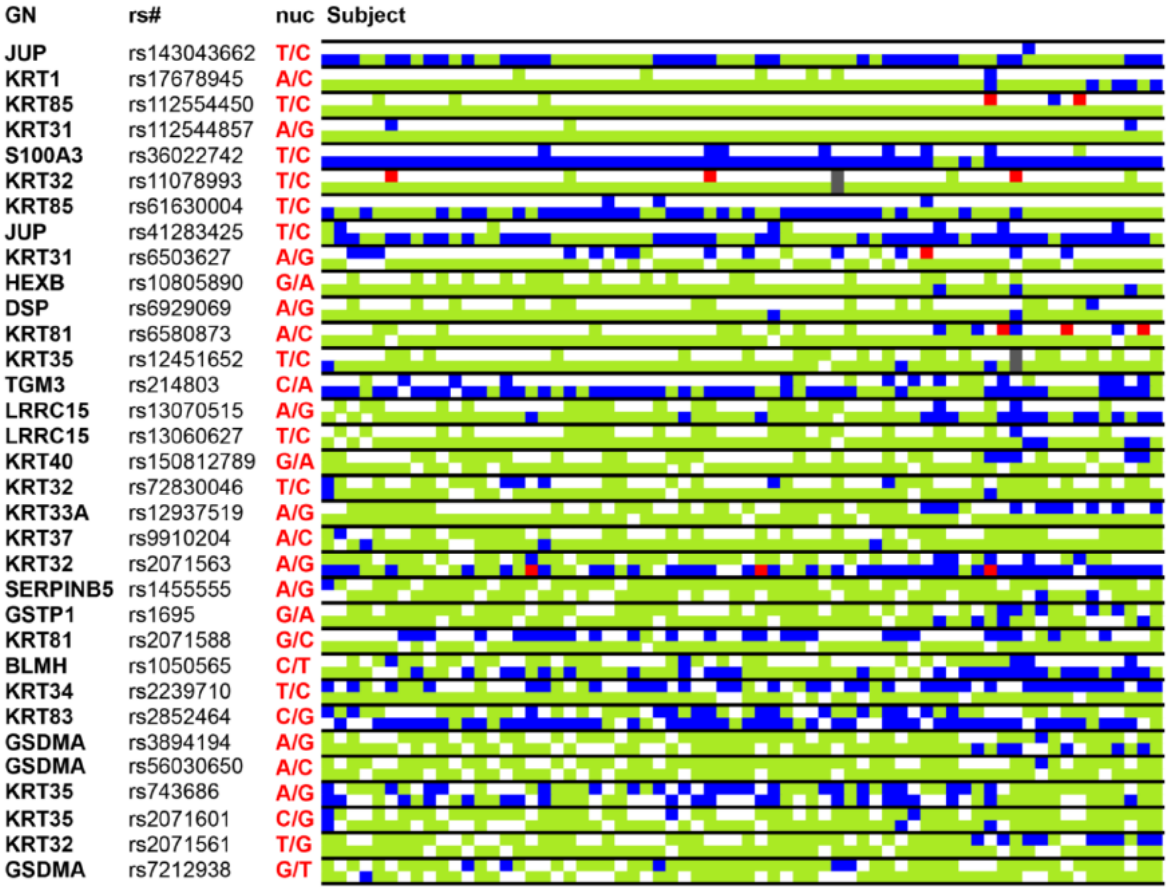


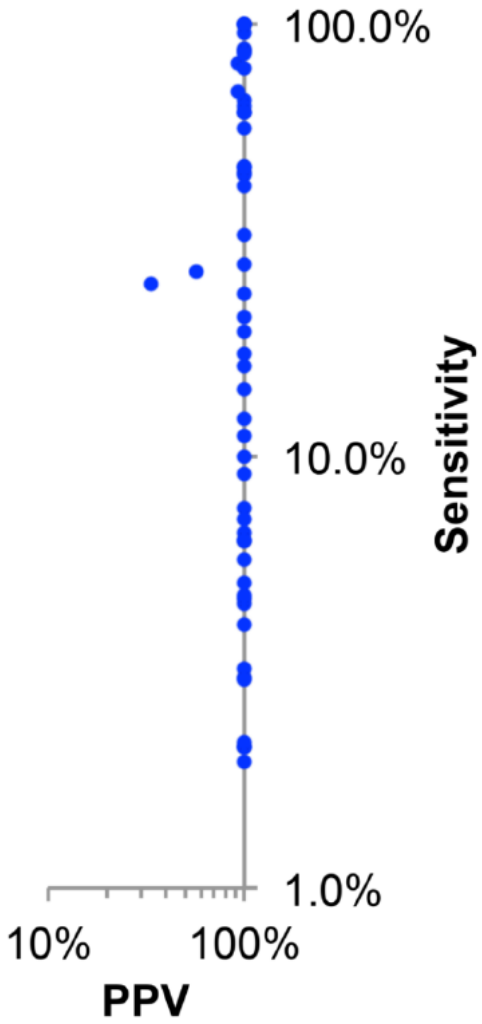
Protein-Based Human Identification

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).