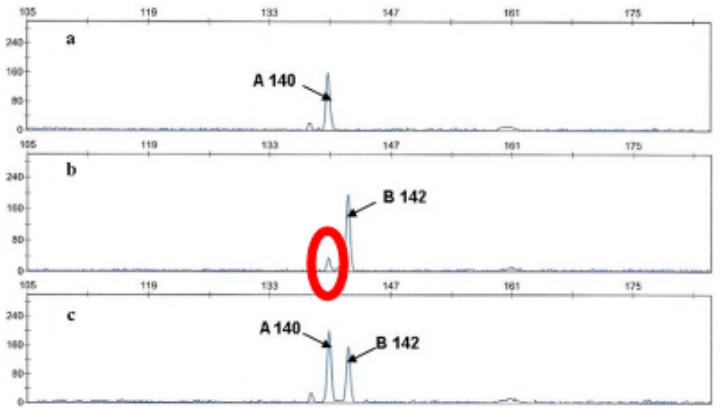
Modeling PCR stutter noise for accurate calling of STRs from short reads

Melissa Gymrek

February 7, 2013

PCR stutter noise complicates STR calling from short reads

Capillary electrophoresis



Sequence reads

ATGTCAACAACAACAACAACAATTGA
ATGTCAACAACAACAACAACAATTGA
ATGTCAACAACAACAACAACAATTGA
ATGTCAACAACAACAACAA***TTGA
ATGTCAACAACAACAACAA***TTGA
ATGTCAACAACAACAA******TTGA

Aims

- Build a model of PCR stutter noise at short tandem repeats (2-6bp motifs)
- Generate likelihood scores for each possible allelotype
- Output information about allelotyping calls in a format (VCF) that is consistent with other bioinformatic tools

Outline

Building a model of PCR stutter noise

Goal: generate likelihood scores for each possible allelotype $\langle A, B \rangle$ given a set of reads $\vec{R}$ where each read shows evidence of a certain allele r_i:

Reference ACAGTCGATCGCAGCAGCAGCAGCAGGATGATCGATCGTAG

Reads

```
GTCGATCGCAGCAGCAGCAG***GATGATCCGT
 CGATCGCAGCAGCAGCAG***GATGATCCGTAG
 CGATCGCAGCAGCAGCAGCAGGATGATCCGTAG
    TCGCAGCAGCAGCAGCAGGATGATCCGTAGTA
     CGCAGCAGCAG******GATGATCCGTAGTAG
```

$$\vec{R} = \{-6, -3, -3, 0, 0\}$$

Building a model of PCR stutter noise

Goal: generate likelihood scores for each possible allelotype $\langle A, B \rangle$ given a set of reads $\vec{R}$ where each read shows evidence of a certain allele r_i:

$$\log P(\vec{R}|\langle A, B\rangle; \theta) = \sum_{i=1}^{|\vec{R}|} \log P(r_i|\langle A, B\rangle; \theta) \quad (1)$$

Assume a read has equal probability of coming from either allele:

$$P(r|\langle A, B\rangle) = \frac{1}{2}[P(r|A) + P(r|B)] \quad (2)$$

Building a model of PCR stutter noise

Model two aspects of stutter noise:

- Probability s_j a read from locus j results from stutter.
- Distribution of error lengths. $D(e)$: probability of error length e given that there is stutter noise.

Calculate the likelihood of each read:

$$P(r|A;\theta) = \begin{array}{ll} 1 - s_j & r = A \\ s_j D(r - A) & r \neq A \end{array} \quad (3)$$

Train stutter model on hemizygous male sex chromosomes

Use chrX and chrY loci with confident calls:

- At least 10x coverage
- At least 50% of reads agree with the modal allele.

Reads not matching the modal allele are assumed to be stutter errors.

Use 22 30x male genomes sequenced with Illumina 100bp PE reads. Obtained 167,154 reads from 10,674 loci for training. 0.56% of 10x loci were rejected for being too messy.

Stutter noise depends on characteristics of the STR locus

These characteristics were found to affect stutter probability:

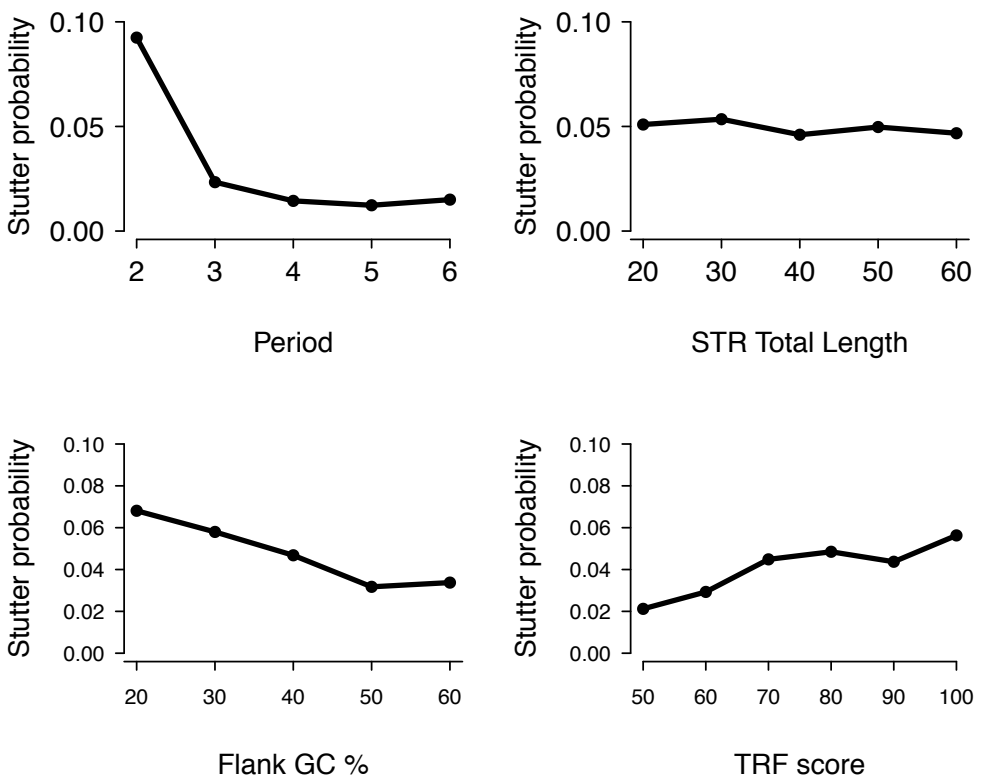
- Motif length
- STR region length
- GC content of flanking regions
- STR purity

Below we use these four features to predict s_j for each locus j.

Example training data for a single locus

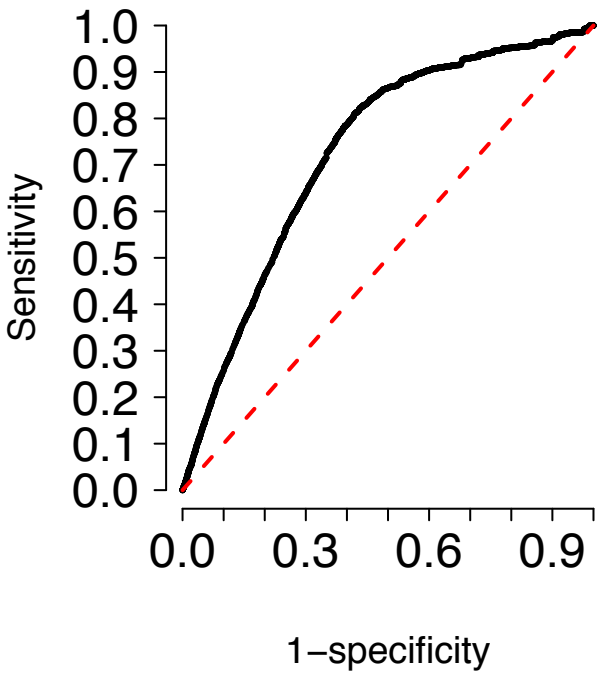
	step.size	stutter	period	gc	score	length
1004	0	0	3	0.36	1	33.9
1005	0	0	3	0.36	1	33.9
1006	0	0	3	0.36	1	33.9
1007	0	0	3	0.36	1	33.9
1008	0	0	3	0.36	1	33.9
1009	0	0	3	0.36	1	33.9
1010	0	0	3	0.36	1	33.9
1011	0	0	3	0.36	1	33.9
1012	0	0	3	0.36	1	33.9
1013	-3	1	3	0.36	1	33.9

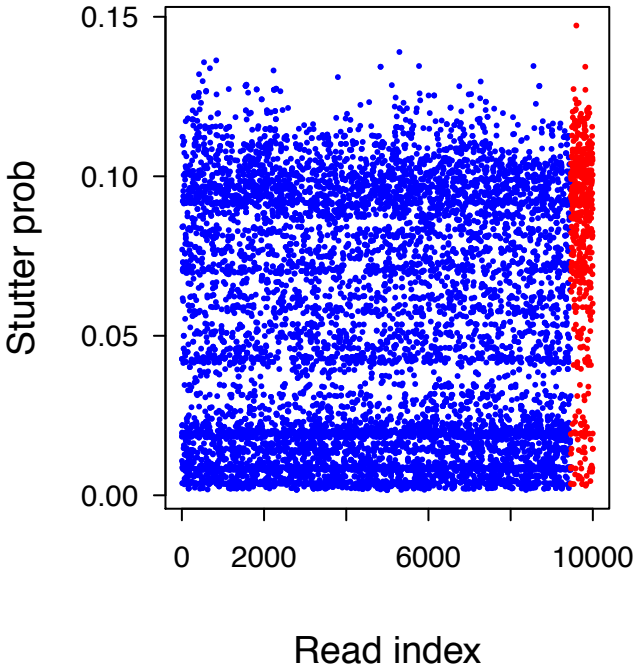
Factors influencing stutter noise



Logistic regression to model stutter probability

$$\text{Stutter} \sim \text{Period} + \text{Score} + \text{GC} + \text{Length}$$





Length distribution of stutter errors at STRs

Goal: What is the distribution of lengths of stutter errors?

Most errors are a multiple of the unit size (**unit errors**):

```
ACAGTCAGCTATCGACTCAGCAGCAGCAGCAGCAGCACTGATC
ACAGTCAGCTATCGACTCAGCAGCAGCAGCAG***CACTGATC
```

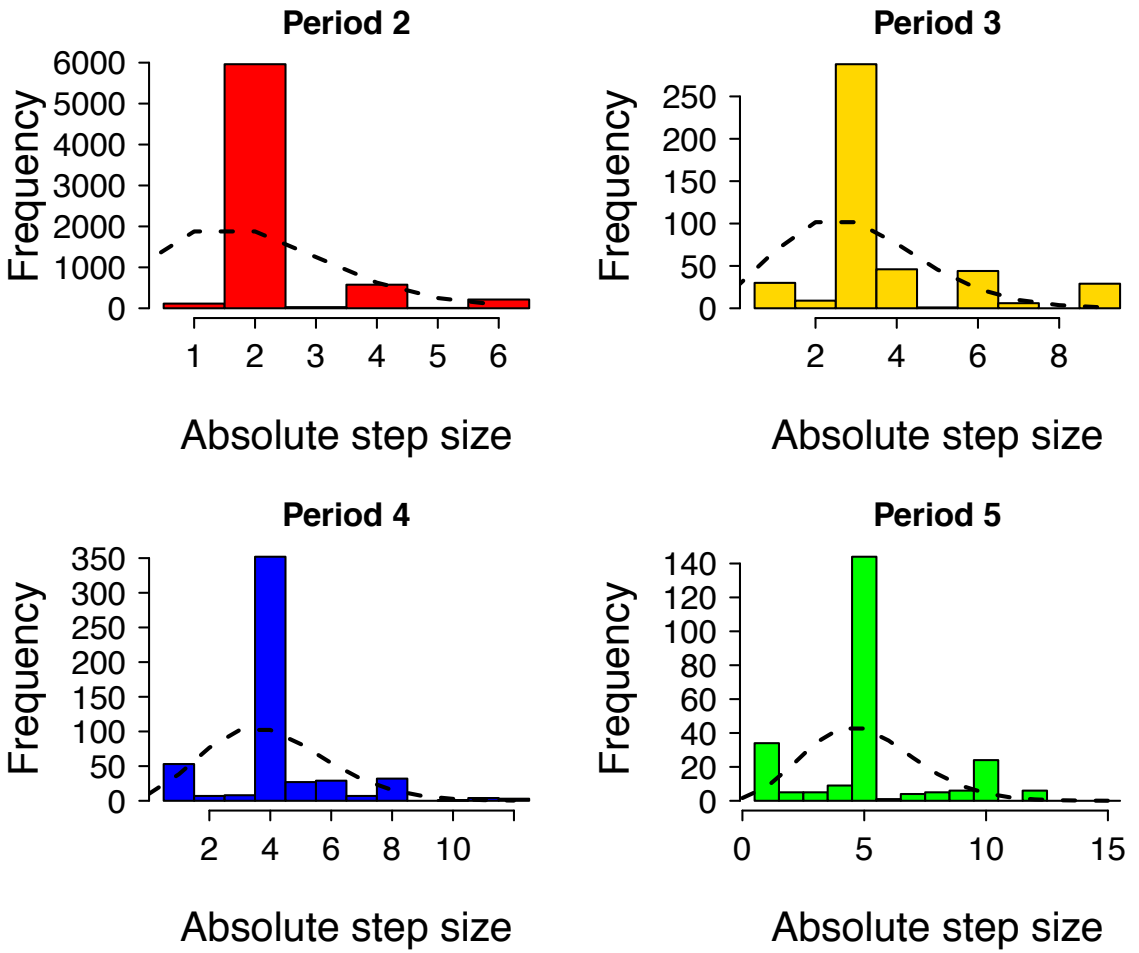
A small number of errors are not (**non-unit errors**):

```
ACAGTCAGCTATCGACTAAATAAATAAATAAATACGACTTACG
ACAGTCAGCTATCGACTAAATAAATAAATAA*TACGACTTACG
```

$D(e, m)$: probability to see stutter error of length e at a locus with period m.

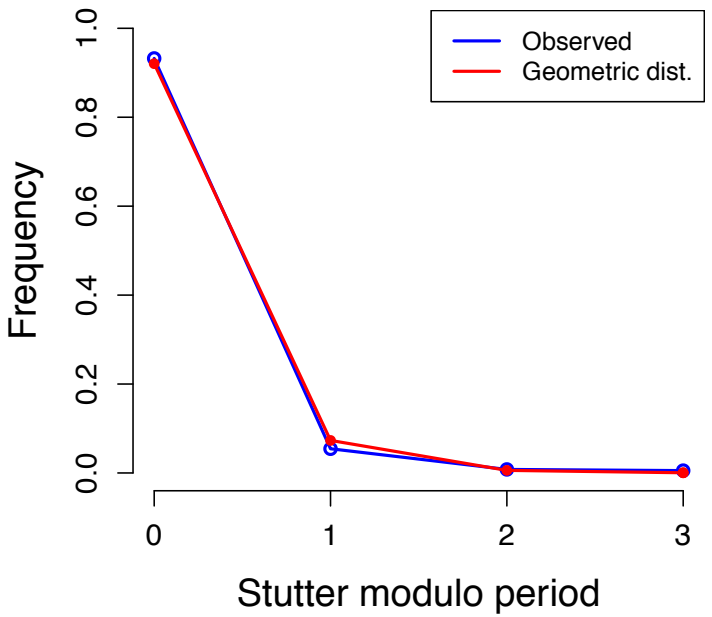
Distribution of step size for unit errors

Unit errors are roughly Poisson with $\lambda = m$ (motif length).



Distribution of step size for non-unit errors

Non-unit errors modulo the unit size follow a geometric distribution with $p = \frac{1}{\overline{x}+1}$ where $\overline{x}$ is the average step size modulo the period.



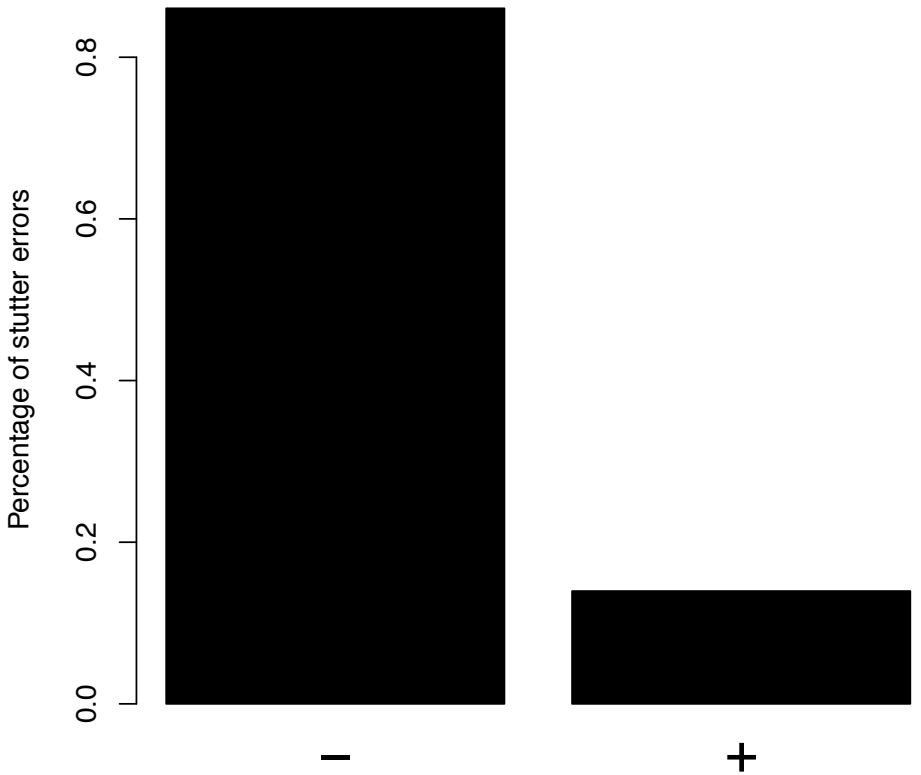
True allele

TCGAAAATAAAATAAAATGATGC

Error%period

0 TCG*****AAAATAAAATGATGC
1 TCG*AAATAAAATAAAATGATGC
2 TCG**AATAAAATAAAATGATGC

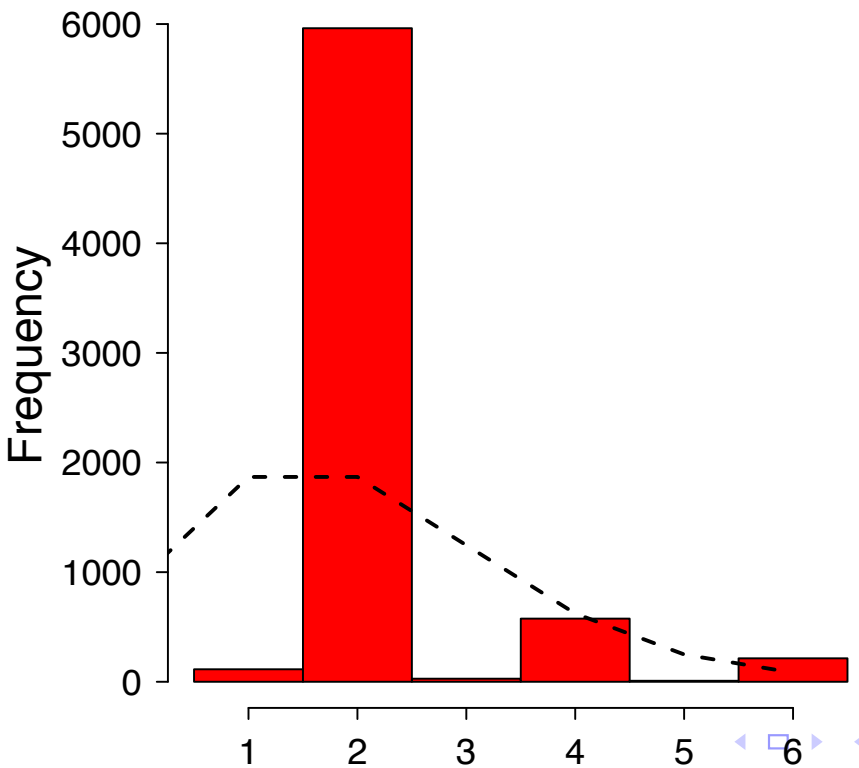
Stutter noise tends to delete repeat units



Constructing the step size PDF

1. Poisson distribution of unit step sizes:

$$D(e, m) = Pois(\lambda = m) \tag{4}$$

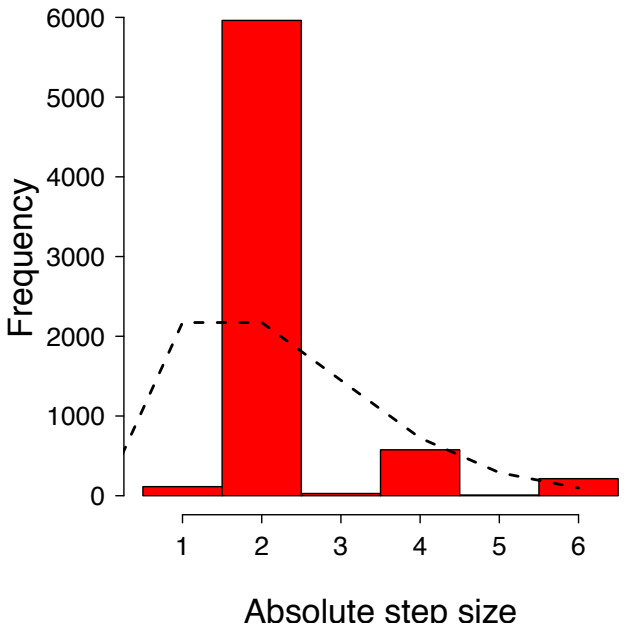


Constructing the step size PDF

2. Set D(0,m) = 0:

$$D(e, m) = Pois(\lambda = m) * Z \tag{5}$$

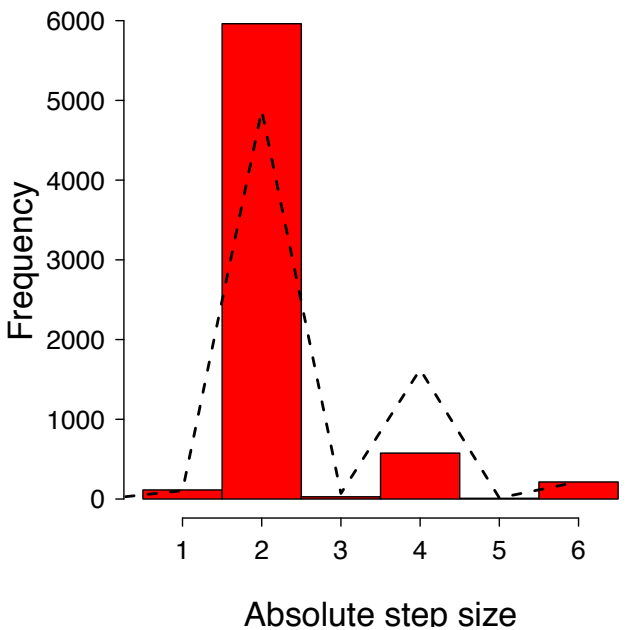
where Z is an indicator variable that is 1 if $e \neq 0$, else 0.



Constructing the step size PDF

3. Geometric distribution of non-unit step sizes modulo m:

$$D(e, m) = Pois(\lambda = m) * Z * Geom(p = \frac{1}{\overline{x} + 1}) \qquad (6)$$

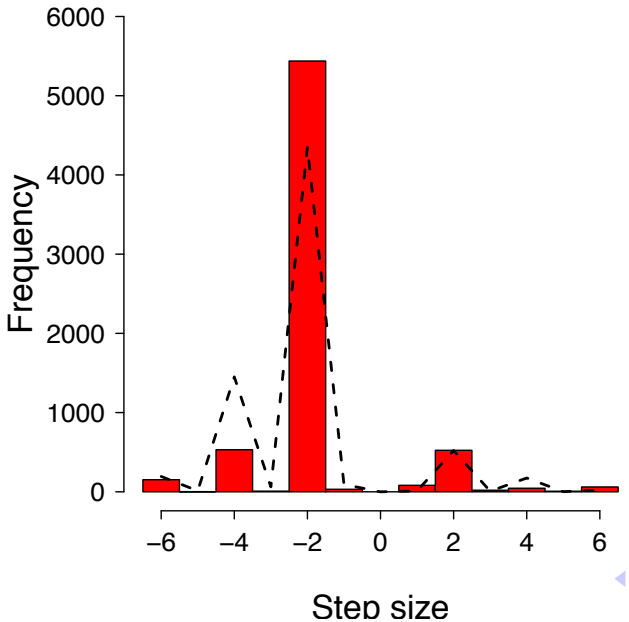


Constructing the step size PDF

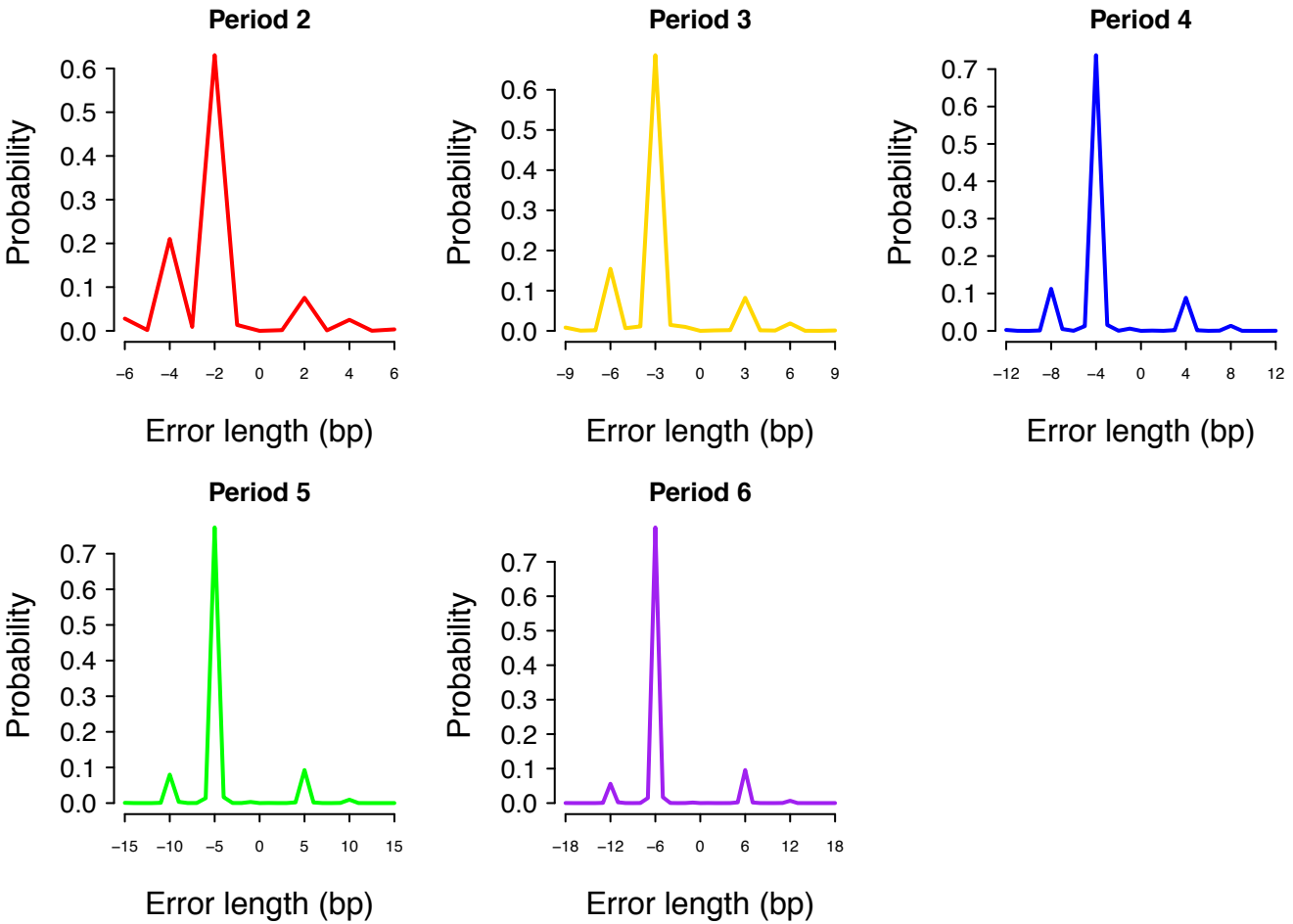
4. Account for increase vs. decrease in length.

$$D(e,m) = Pois(\lambda = m) * Z * Geom(p = \frac{1}{\overline{x}+1}) * (e > 0?q : 1-q) \quad (7)$$

where q is the probability that stutter will increase the true allele length.



Stutter error length PDFs for each motif length



Training the stutter model using lobSTR

```
lobSTR --command train --bam test.bam --haploid
chrX,chrY --noise_model test --strinfo
strinfo.hg19.tab
```

- Build logistic regression model to predict the probability of stutter noise at each locus, s_j.
- Estimates $\overline{x}$ and q and uses these to generate the error length PDF for each motif size.

Scoring STR allelotypes

Goal: determine the most likely allelotype at each STR locus and determine confidence scores.

$$\arg\max_{\langle A,B\rangle} \log P(\vec{R}|\langle A,B\rangle;\theta) = \sum_{i=1}^{|\vec{R}|} \log \frac{1}{2}[P(r_i|A;\theta) + P(r_i|B;\theta)] \tag{8}$$

where as stated above:

$$P(r|A;\theta) = \begin{array}{ll} 1-s_j & r = A \\ s_j D(e, m_j) & r \neq A \end{array} \tag{9}$$

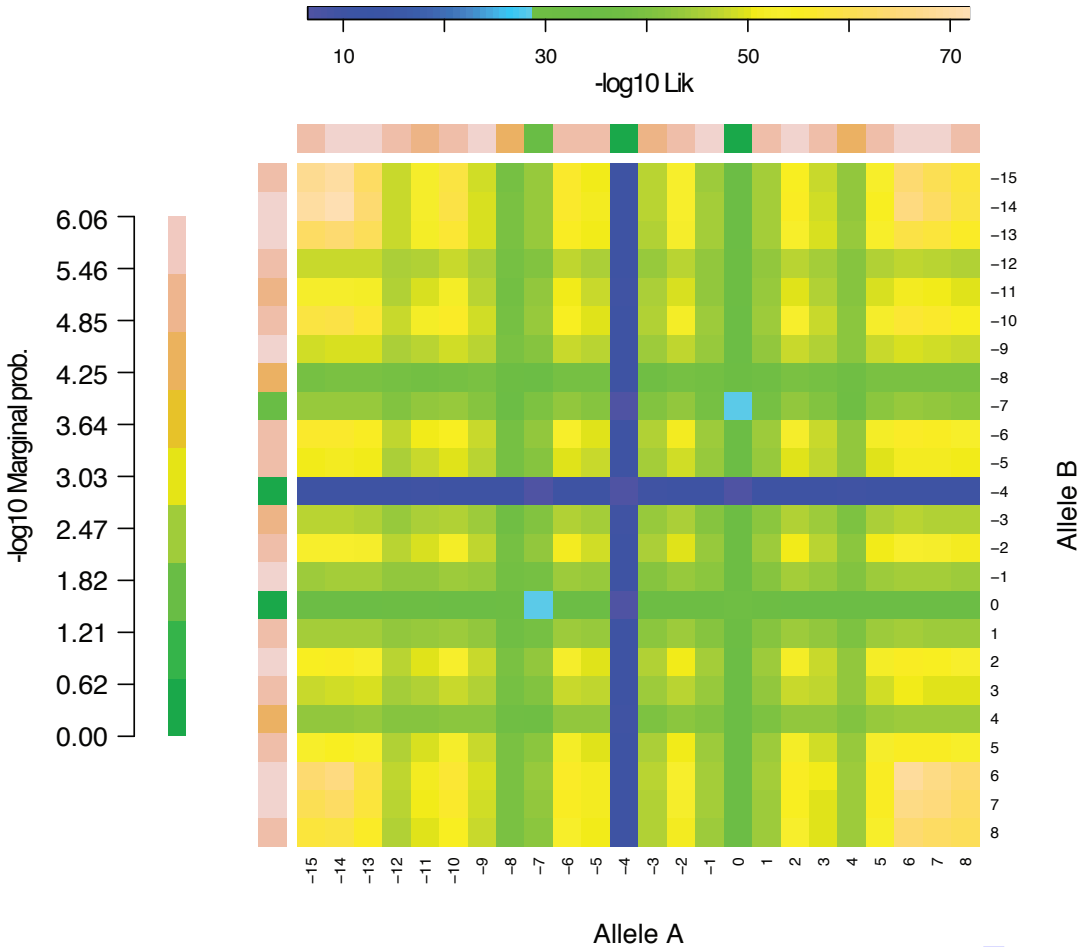
Method outline: perform a grid search over possible allelotypes $\langle A, B\rangle$. Determine the maximum likelihood allelotype given the observed reads.

Scoring STR allelotypes

- Determine likelihood for each possible allelotype.
  - Return maximum likelihood allelotype $\langle A, B \rangle$.
  - Return likelihood and confidence score of $L(\langle A, B \rangle) / \sum_{\langle A', B' \rangle} L(\langle A', B' \rangle)$
  - Return marginal likelihood score for each allele
- If prior allele frequencies known, give posterior probability for each allelotype.

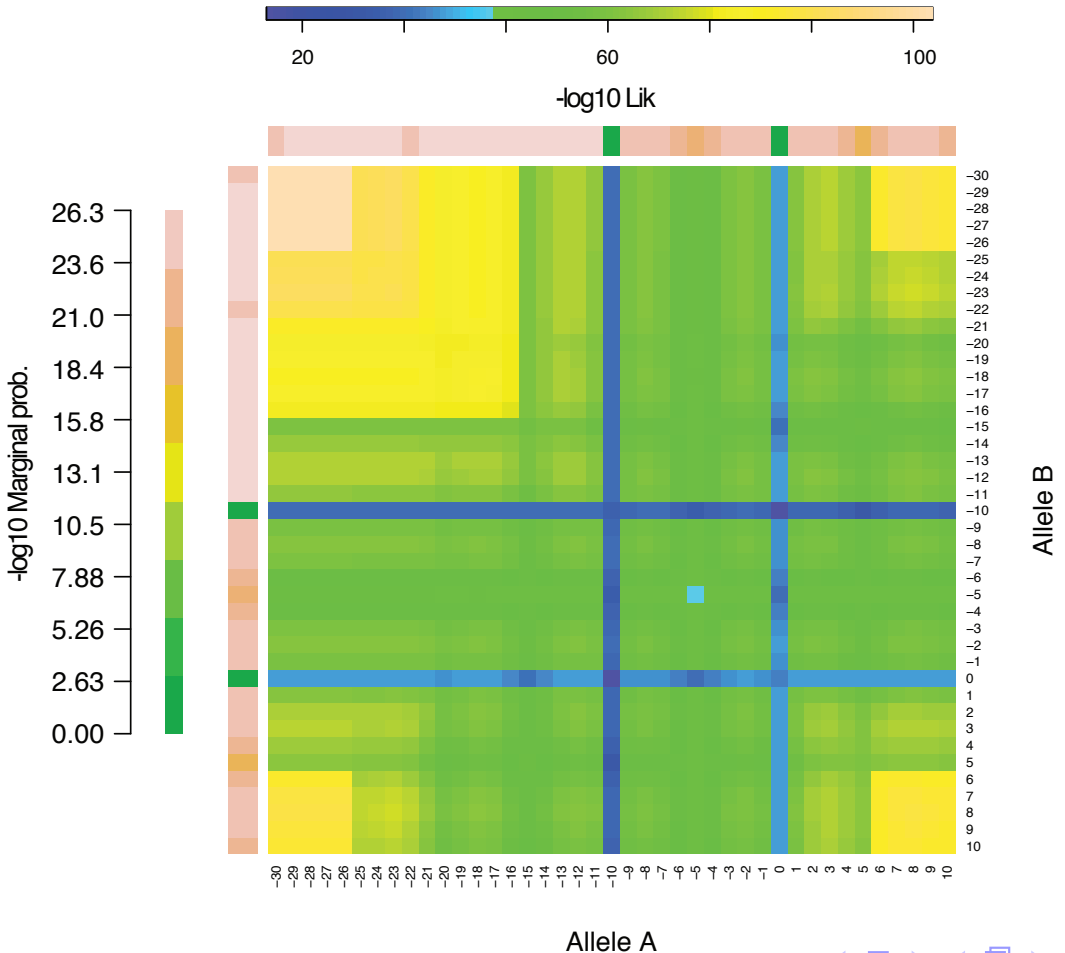
Example homozygous locus

Period 4, Reads: (-7,-4,-4,-4,-4,-4,-4,-4,-4,-4,-4) **Call**: -4,-4



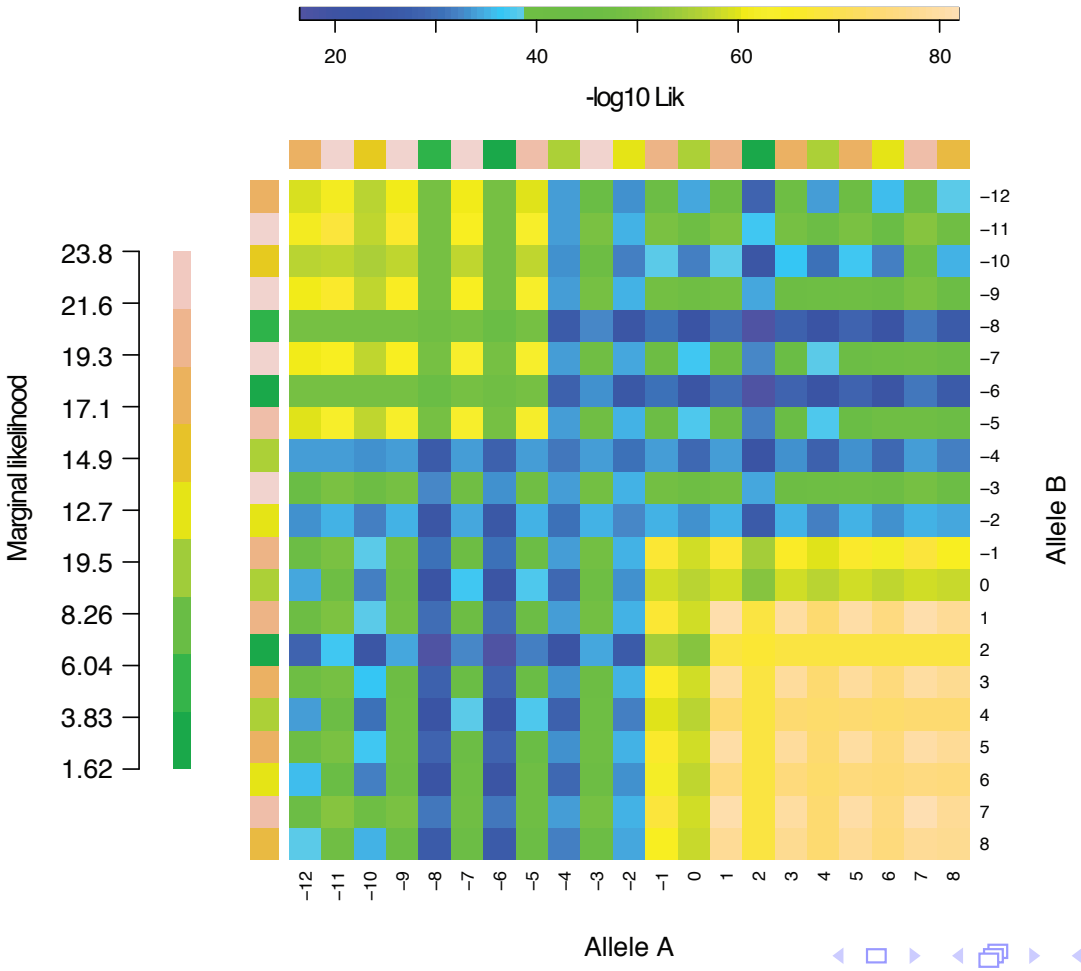
Example heterozygous locus

Period 5, Reads: (-10,-10,-10,-10,-10,-10,0,0,0,0) **Call**: -10,0



Example messy locus

Period 2, reads (-8,-8,-8,-6,-6,-6,2,2,2,2) **Call**: -6,2



TODO list

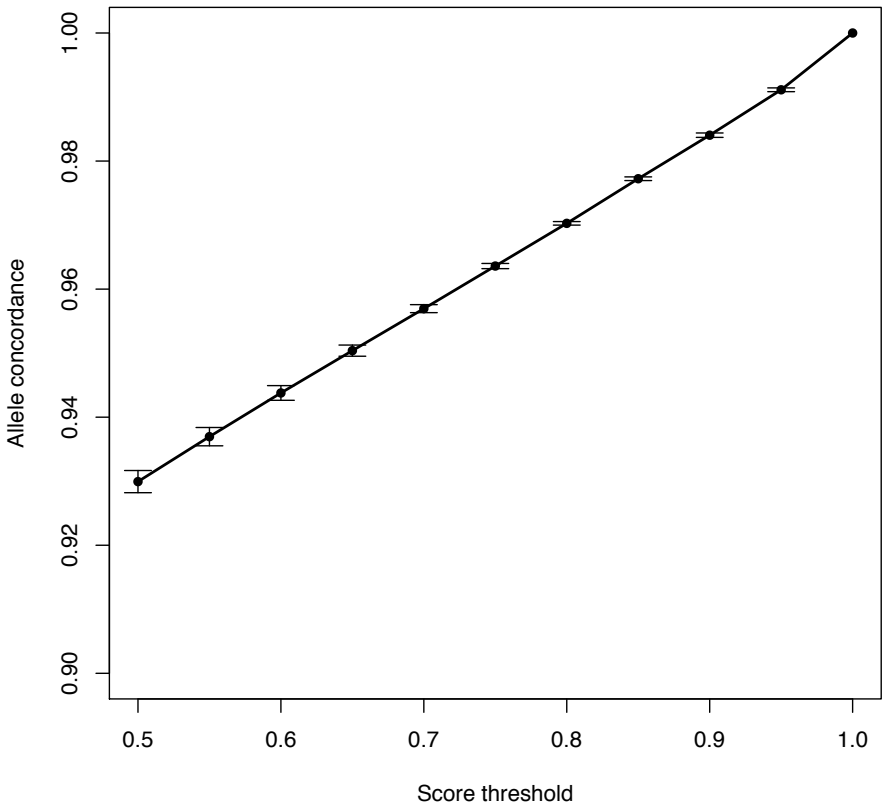
- Incorporate read alignment quality scores
- Model the bias toward covering shorter alleles:

$$P(r|\langle A, B\rangle) = fP(r|A) + (1-f)P(r|B) \tag{10}$$

Currently f is set to 0.5.

Confidence score indicates call quality

Call concordance in 26 monozygotic twin pairs:



Challenges in representing lobSTR output

- How to annotate loci with many many alleles?
- Which alternate alleles to list?
- Which scores to return?

lobSTR VCF output format

Steps to create VCF output:

- Initial round of lobSTR allelotyping to generate prior allele frequencies and to determine which alternate alleles to include
- Second round of lobSTR allelotyping to generate genotype likelihoods and posteriors for all possible allelotypes
- Merge VCFs from each sample with GATK
- Validate VCF with GATK

ALT field

##ALT=<ID=STRVAR,Description="Short tandem variation">

Alternate alleles are given as <STRVAR:$ALLELE>, where $ALLELE is the number of base pairs length difference of the alternate allele from the reference sequence

e.g.

ALT

<STRVAR:-4>, <STRVAR:-2>, <STRVAR:2>

GL field

##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype likelihoods (log10 scaled)">

- ALT: <STRVAR:-16>,<STRVAR:-12>,<STRVAR:-8>,<STRVAR:-4>
- Call: -8,-8
- GL: -7.27172,-7.5391,-7.92738,-6.65142,-6.77211,-6.29324,-0.610734,-0.610842,-0.610311,-0.00887698,-6.11645,-6.17955,-5.90496,-0.609606,-5.63758

Example

```
##fileformat=VCFv4.1
##fileDate=20130206
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of variant">
##INFO=<ID=MOTIF,Number=1,Type=String,Description="Repeat motif">
##INFO=<ID=REF,Number=1,Type=Float,Description="Reference copy number">
##ALT=<ID=STRVAR,Description="Short tandem variation">
##FORMAT=<ID=GB,Number=1,Type=String,Description="Genotype given in bp difference from reference">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype likelihoods (log10 scaled)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods \
  for genotypes as defined in the VCF specification">
##FORMAT=<ID=GPP,Number=G,Type=Float,Description="Genotype Posterior probabilities \
  (phred scaled, -10log10)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PP,Number=1,Type=Float,Description="Posterior probability of call">
##FORMAT=<ID=S1,Number=1,Type=Float,Description="Allele 1 marginal posterior">
##FORMAT=<ID=S2,Number=1,Type=Float,Description="Allele 2 marginal posterior">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA07037
20      241827  .       T       <STRVAR:-12>,<STRVAR:-8>,<STRVAR:-4>,<STRVAR:4> 4.67365 .    \
AC=0,0,2,0;AN=2;END=241870;MOTIF=AAAC;REF=11;VT=STR     \
GT:ALLREADS:ALLPARTIALREADS:CONFLICT:DP:GB:GL:PL:GPP:MP:PC:PP:S1:S2:STITCH:SUPP  \
3/3:-4|1:NA:0:1:-4/-4:-2.85876,-3.12974,-4.00366,-2.99245,-3.42602,-3.18659,\
-0.30447,-0.305033,-0.304791,-0.00404644,-3.09819,-3.80952,-3.36568,-0.304984,\
-3.67583:28,31,39,29,34,31,3,3,3,0,30,38,33,3,36:32.0814,47.9087,69.7654,44.7749,\
62.2282,58.0729,5.67662,18.7998,17.0365,1.81053,45.8323,66.0631,59.8638,17.0384,62.9653:\
.:0:0.659094:0.999243:0.999243:0:1
```

Acknowledgements

- **Yaniv Erlich**
- David Golan
- Saharon Rosset
- 1000 Genomes analysis group