Faculty of Health and Medical Sciences

Sequence variation in the short tandem repeat system SE33 discovered by next generation sequencing

Eszter Rockenbauer, MSc, PhD and Line Møller, MSc

Forensic Geneticist
Section of Forensic Genetics
Department of Forensic Medicine
Faculty of Health and Medical Sciences
University of Copenhagen
Denmark

Aims

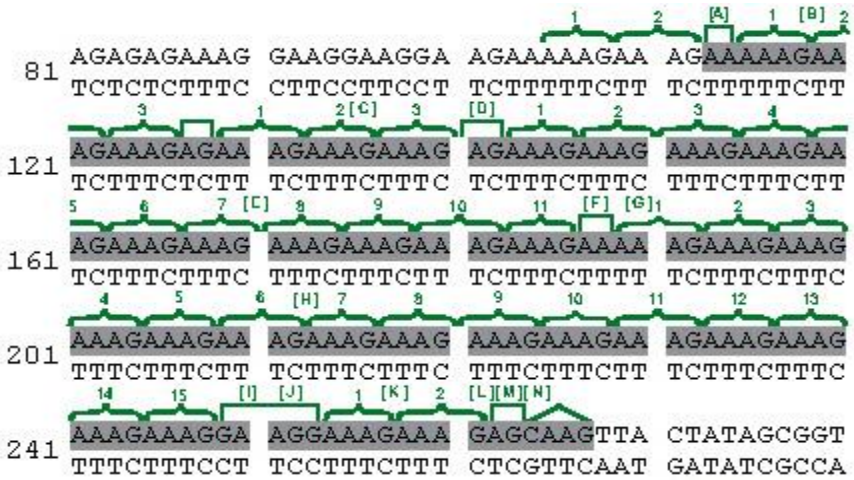
- Sequence SE33 with an NGS assay
- Use Roche Junior Platform
- Search for sequence variation
- Identify new allele variants
- Find population specific alleles
- Test usefulness of higher allele diversity in paternity cases

SE33

- Core locus in the European Standard Set of STRs
- Included in several commercial STR typing kits
  - E.g. NGM Select, ESI17
- Complex and highly variable sequence
- Consisting of four nucleotide AAAG units interrupted by AA and AG units
- Wide span of allele length variation (3 to 39 four repeat units)

Challenges with SE33

- No commercially available NGS kits to sequence SE33
- Several polyA stretches difficult to sequence with presently used NGS techniques
- Alleles can be up to 343 bp ↔ most NGS machines produce ≤ 300 bp reads



GS Junior 454

- Pyrosequencing technology
- 200 cycles of nucleotide addition (A, T, C and G in fixed order)
- Read length ≤ 800 bases
- Accuracy of 99% for reads of ≤ 400bp
- Several successful STR sequencing studies*

*Fordyce et al. 2011, Van Neste et al. 2012, Dalsgaard et al. 2014; Rockenbauer et al. 2014; Gelardi et al. 2014; Scheible et al. 2014

Samples

Samples

203 samples from Danes(144), Somalis(81) and Greenlanders(58)

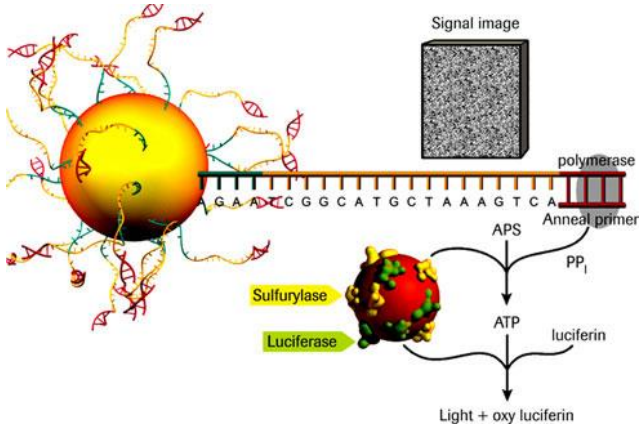
- 188 from unrelated individuals
- 5 trios with an unsolved genetic inconsistency in SE33 between parent and child

Control

- ISO17025 accredited STR typing (AmpFLSTR® NGMSelect® PCR Amplification Kit)

STR sequencing

- Amplicon sequencing with Multiplex IDentifiers (MIDs) in the PCR primer sequence
- Quantification and pooling amplicons into a library
- emPCR and sequencing



Data analysis

- Generally low sequencing quality
- Standard filters were switched off to get more reads in output file
- "Filters_off_amplicons"-pipeline*:
  - Sorting by MIDs
  - Filtering by flanking sequences
- Both flanking sequence must be present (optional)
- Alignment in BioEdit (Ibis Biosciences)

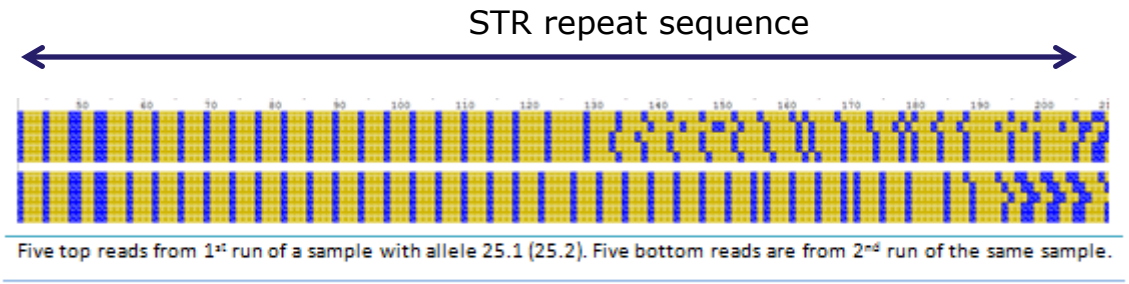
*in-house Python-based algorithm

Results of allele calling

- Total sample coverage 4,643 (before) and 318 (after filtering)
- 10-20% forward 80-90% reverse strand reads
- 57 samples were sequenced twice
- Low-quality samples were reanalysed with only one flanking sequence
- 295 correct allele calls (from 394 expected alleles)
- ~20% called with an insecurity of +/-1bp
- Most correctly called alleles had ≤ 30 repeats (~300bp incl. flanks)

Results of allele calling

- Uncertain calls:
  - Only few reads for long alleles (>30 repeats)
  - Uncertain number of ”A’s” (”T” in reverse strand read)
- Undetermined alleles:
  - lack of sequencing reads
  - inconsistent reads
  - reads deviating from PCR-CE



Five top reads from 1st run of a sample with allele 25.1 (25.2). Five bottom reads are from 2nd run of the same sample.

Results of allele calling

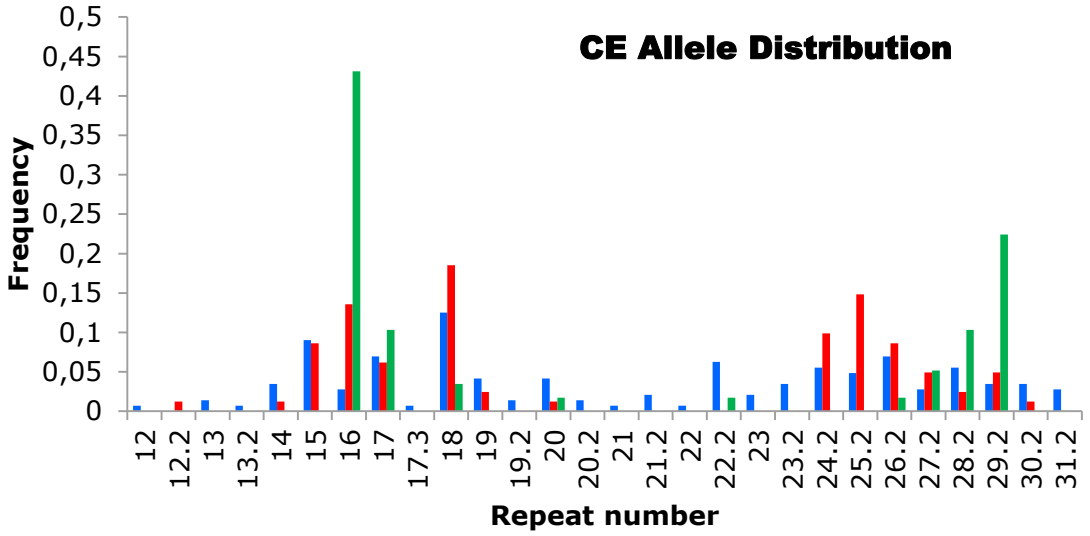
- High increase in allele variation
- 27 novel alleles not reported in STRbase
- Raised the power of discrimination by 282% compared to CE
- Identified the mutated allele in 3 out of 5 family trios

Novel alleles

Table 19 – New Alleles		
1	SE33[13]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{13}$G[AAAG]$_3$AG
2	SE33[16]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{16}$G[AAAG]$_3$AG (4bp del in flanking region)
3	SE33[17]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{17}$G[AAAG]$_3$AG (4bp del in flanking region)
4	SE33[17.3]	[AAAG]$_2$AG[AAAG]$_3$---[AAAG]$_{15}$G[AAAG]$_3$G[AAAG]$_3$AG
5	SE33[18]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{17}$GAAGG[AAAG]$_3$AG
6	SE33[21.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_8$AA[AAAG]$_{13}$GAAGG[AAAG]$_2$AG
7	SE33[21.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{12}$AA[AAAG]$_9$GAAGG[AAAG]$_2$AG
8	SE33[23]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{23}$G[AAAG]$_3$AG
9	SE33[24.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$A**A**[AAAG]$_{15}$GAAGG[AAAG]$_2$AG
10	SE33[25.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_7$A**A**[AAAG]$_{18}$GAAGG[AAAG]$_2$AG
11	SE33[25.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_8$AA[AAAG]$_{17}$GAAGG[AAAG]$_2$AG
12	SE33[25.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{13}$A**A**[AAAG]$_{12}$GAAGG[AAAG]$_2$AG
13	SE33[26.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_7$A**A**[AAAG]$_{19}$GAAGG[AAAG]$_2$AG
14	SE33[26.2]	[AAAG]$_2$AG[AAAG]$_3$AAG[AAAG]$_{10}$A[AAAG]$_{17}$GG----[AAAG]$_2$AG
15	SE33[26.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{12}$A**A**[AAAG]$_{14}$GAAGG[AAAG]$_2$AG
16	SE33[27.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_5$A[AAAG]$_{20}$[GAAG]$_3$G[AAAG]$_2$AG
17	SE33[27.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_5$A[AAAG]$_{22}$GAAGG[AAAG]$_2$AG
18	SE33[27.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_7$A**A**[AAAG]$_{20}$GAAGG[AAAG]$_2$AG
19	SE33[27.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$A**A**[AAAG]$_{18}$GAAGG[AAAG]$_2$AG
20	SE33[28.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$AA[AAAG]$_{17}$G[AAGG]$_3$[AAAG]$_2$AG
21	SE33[29.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$A[AAAG]$_{19}$G[AAGG]$_2$[AAAG]$_2$AG
22	SE33[29.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{10}$A**A**[AAAG]$_{19}$GAAGG[AAAG]$_2$AG
23	SE33[30.2]	**A**AAG]$_2$AG[AAAG]$_3$AG[AAAG]$_8$AA[AAAG]$_{22}$GAAGG[AAAG]$_2$AG
24	SE33[31.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{11}$AA[AAAG]$_{20}$GAAGG[AAAG]$_2$ (4bp del in flanking region)
25	SE33[30.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{13}$A[AAAG]$_{18}$G----[AAAG]$_2$AG
26	SE33[31.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_4$A[AAAG]$_{10}$A[AAAG]$_{17}$GAAGG[AAAG]$_2$AG
27	SE33[31.2]	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_8$AG[AAAG]$_{23}$GAAGG[AAAG]$_2$AG

Bold black bases are missing in some reads.
Bold red bases are appear duplicated in some reads.

Allele distribution



Number of called alleles:

DK	144
SOM	81
GRL	58

PCR-CE / NGS

No. of unique alleles

	Danes	Somali	Greenlanders
Danes	12 / **43**	27 / **61**	
Somali	1 / **10**	14 / **14**	15 / **28**
Greenlanders	0 / **4**	9 / **9**	8 / **7**	9 / **15**

Mutation studies

	Rolle	NGS Allele sequence	CE length
1	M	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{11}$AA[AAAG]$_{10}$GAAGG[AAAG]$_2$	21.2
1	M	No usable sequence information obtained	(31.2)
1	C	[AAAG]$_2$A[AAAG]$_3$AAG[AAAG]$_3$AAG[AAAG]$_{19}$G[AAAG]$_3$	19
1	C	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AAG[AAAG]$_{11}$A$_x$[AAAG]$_{10}$GAAGG[AAAG]$_2$	21.2
1	F	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{18}$G[AAAG]$_3$	18
1	F	[AAAG]$_2$A[AAAG]$_3$AAG[AAAG]$_3$AAG[AAAG]$_{20}$G[AAAG]$_3$	20
2	M	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{17}$G[AAAG]$_3$	17
2	M	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{18}$G[AAAG]$_3$	18
2	C	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{17}$G[AAAG]$_3$	17
2	C	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{16}$GAAGG[AAAG]$_3$	17
2	F	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{17}$GAAGG[AAAG]$_3$	18
2	F	No usable sequence information obtained	(20)
5	M	[AAAG]$_2$A$_x$[AAAG]$_3$A$_x$G[AAAG]$_3$A$_x$G[AAAG]$_{11}$A$_x$[AAAG]$_{18}$GAAGG[AAAG]$_2$	29.2
5	M	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{13}$A[AAAG]$_{18}$G[AAAG]$_2$	30.2
5	C	[AAAG]$_2$A$_x$[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{13}$A$_x$[AAAG]$_{15}$GAAGG[AAAG]$_2$	28.2
5	C	[AAAG]$_2$A[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{13}$A[AAAG]$_{18}$G[AAAG]$_2$	30.2
5	F	[AAAG]$_2$A[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{11}$A[AAAG]$_{20}$G[AAAG]$_2$	30.2
5	F	[AAAG]$_2$A$_x$[AAAG]$_3$AG[AAAG]$_3$AG[AAAG]$_{14}$AA[AAAG]$_{15}$GAAGG[AAAG]$_2$	29.2

M: mother, F: father, C: child. ( ) Alleles with no sequence information were not sequenced correctly in comparison to CE results and were left out of analysis.

Conclusions

- SE33 is a promising candidate locus for NGS based genotyping
- Large sequence variation is observed
- Difficulties owing to poly-A stretches and long alleles
- The Roche Junior platform does not perform optimally
- Similar difficulties expected on other available NGS platforms
- Alternative sequencing technique/s and platform/s are needed

Acknowledgements

Professor and Director of the department:
Niels Morling

Forensic geneticists:
Claus Børsting

Bioinformatics:
Carina G. Jønck

Lab:
Nadia Jochumsen

Thank you for your attention