Section 2: STR Typing Characteristics

STR Typing

- Forensic DNA interpretation has been centered on the analysis of STRs (*short tandem repeats*), i.e. short DNA sequences that are repeated several times.

- These repeat patterns are located in areas called *loci* and vary among individuals.

- Variants for a given locus are called *alleles* and it is this variation (called polymorphism) that allows us to associate a particular DNA sample with an individual person.

Mutations

Mutations are the cause of the variation encountered in DNA and one of the reasons that STR loci render highly informative markers in forensic genetics. Most of the mutations are caused by an error during DNA replication (although other mechanisms and external influences can also lead to a change in DNA sequence).

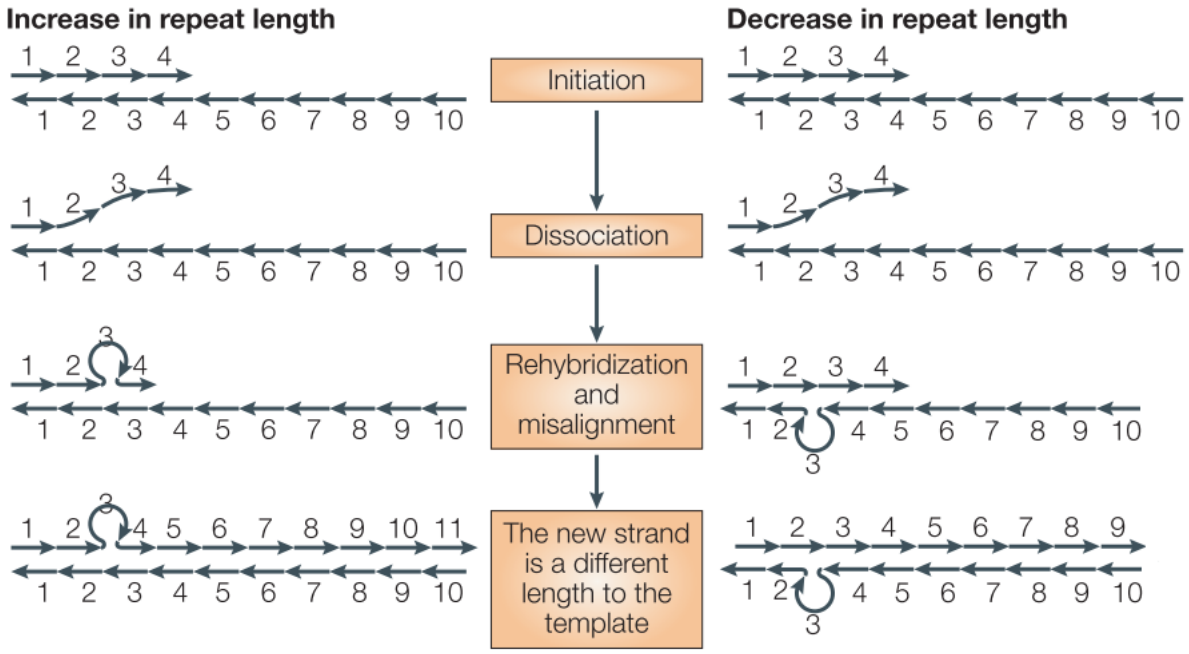
Examples of mutations:

- **Substitutions:** A point mutation where one base is substituted for another, such as a SNP.

- **Indels:** Small insertions/deletions due to the addition of one or more extra nucleotides into the DNA or the loss of a section of DNA.

Slipped Strand Mispairing

STR polymorphisms derive mainly from variability in length. A proposed mechanism for these genetic variations is the *slipped strand mispairing* (SSM) mechanism: the dissociation of replicating DNA strands followed by misaligned re-association.



Source: Microsatellites: simple sequences with complex evolution (Ellengren, 2004).

Mutations

The average in vivo mutational rates of the core STR loci are estimated to be between 0.01% and 0.64%, although the exact mutation rate of a locus is associated with the base composition of the repeats and the length of the allele.

Meiotic mutations, occurring in the process of transmitting an allele from a parent to a child, can cause the child's allele to differ from its parental type and can be important for paternity and other relatedness testing.

Mitotic mutations, or somatic mutations, occur within an individual and are of importance for identification and, although rare, could result in different profiles being recorded from the same individual (and hence possibly lead to a false exclusion).

STR Typing

To effectively interpret DNA evidence, we need to understand STR typing characteristics such as

- the PCR-CE process
- anomalies (like drop-ins/drop-outs)
- peak height variability
- stutter

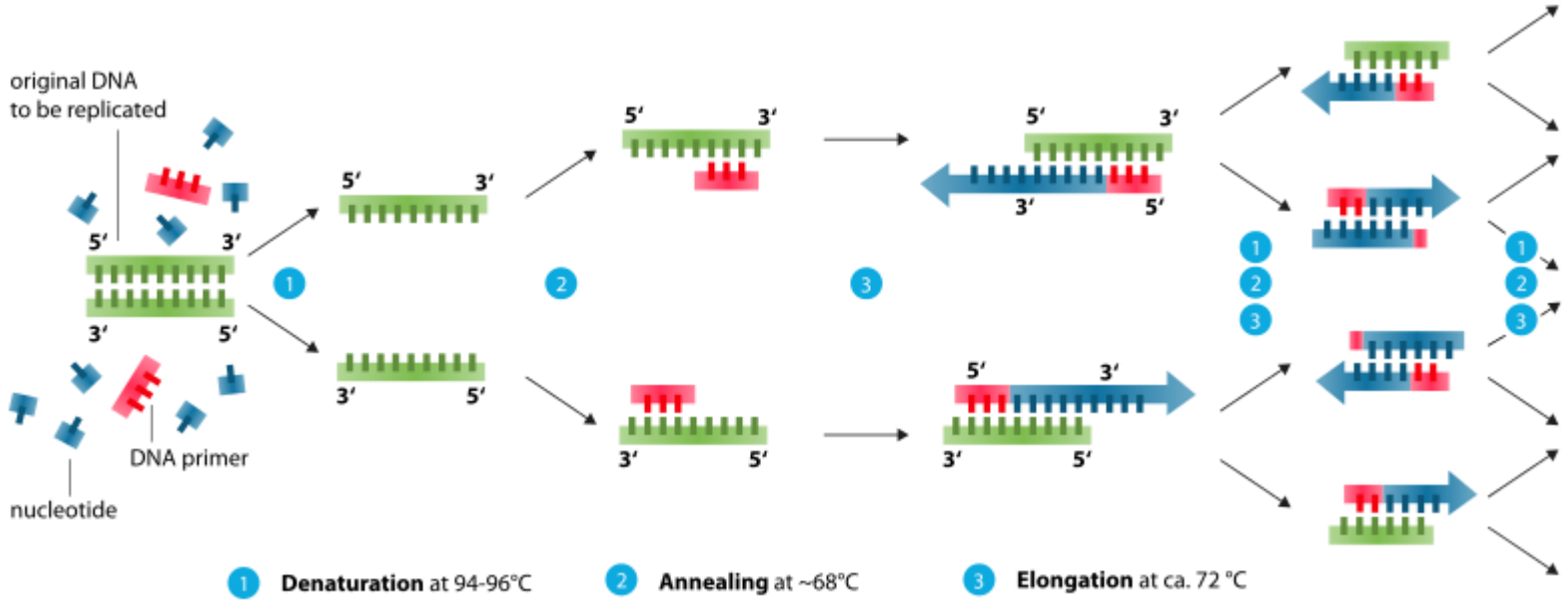
Understanding PCR

To produce an STR profile from a biological sample many identical copies of the DNA molecules within the target region (i.e. the DNA *template*) are needed. PCR (*polymerase chain reaction*) can be used to copy, or amplify, DNA through the following steps:

- **Denaturation:** Melting DNA such that the double-stranded template separates into two single-stranded DNA molecules.
- **Annealing:** Cooling the mixture to let *primers* bind to the strands.
- **Elongation:** DNA polymerase (a special copier molecule) completes missing sequences using available nucleotides.

Understanding PCR

These basic steps constitute one cycle, so by repeating this process, the DNA target gets amplified to millions of copies.



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

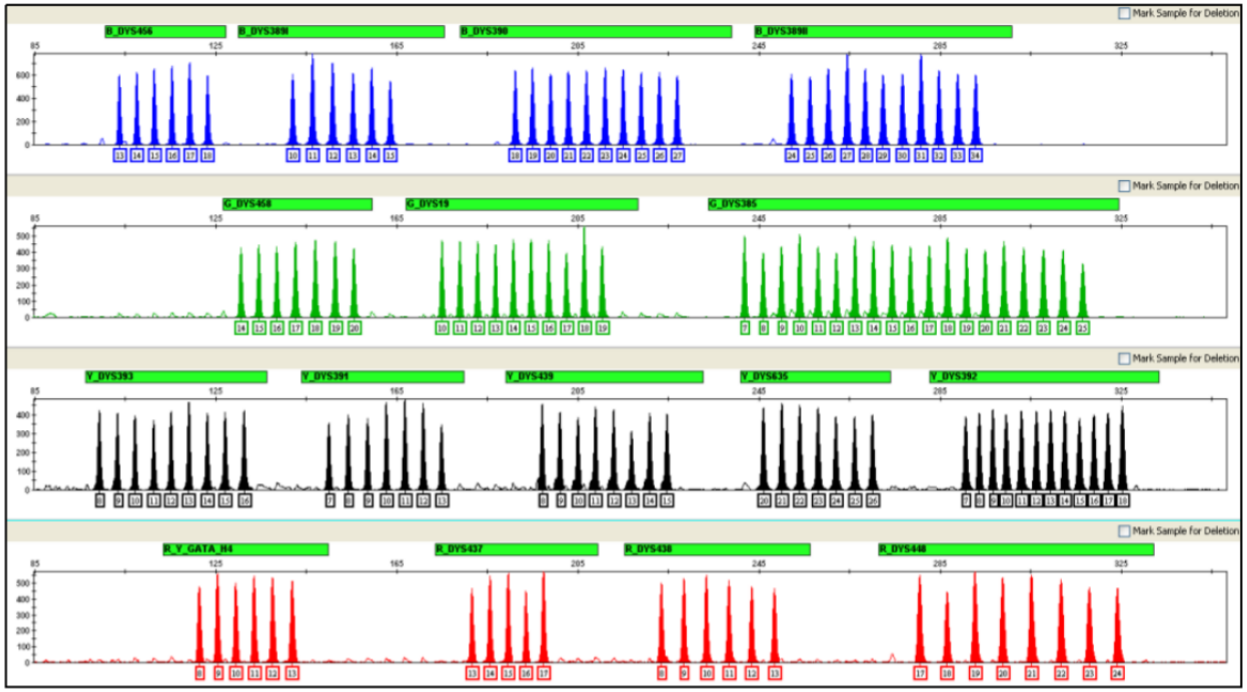
Capillary Electrophoresis

To obtain meaningful results from the PCR process, *capillary electrophoresis* (CE) has traditionally been used, allowing forensic scientists to gain access to the allele numbers contained in a DNA sample.

- DNA products are injected into the capillary where they travel in the direction of a positive charge;
- The travel time depends on the fragment size and can thus be used to infer the number of repeats;
- Primers are labeled with fluorescent dye, which will emit visible light at the detector window of the capillary.
- The fluorescence, measured in relative fluorescence units (RFU), is recorded over time and can be visualized with an *electropherogram* (epg).

Allelic Ladders

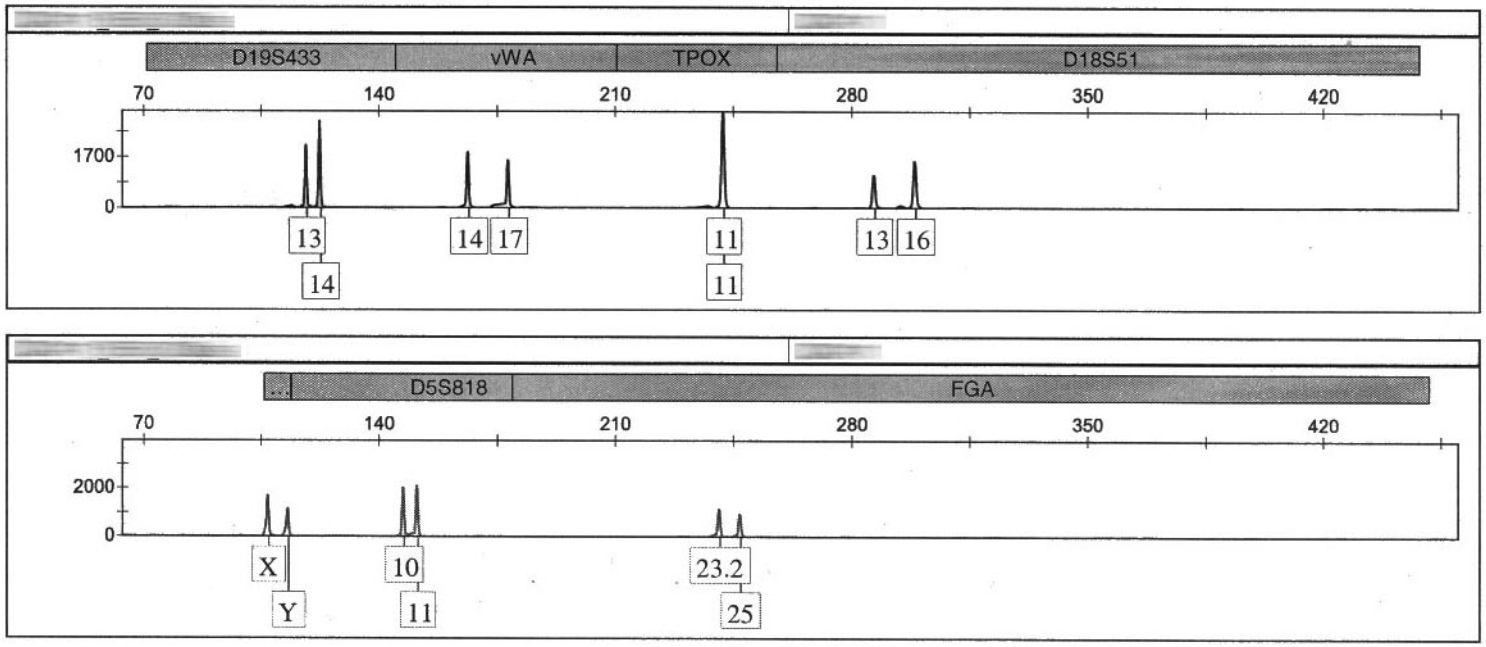
PCR-CE output can be compared to *allelic ladders* to determine allele designations.



Source: *AmpFℓSTR Yfiler PCR Amplification Kit User Guide*.

Example of an Electropherogram

An epg shows allelic designations, represented by peaks, with integer values indicating the number of complete repeat motifs and additional nucleotides separated by a decimal point.



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

STR Classes

STR loci may be categorized in three different classes, based on how well alleles conform to the core repeat pattern:

- **Simple STRs**: only show variation in the number of repeats without additional sequence variation.
- **Compound STRs**: consist of several adjacent repeats of the same repeat unit length.
- **Complex STRs**: contain repeats of variable length as well as sequences.

Examples of STR Classes

STR loci may be categorized in three different classes, based on how well alleles conform to the core repeat pattern:

Class	**Locus**	**Allele sequence**
Simple	CSF1PO	$[TCTA]_8$
Simple	Penta D	$[AAAGA]_{12}$
Compound	vWA	$[TCTA][TCTG]_4[TCTA]_{13}$
Compound	D22S1045	$[ATT]_7ACT[ATT]_2$
Complex	FGA	$[TTTC]_3TTTTTTT[CTTT]_{11}CTCC[TTCC]_2$
Complex	D1S1656	$[TAGA]_4TGA[TAGA]_{13}TAGG[TG]_5$

Micro-variants and sequence variations

An allele that contains an incomplete repeat unit is called a *micro-variant*. They are usually rare, with the exception of allele 9.3 at THO1, and can be reliably distinguished if the variant alters the allele length.

However, same-length variants (i.e. *isoalleles*) will be recorded as matching alleles even if they differ at sequence level. This means that CE-based methods have less discriminatory capability than is potentially available via sequencing techniques.

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	$[AGAA]_{16}GGAA[AGAA]_3$

Anomalies

If DNA profiling technologies were flawless, and no other (human) errors have been introduced, an STR profile would provide a perfect representation.

For good-quality samples, this is a reasonable assumption and STR allele calling is usually pretty straightforward.

However, a number of anomalies may still arise. And more importantly, crime scene profiles rarely belong to this category and usually consist of low template samples that may be contaminated and/or degraded, making them even more prone to typing errors.

PCR-CE Method Response Categories

PCR-CE method response can be classified into several categories:

- **Analyte signal**: peaks corresponding to one or both authentic alleles at a locus.

- **Molecular artifacts**: peaks identifiable as systematic method error, such as stutter.

- **Background noise**: method response resulting from negative controls or that cannot be classified as analyte signal or molecular artifact.

Pull-up Peaks

Small *pull-up* peaks may be observed as a consequence of the spectral overlap of the different dye colors.

- Typically, a blue peak may pull up a green peak directly below it, usually as a result of overloading.

- This can be problematic if the minor peak coincides with the position of a potential allele.

Drop-ins

Allelic peaks that do not come from any of the assumed contributors to a DNA sample are termed *drop-ins*.

- Drop-ins may arise from airborne DNA fragments in a laboratory, or due to environmental exposure at the crime scene, and can typically not be reproduced on subsequent analysis of the same DNA extract.

- Verification of the source of drop-ins is not usually possible, although the existence of drop-ins can be confirmed through negative controls.

- As techniques become more sensitive, more drop-ins will occur, and potential difficulties may arise when they are incorrectly classified as analyte signal.

Contamination

Drop-ins are related to the concept of *contamination*.

- Contamination is one of the causes for drop-ins, as a result of DNA that got into a sample during collection or subsequent analysis.

- Databases of lab and scene staff can facilitate the identification of certain kinds of contamination.

- The most dangerous form of contamination is between different evidence samples, from either the same or different crime scenes.

- The observation of a more complete profile resulting from contamination is referred to as gross contamination.

Drop-outs

A *drop-out* occurs when an allele from a contributor to the crime scene sample is not reported in the STR profile.

- This happens when a peak fails to reach the detection threshold, meaning that they cannot be reliably distinguished from background noise.

- Low template DNA samples and degradation increase the drop-out rate, which is believed to be associated with DNA fragment length.

Silent or Null Alleles

Drop-outs should not be confused with *silent* alleles, in which a system is unable to visualize an allele.

- Differences in protocol could generate a silent allele in one laboratory that is non-silent in another laboratory and may be problematic in relatedness testing.

- Causes are PCR failure (e.g. due to a primer binding site mutation) or a copy number variant deletion.

- They may be detected when the peak height variability falls outside the normal range.

Copy Number Variants

When mutations affect relatively large segments of DNA (1 kb or larger) the resulting difference is called a *copy number variant* (CNV).

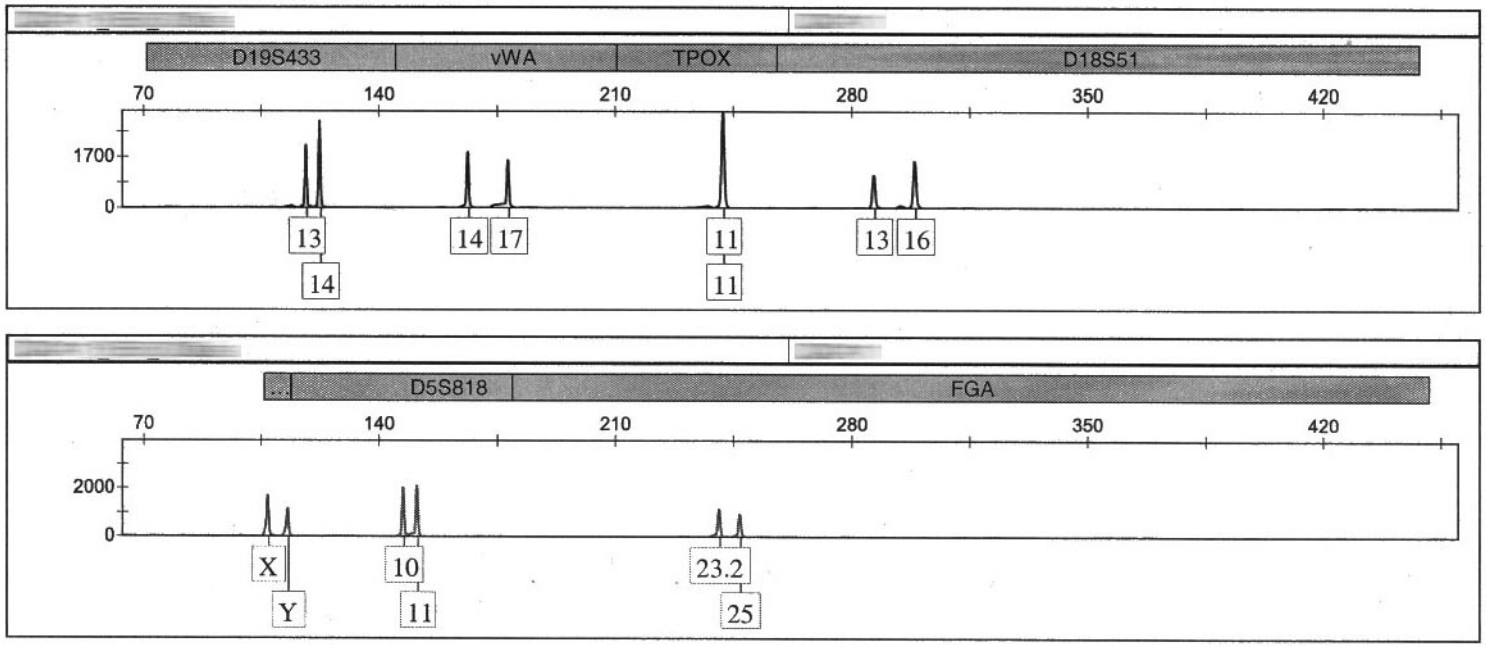
This can lead to difficulties in forensic applications when:

- a deletion or duplication leads to an unusual pattern of peak heights (single peak with height similar to heterozygote alleles or unbalanced heights because of overlap);

- a single-contributor profile contains a locus displaying three peaks (and may be confused with a low-template second contributor).

Peak Height Variability

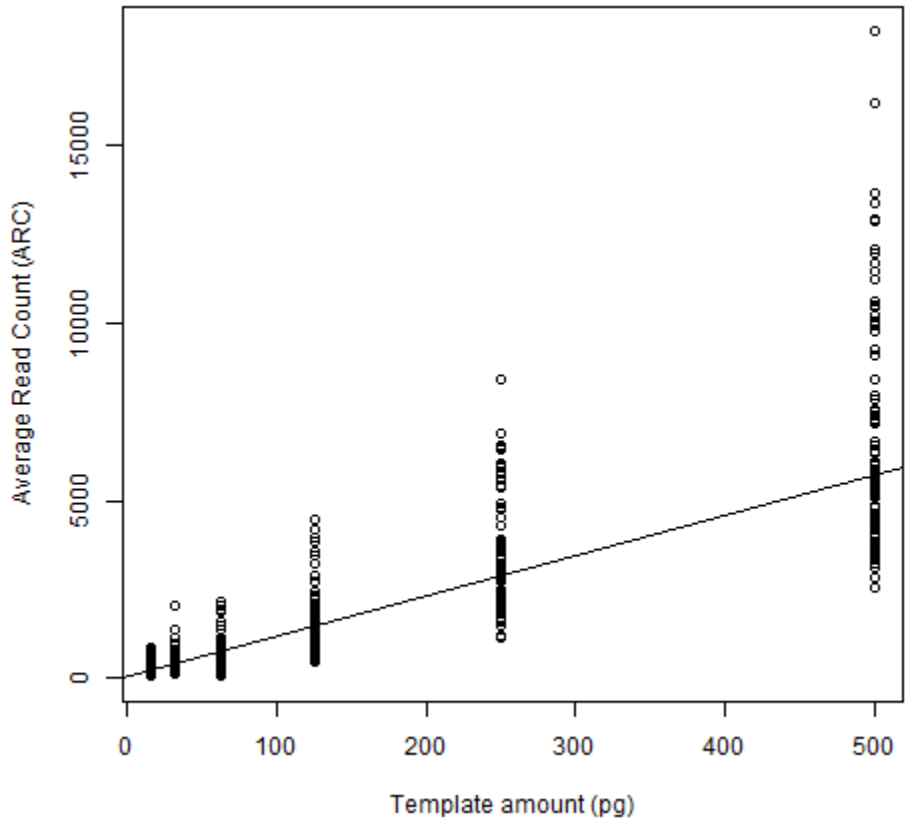
Besides the anomalies already discussed, several other factors play a role in observed variations within STR profiles.



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

Template

In theory, peak heights from a single contributor are expected to be approximately proportional to the amount of undegraded DNA template.



Template

The amount of DNA for each contributor to a sample will therefore directly relate to the peak height of contributors.

In practice, there exists some stochastic variation in peak height.

Nowadays, only a couple of picograms of DNA is sufficient to produce results. However, for these *low template DNA* (LTDNA) samples, stochastic effects can play a major role and will invariably influence the analysis (and likely decrease the statistical weight of the evidence).

Degradation

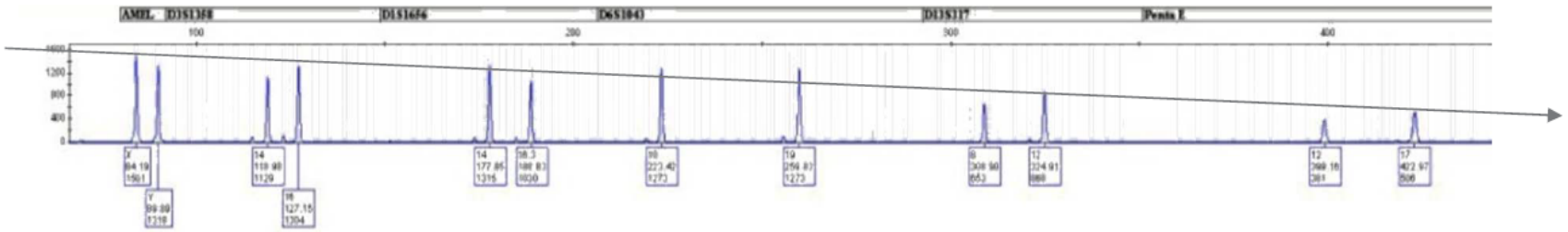
DNA evidence is prone to degradation due to a variety of mechanisms and circumstances, including chemical processes and environmental conditions, causing breakage of previously intact DNA molecules.

If breakage occurs in regions where primers anneal, or between the forward and reverse primers, target regions may not amplify efficiently or fail to amplify at all.

Degradation

Studies suggest that degradation leads to peak heights showing a downward trend with increasing molecular weight, supposedly because smaller alleles are more resistant to degradation.

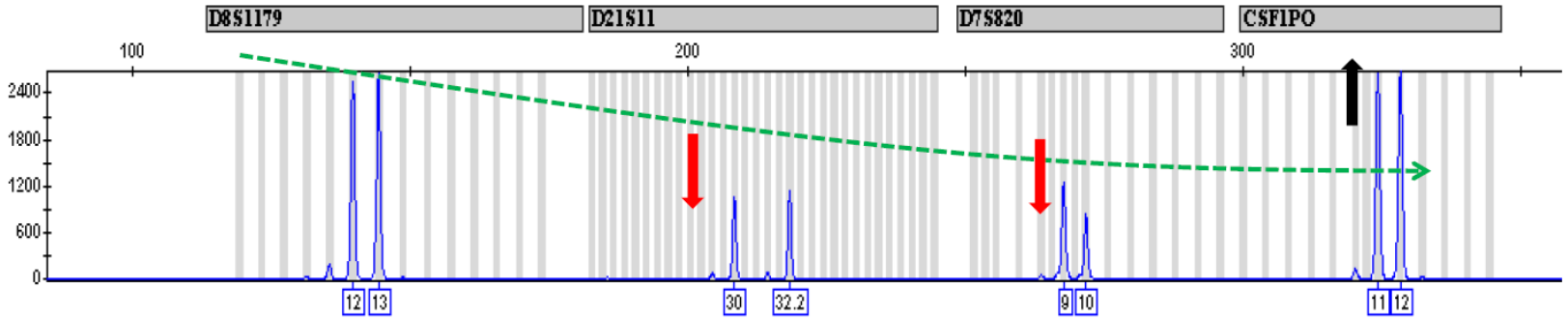
This observation is sometimes referred to as the degradation slope or the ski slope.



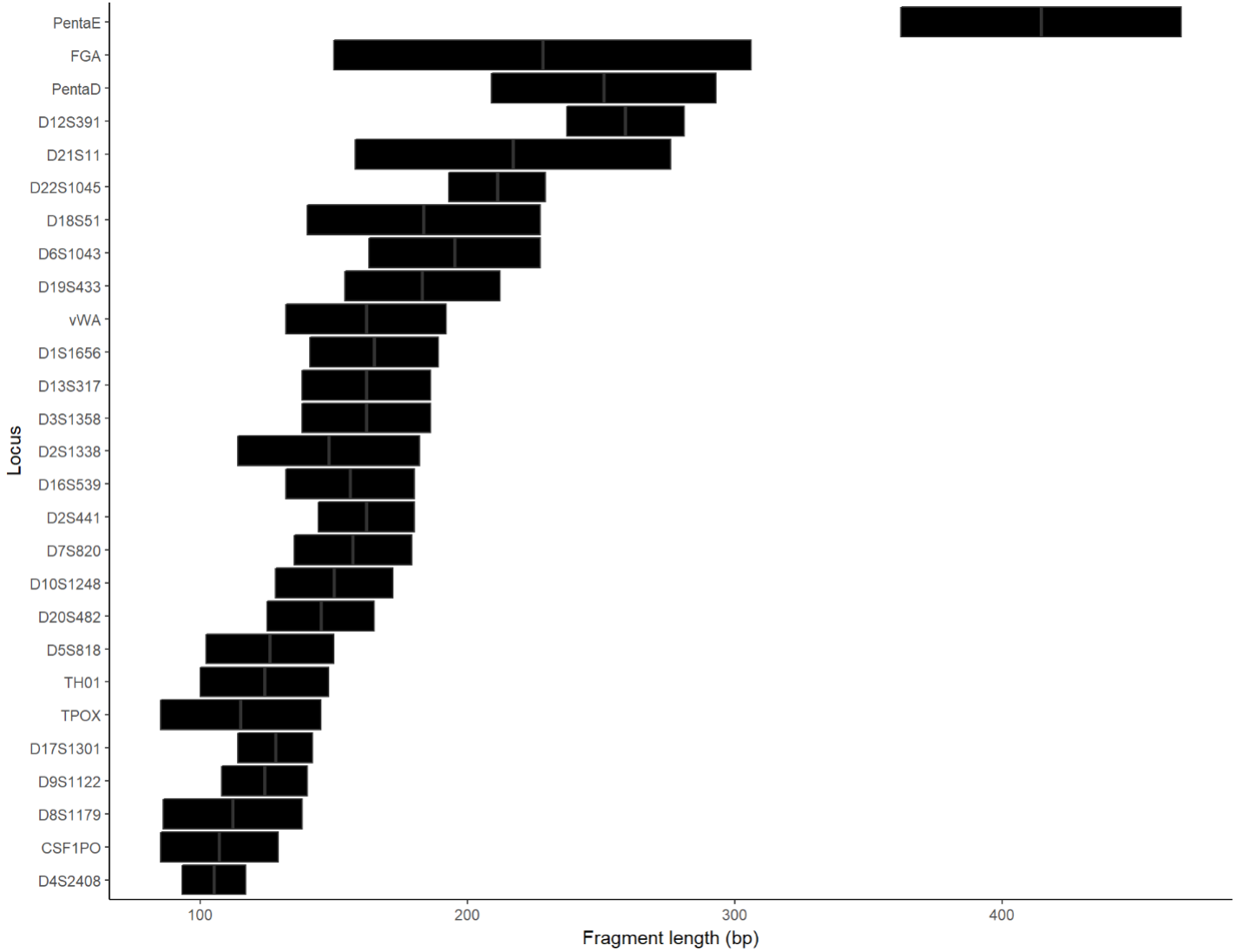
Locus Specific Amplification Efficiency

Additional variability arises from differences in amplification efficiency per locus. Observations show that some loci amplify more efficiently than others, and that these differences appear to vary over time.

Amplification bias is thought to be a result of the large variation in target loci length.



Target Loci Length

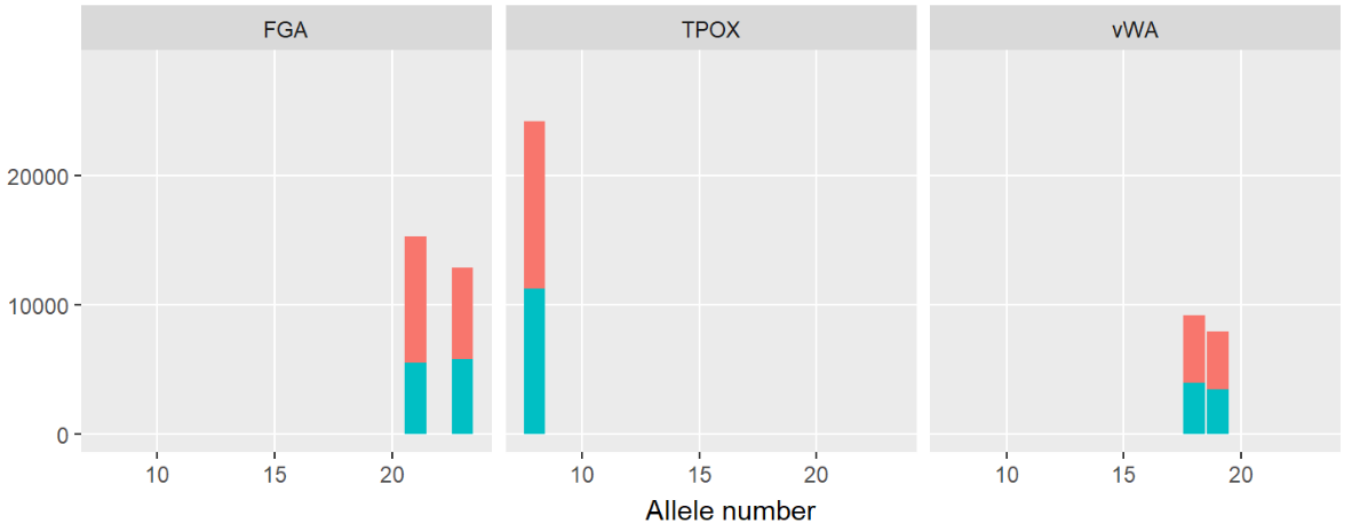


Fragment ranges of NGS products for ForenSeq autosomal loci.

Replicates

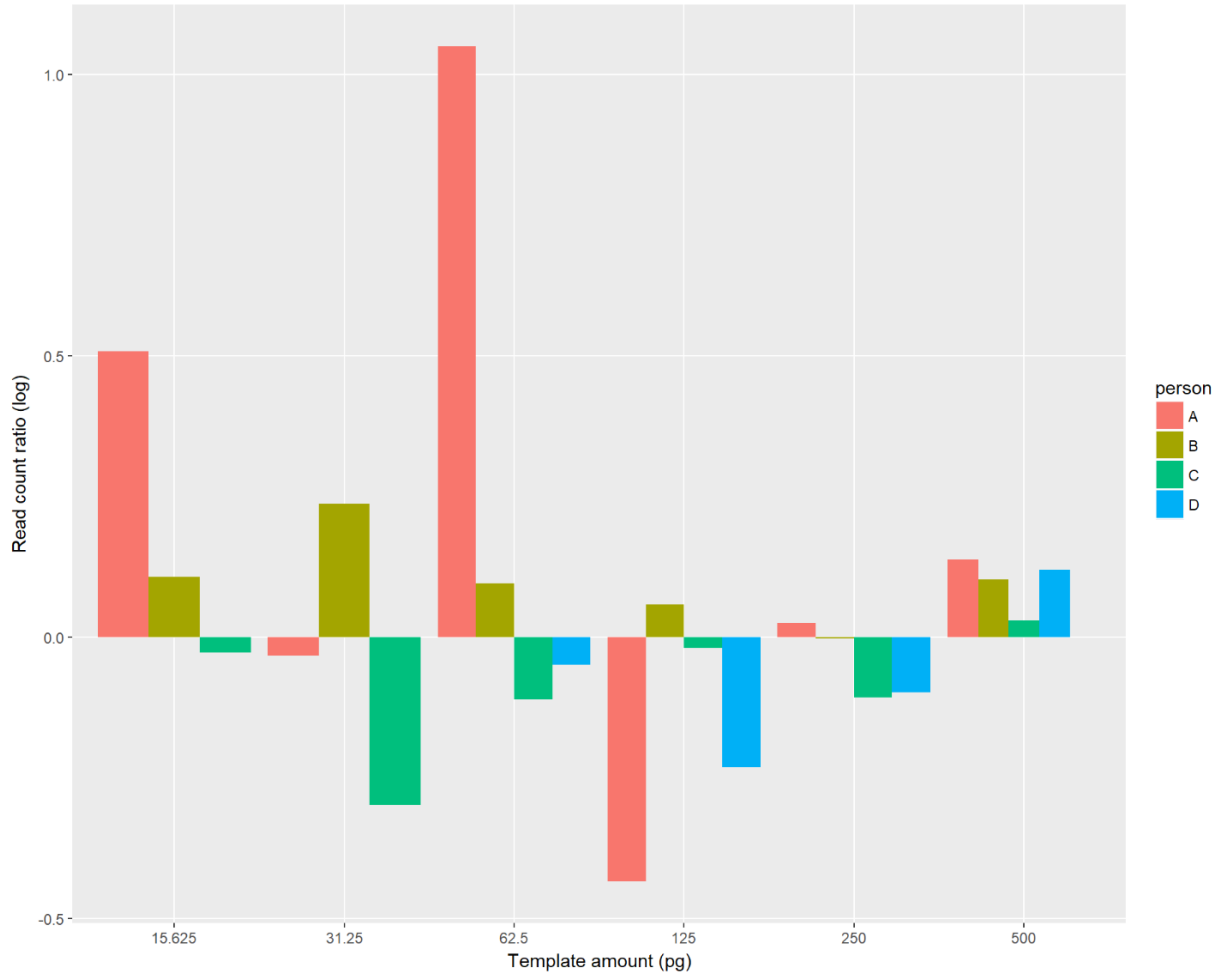
Replicates may show different replicate amplification efficiencies, but can be consolidated into a single analysis, even for different amounts of template DNA. As long as replicates originate from the same DNA extract, they can be used to obtain a more accurate genotype profile.

Replication is not always possible, and in case of a LTDNA sample it would probably be preferable to use as much as possible of the available DNA to give the best possible single-run profile.



Replicate Consistency vs. Template Amount

Higher template amounts result in more balanced peak heights between replicates.



Heterozygote Balance

A consequence of all the stochastic variations that have been introduced into the process, is that the two peaks of heterozygous alleles will also show variability, termed the *heterozygote balance*.

The difference is thought to be affected by the number of repeat sequences, since high molecular weight alleles:

- Stutter more;
- And amplify less.

Heterozygote Balance

Understanding the variability in heterozygous balance is important for the interpretation of mixed profiles and low template DNA:

- For LTDNA, peaks may be so imbalanced that it leads to alleles not exceeding the allelic threshold or even a drop-out.

- It may be used to classify combinations of alleles (or genotypes) as possible or impossible when considering a mixture.

Stutter

Since STR typing methods make use of the PCR process, which relies on DNA replication characteristics, replication slippage also exists during DNA amplification of STRs in vitro.

This phenomenon manifests itself in an epg in the form of a *stutter* peak, i.e. a non-allelic peak that differs in size from the main product, usually by multiples of the length of the repeat unit, appearing adjacent to an allelic peak.

As a consequence, most profiling techniques cannot be used to study in vivo mutational dynamics.

Stutter Characteristics

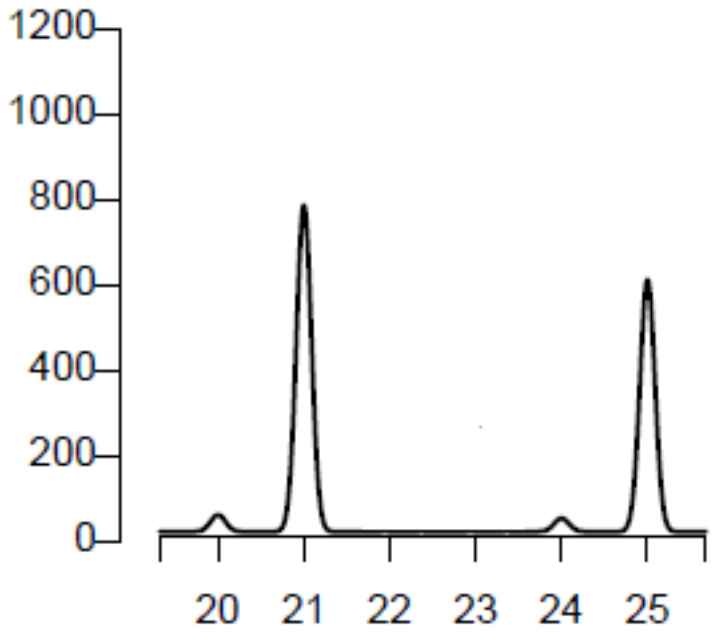
The characteristics shared by mutations and stutter are considerable:

- Rates increase with the number of repeat units (i.e. less stutter for shorter alleles, more stutter for longer alleles);
- Are inversely correlated with repeat unit length (i.e. more stutter for dinucleotide repeats, less stutter for tetranucleotide repeats);
- And typically involve the insertion or deletion of a complete repeat unit.

Stutter Categories

Stutter can be primarily recognized as peaks whose length places them in 'stutter position' of other peaks present within a sample.

- Back stutter
- Forward stutter
- Double back stutter
- Two bp stutter



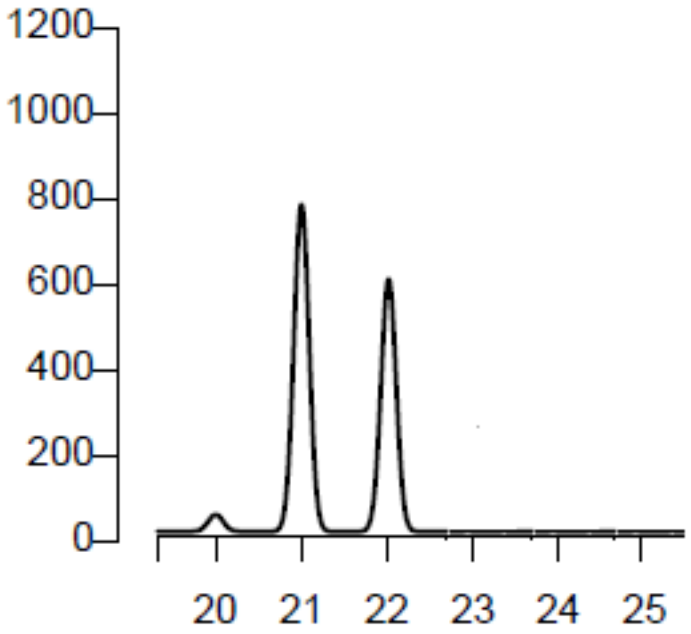
Stutter Difficulties

It is not always possible to distinguish stutter from other molecular artifacts or analyte signal:

- Stutter affected heterozygous genotypes;
- Composite stutter;
- Increase in repeat motif canceled out by a contraction;
- Compound repeats differing one nucleotide in repeat motif.

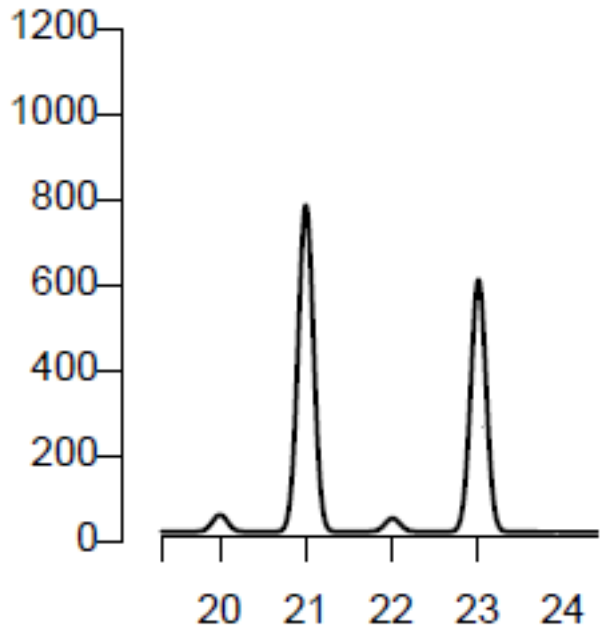
Stutter Affected Heterozygotes

Stutter affected heterozygous genotypes occur when two authentic alleles are separated by one repeat, and the total peak heights are a combination of analyte signal and stutter.



Composite Stutter

Composite stutter arises when the difference between two authentic alleles consist of two repeats and forward stutter of the low molecular weight allele coincides with back stutter of the high molecular weight allele.



Stutter Expansion and Contraction

In rare situations, an increase in repeat motif may cancel out a repeat contraction. This artifact would not be in stutter position and can only be recognized if the expansion and contraction involve different repeats.

$$[\mathrm{TCTA}][\mathrm{TCTG}]_3[\mathrm{TCTA}]_{11} \quad \Leftrightarrow \quad [\mathrm{TCTA}][\mathrm{TCTG}]_2[\mathrm{TCTA}]_{12}$$

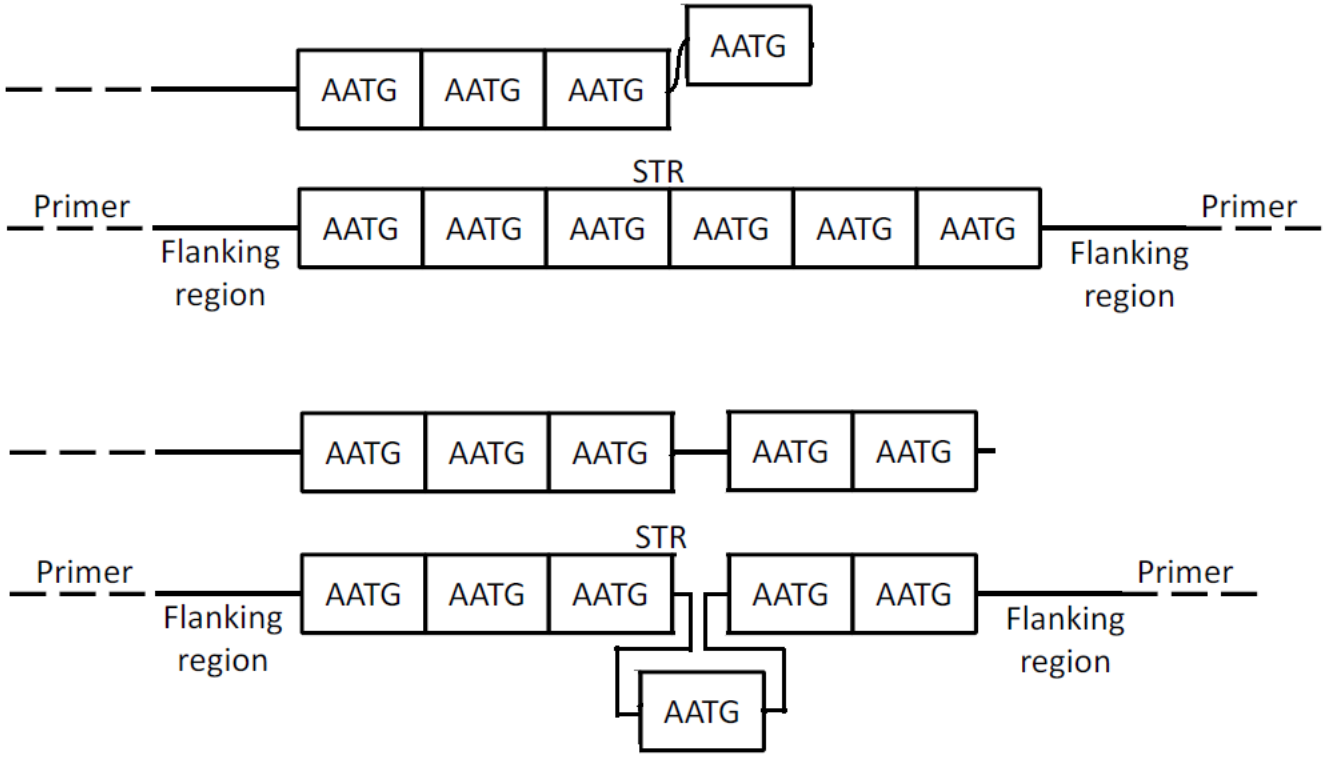
Stutter vs. Substitutions

If adjacent repeats of compound STR loci differ by a single nucleotide and are repeated only once or twice, stutter products can possibly not be distinguished from substitution errors.

$$[\mathrm{TCTA}][\mathrm{TCTG}]_{1-2}[\mathrm{TCTA}]_{13-15} \quad \Rightarrow \quad [\mathrm{TCTA}]_{13-17}$$

Back Stutter

Back stutter is the most prevailing type of stutter, suggesting a preference for repeat contractions over expansions (which are energetically less favorable).



Double Back Stutter

There are two possible mechanisms for the creation of double back stutter:

- Direct creation caused by a double loop during slipped strand mispairing;
- Stutter of a previously formed stutter product.

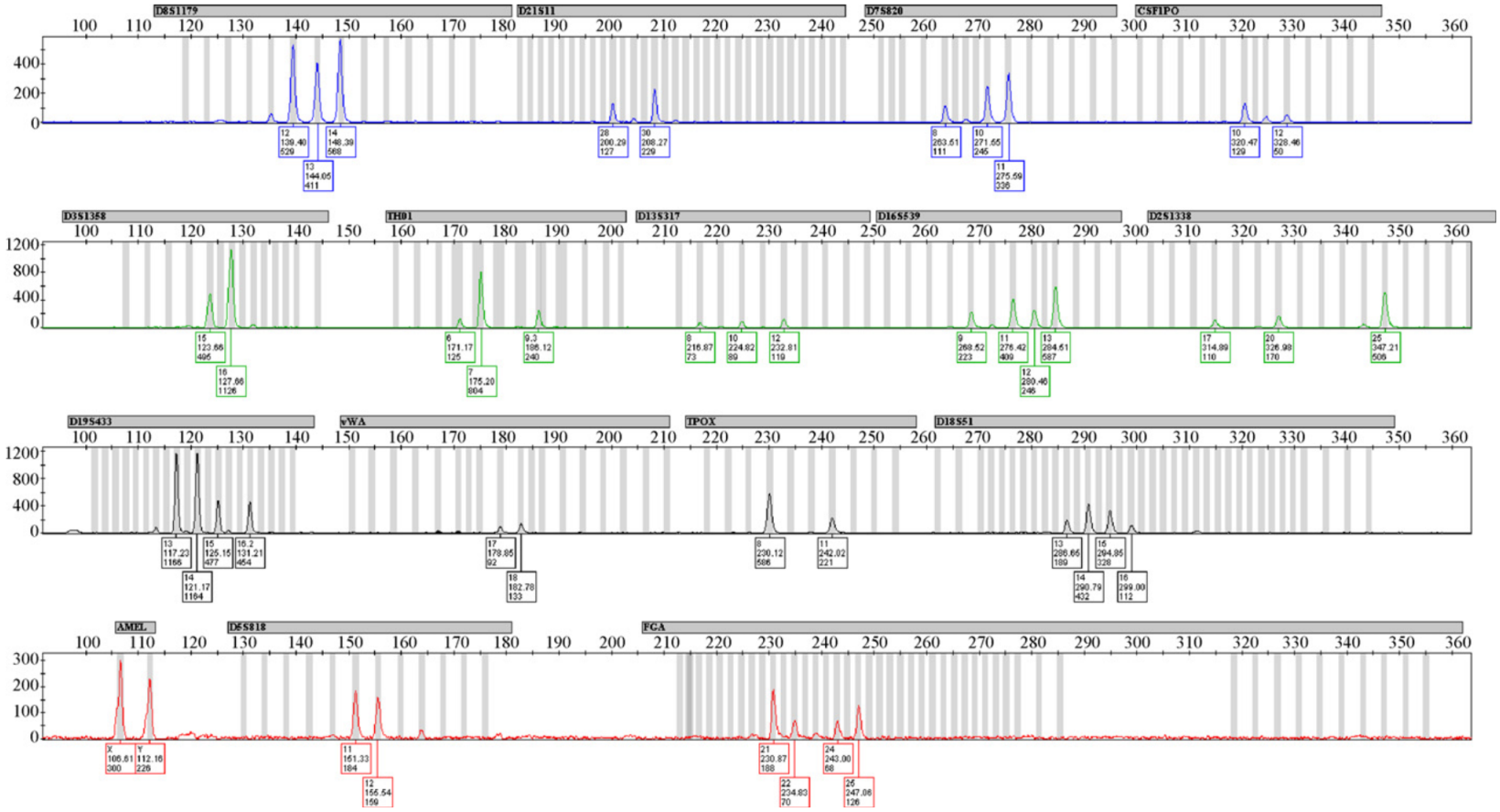
It is suggested that a double loop is more likely than stutter of stutter, at least for Y-STR data.

$$\begin{aligned} [AATG]_6ATG[AATG]_3 &\Rightarrow [AATG]_4ATG[AATG]_3 \\ &\Rightarrow [AATG]_5ATG[AATG]_2 \end{aligned}$$

Source: Modelling PowerPlex Y stutter and artefacts (Bright et al., 2011).

Mixtures

Most forensic stains contain DNA from different individuals. The number of contributors (NoC) is usually unknown.



Source: The interpretation of low level DNA mixtures (Kelly et al., 2012).

Mixtures

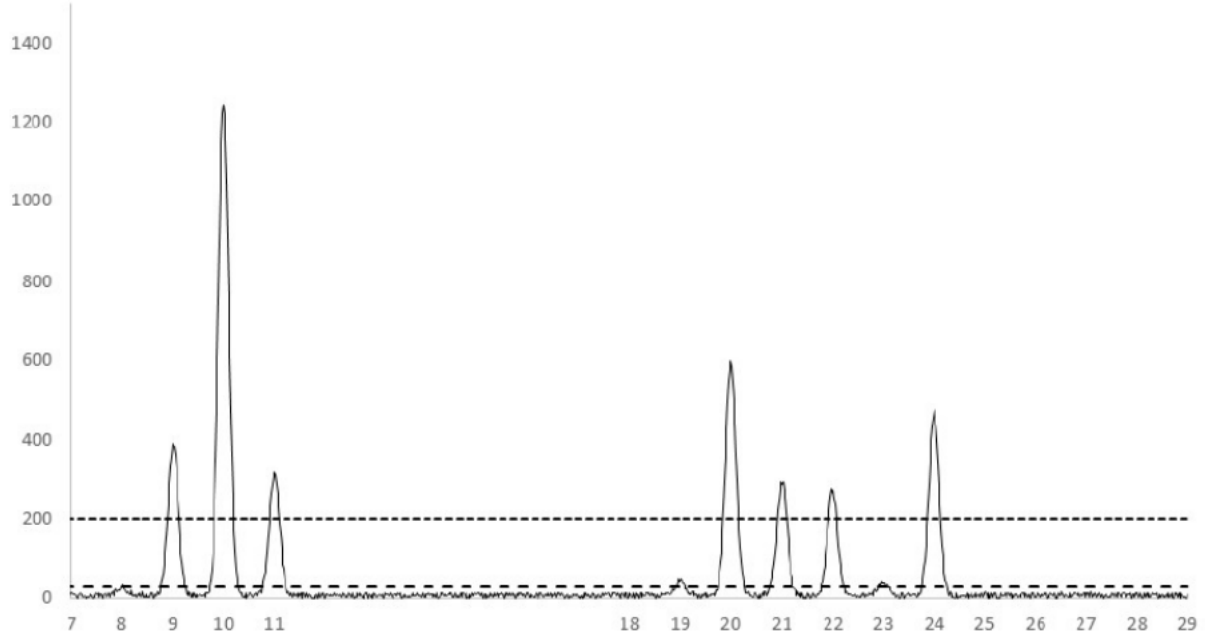
The NoC can technically be any number of contributors. Some guidelines can be used:

- If the number of alleles is known, the maximum allele count (MAC) method can be used to estimate NoC $\geq$ MAC/2.
- If the heterozygote balance for two peaks is outside a certain range they cannot be a heterozygote.
- If a peak is in stutter position it may be classified as stutter if it falls within an expected range.

These may need modification when multiple effects play a role.

Mixtures

Mixing proportions or ratios can range from the contributors being approximately equal to each other to one being in great excess (major vs. minor contributors). Qualitative data cannot distinguish contributors in such case, so a quantitative approach may be preferred if possible.



STR Typing Characteristics

To effectively interpret DNA evidence, phenomena and factors like **mutations, CNVs, contamination, template amount, replicates, amplification efficiency, and degradation** should be considered.

These lead to observations in the form of **stutter, drop-ins, drop-outs, peak height variability and heterozygote balance**, that may need to be incorporated in weight-of-evidence calculations.