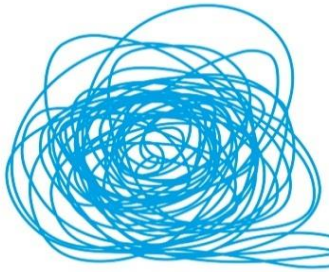
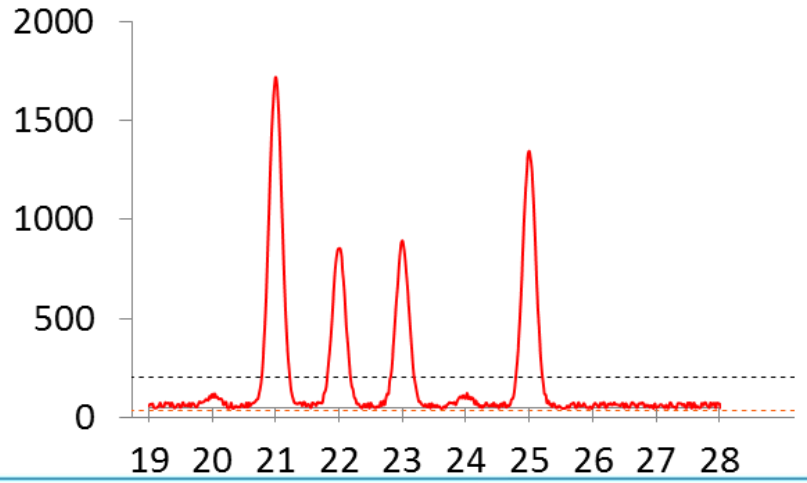
STRmix™

Dr John Buckleton



Profile interpretation

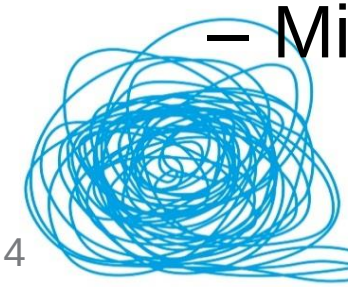
- Consider the locus below. What are the possible genotypes of the contributors?
- Information such as the apparent number of contributors and the approximate mixture proportions would help…



Profile interpretation

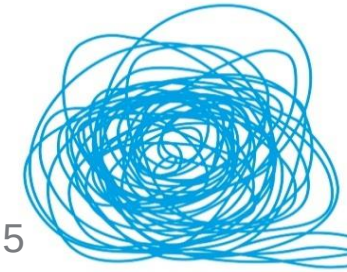
- We are considering p($O|S_j$)…

  …what is the probability of the observed profile given a proposed genotype combination, S_j?
- We might use some heuristics to limit the list of possible combinations
  - Peak height ratio (aka heterozygote balance)
  - Stutter ratios
  - Mixture proportion/ratio



Probability and likelihood

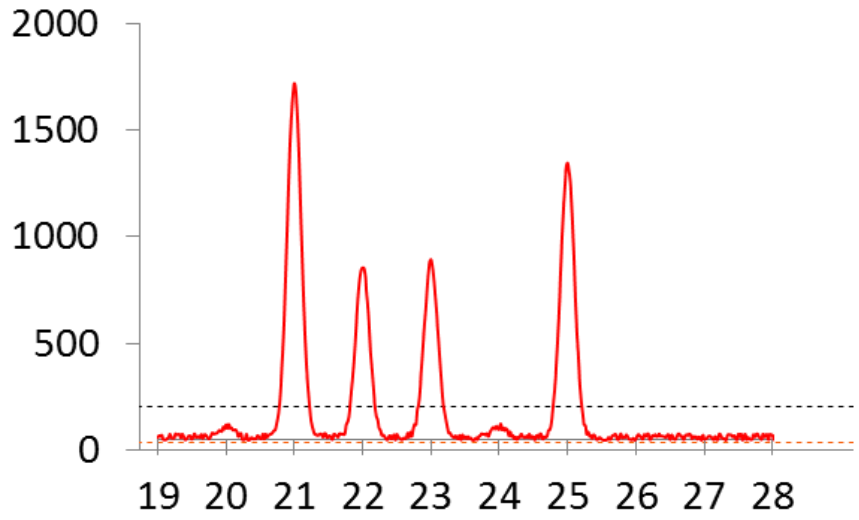
- The probability of the observed profile given a proposed genotype set: $p(O|S_j)$
- This is equivalent to the *likelihood* of the proposed genotype set given the observed profile: $L(S_j|O)$



Profile interpretation

- Assuming two contributors, there are six possible genotype combinations:

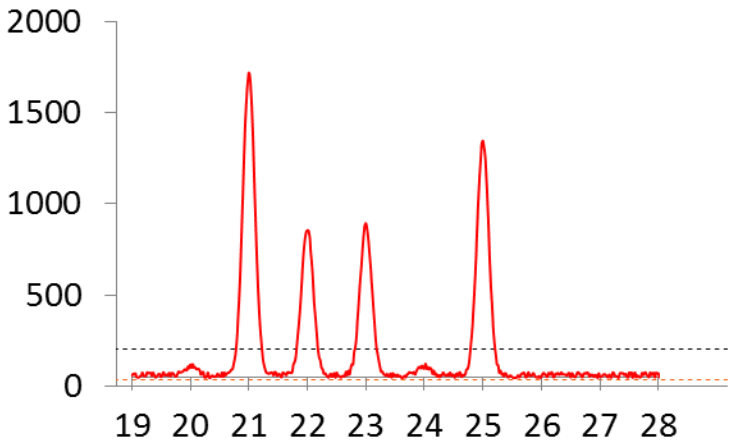
Major	Minor
21,22	23,25
21,23	22,25
21,25	22,23
22,23	21,25
22,25	21,23
23,25	21,22



Binary methods

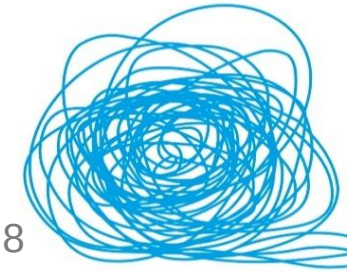
- Using a binary model we assign a weight of 1 when a genotype set explains the profile well
- And we assign a weight of 0 when a genotype set provides a poor explanation of the profile

Major	Minor	p($O \mid S_j$)
21,25	22,23	1
21,23	22,25	0
21,22	23,25	0
23,25	21,22	0
22,25	21,23	0
22,23	21,25	0



‘Binary’ methods

“The binary model assigns the values zero and one to the unknown probabilities, Pr($O|S_j$) … In essence Pr($O|S_j$) is assigned a value of zero if it is thought that this probability is very small relative to the other probabilities. Pr($O|S_j$) is assigned a value of one if it is thought that this value is relatively large.”



A comparison of statistical models for the analysis of complex forensic DNA profiles

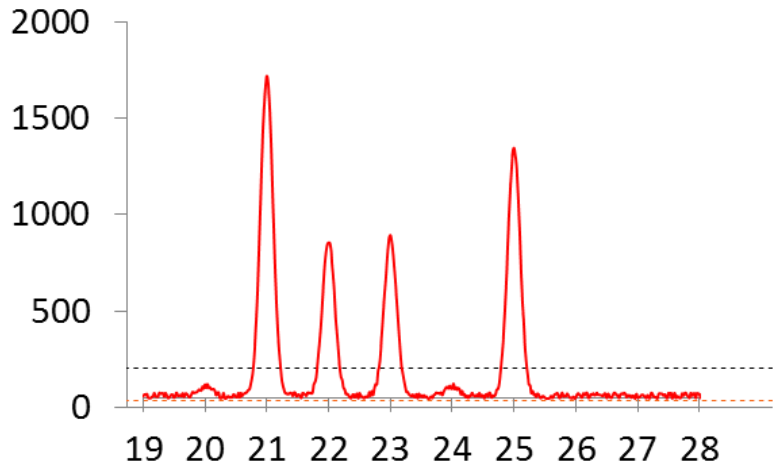
Hannah Kelly [a,b,*], Jo-Anne Bright [a], John S. Buckleton [a], James M. Curran [b]

Science and Justice 54 (2014) 66–70

Continuous methods

- In reality these probabilities are *continuous*
- The probabilities range between 0 and 1 inclusive

Major	Minor	$p(O \mid S_j)$
21,25	22,23	0.79
21,23	22,25	0.05
21,22	23,25	0.05
23,25	21,22	0.05
22,25	21,23	0.05
22,23	21,25	0.01



STRmix™

- Calculates p($O|S_j$)
  - Takes into account number of contributors, degradation, template, stutter, replicates…
- Called *weights*
- The weights are the primary output of STRmix™
- The weights are used when calculating a likelihood ratio

Calculation of the weights

- Determined using Markov chain Monte Carlo (MCMC)
- Likened to a game of hot and cold

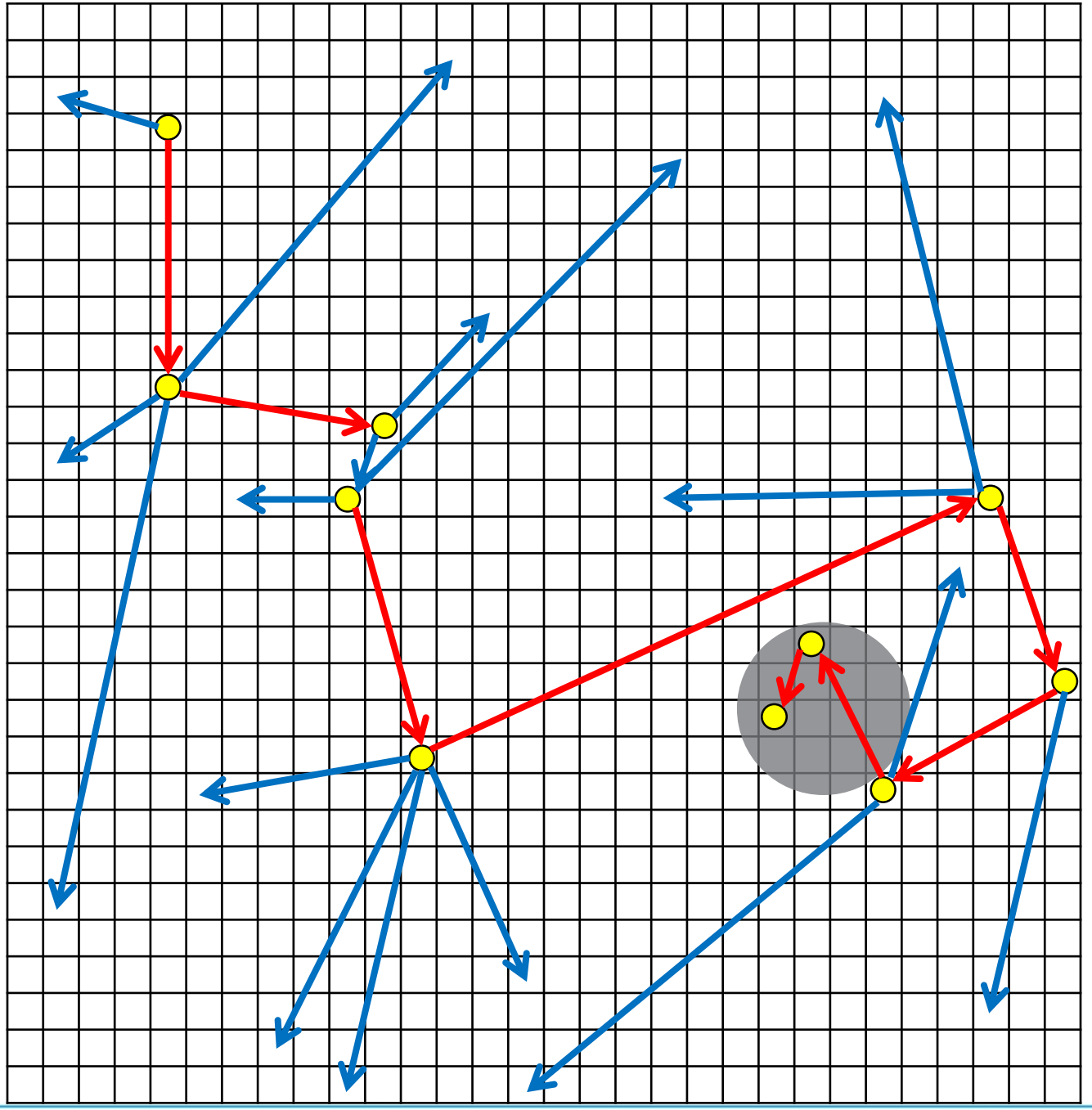
MCMC

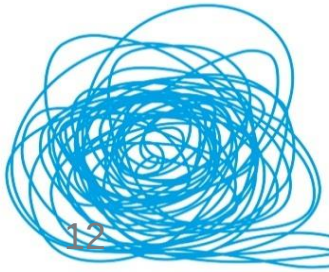
A possible answer

MCMC ‘chain’

Hot guess

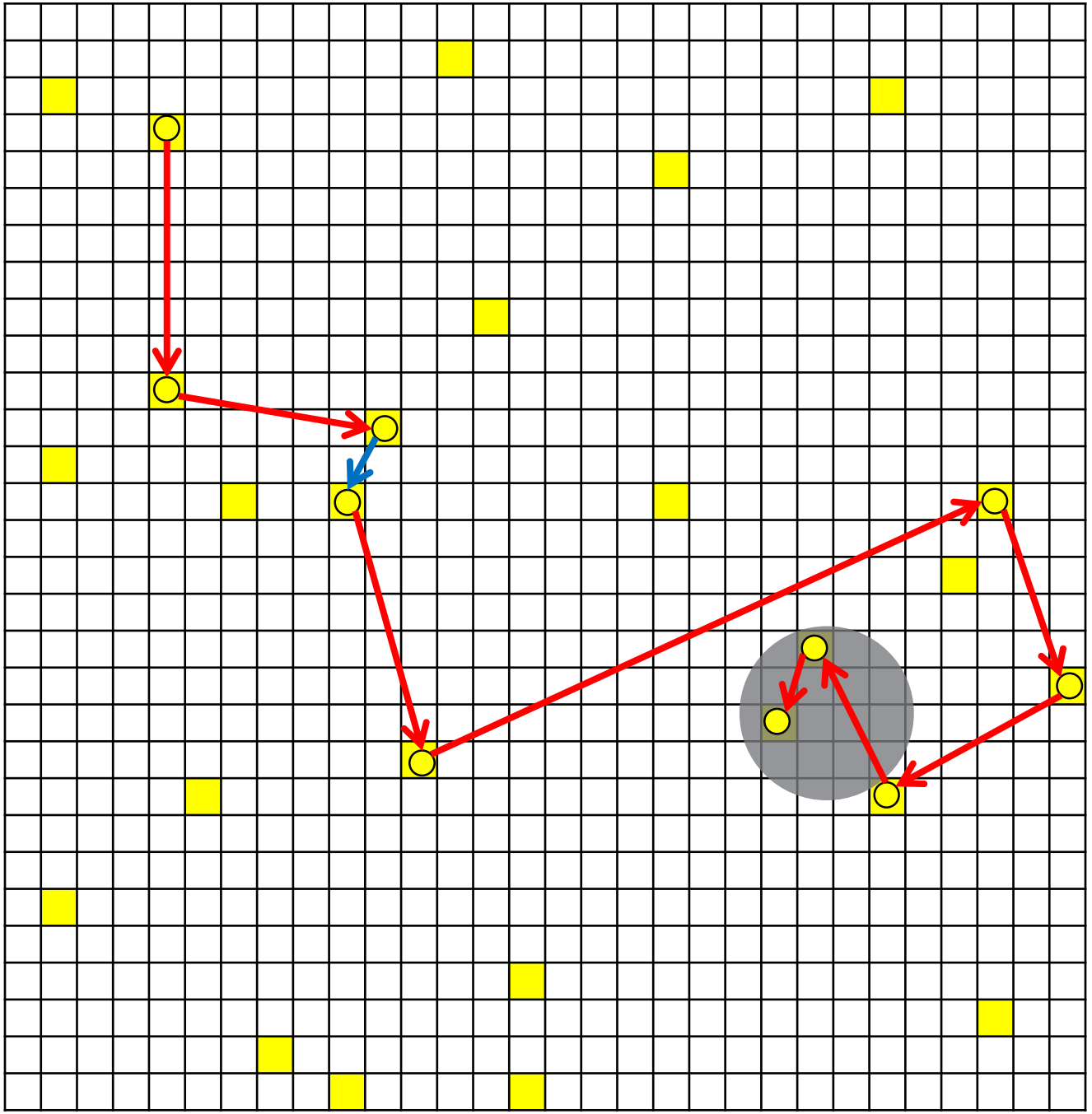
Cold guess





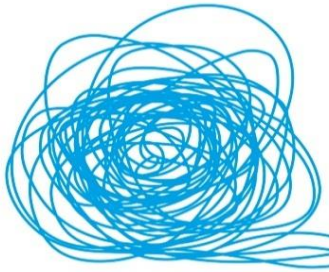
MCMC – maths

- We ran one chain for 10 moves
- Notice that we only had to test a small fraction of all the possible answers to reach the good answers
- Imagine if we ran the MCMC again for 10 moves: it would likely reach the good space, but via a different route (“random walk”)



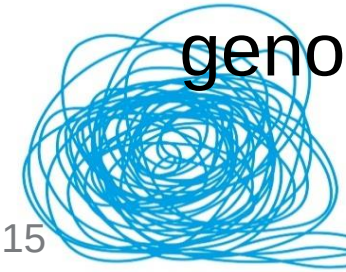
Iterations and accepts/moves

- *Iteration*: A proposed explanation for the recovered profile
  - The number of iterations can be in the billions depending in the complexity of the problem
- *Accepts* or *Moves:* A proposal (iteration) that is accepted (either hot or cold). By default, we run each chain for 50,000 accepts. *Accepts* are a subset of *iterations*



Markov chain Monte Carlo

- Formally: “Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution.”
- Less formally: An algorithm based on standard mathematical principles that assigns a likelihood for each genotype combination
- In STRmix™ the MCMC is ‘solving’ the equation for genotype weights

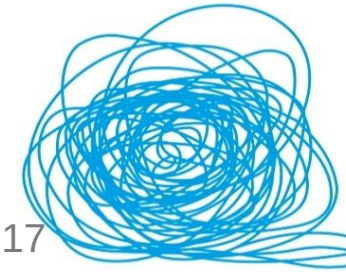


MCMC

- Monte Carlo methods rely on random sampling
- Named after Monte Carlo casino

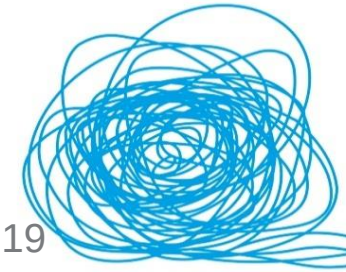
Examples of MCMC in common use

- Physics
  - Predicting behaviour of radiation particles
- Genetics
  - Creating phylogenetic trees
- Engineering
  - Predicting behaviour of buildings in earthquakes
- Aeronautics, social science, computational linguistics, weather predicting, stock market, betting



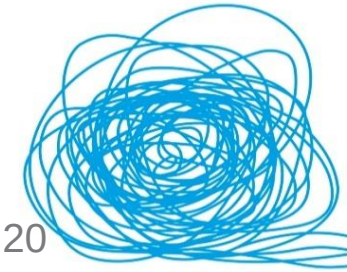
MCMC variability

- Each time we run the same problem, STRmix™ gives a different answer
- Importantly, these answers are all **clustered** around each other and the amount that they would vary is small in relation to the size of the number itself



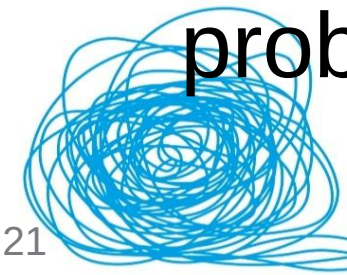
What is STRmix™?

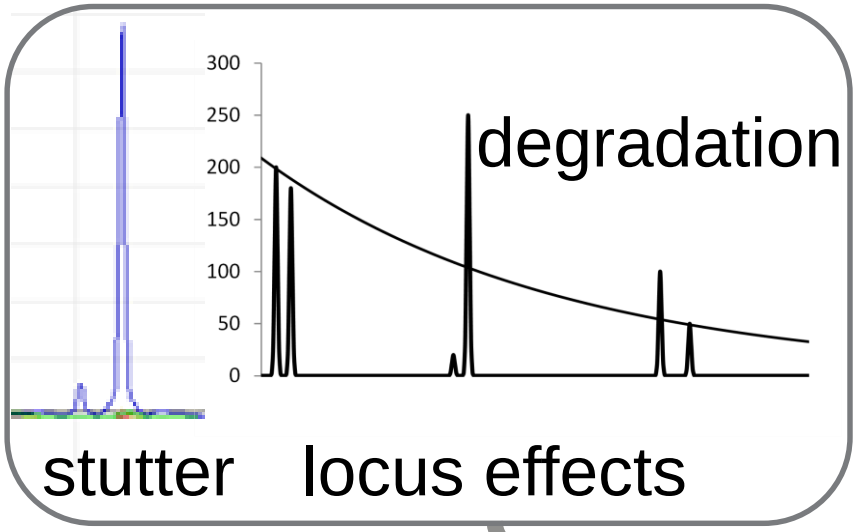
- Probabilistic software for the resolution of forensic DNA profiles
- A highly validated software that implements the most modern DNA interpretation methods
- Fully published and supported expert software



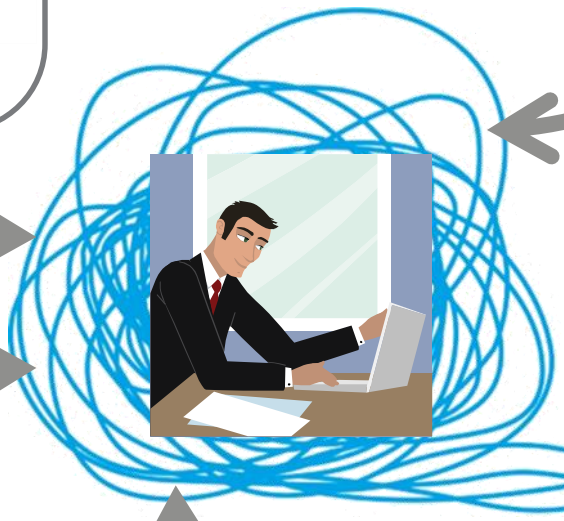
Input file

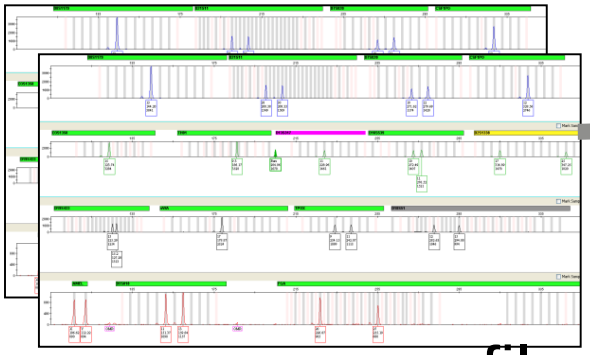
- Analysed text file into STRmix
  - Containing locus names, peaks, heights and their size
- Remove artifacts at analysis
  - Pull up and dye blobs
  - User laboratory determined AT
- STRmix models any type of stutter probabilistically





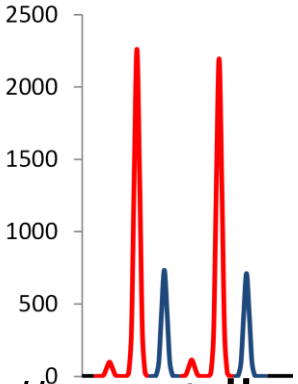
laboratory specific peak height variance





profiles

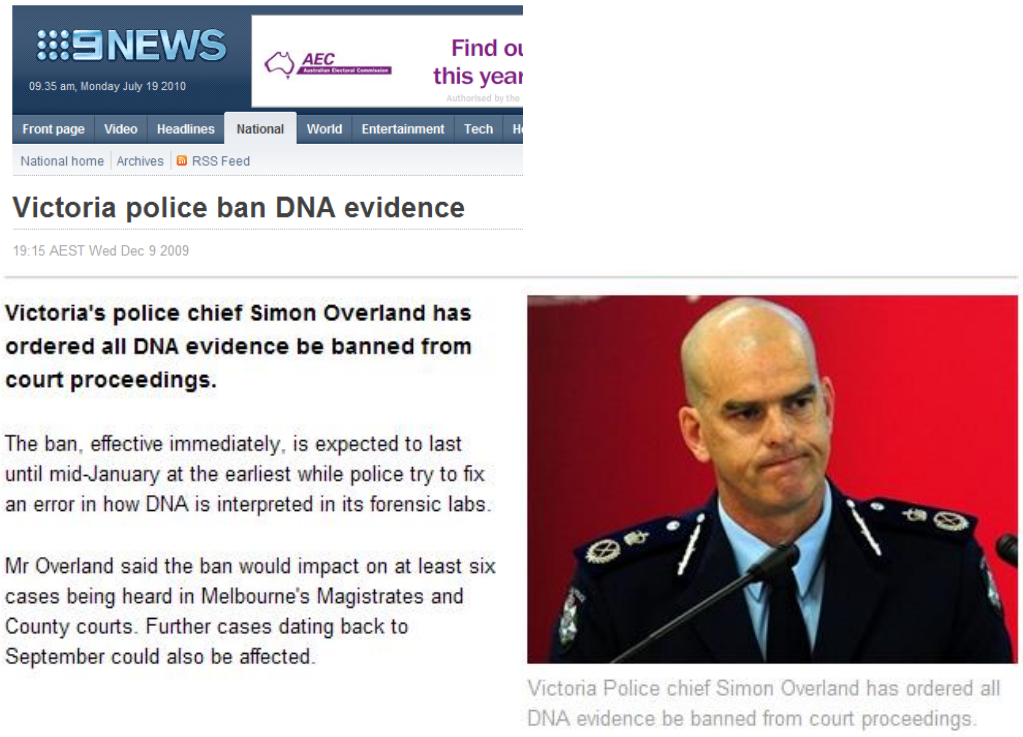
weights → *LR*



contributors

Victoria police ban DNA evidence

2009



9NEWS

09.35 am, Monday July 19 2010

AEC Australian Electoral Commission

Find ou this year

Authorised by the

Front page | Video | Headlines | National | World | Entertainment | Tech | H

National home | Archives | RSS Feed

Victoria police ban DNA evidence

19:15 AEST Wed Dec 9 2009

Victoria's police chief Simon Overland has ordered all DNA evidence be banned from court proceedings.

The ban, effective immediately, is expected to last until mid-January at the earliest while police try to fix an error in how DNA is interpreted in its forensic labs.

Mr Overland said the ban would impact on at least six cases being heard in Melbourne's Magistrates and County courts. Further cases dating back to September could also be affected.

Victoria Police chief Simon Overland has ordered all DNA evidence be banned from court proceedings.

Development starts

2010

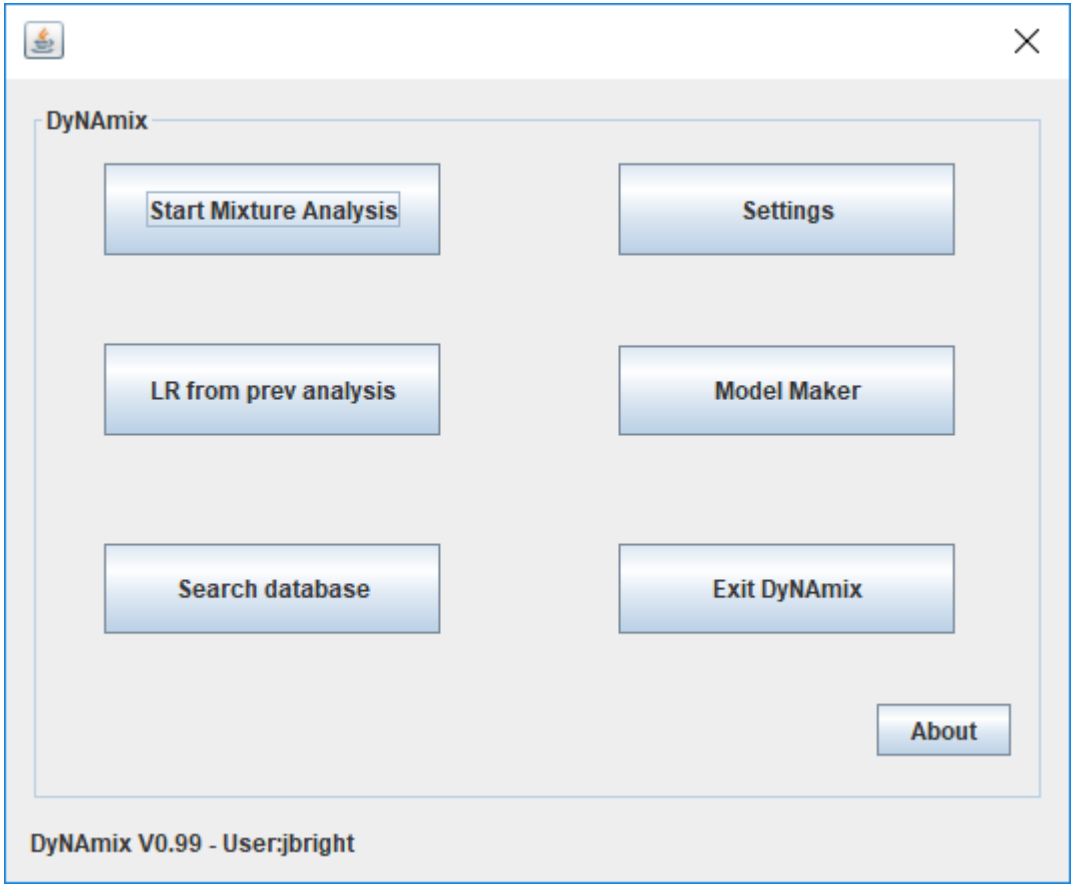
Government of South Australia

Science for Communities

PRIME MINISTER'S SCIENCE PRIZE WINNER 2018

2011

Prototype developed



August 2012

2012

ESR and FSSA go live casework

STRmix V1.08 released in Australia

2013

2014

STRmix V2.0 released internationally

2015
2016
2017

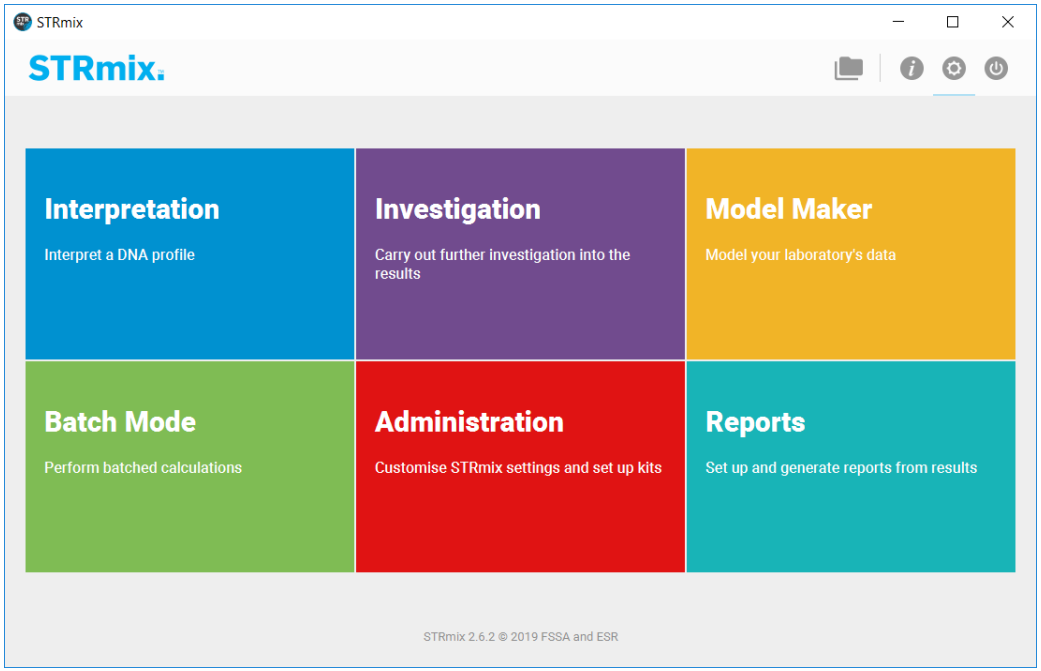
STRmix V2.3
STRmix V2.4
STRmix V2.5

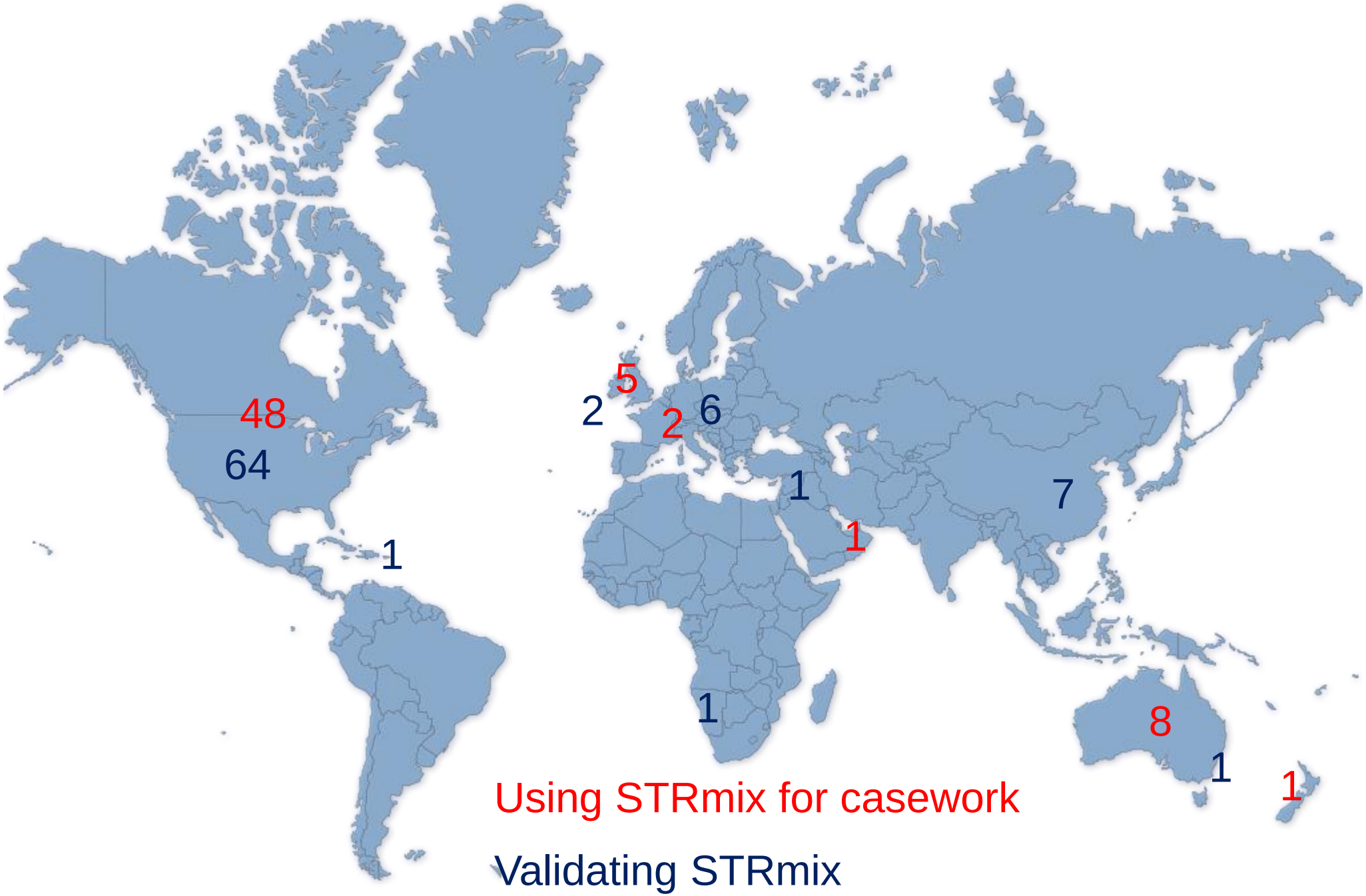
Improved modeling
Improved workflows
Run time improvements
Memory improvements

STRmix V2.6

2018

Current version



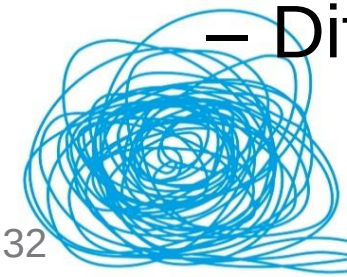


Using STRmix for casework

Validating STRmix

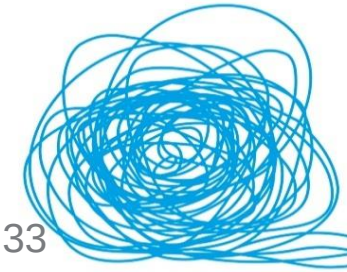
STRmix features

- Fully configurable
  - Common kit and allele frequency file exemplars
  - Configurable defaults/settings
  - Configurable reports
- Deconvolution of profiles
  - 1 through 5 contributors
  - Replicate amplifications
  - Different kits



STRmix features

- Variable number of contributors
- Hd true tester
- LR from previous deconvolution
- Batch mode (setup and queue processes)
- Model maker
- CODIS report

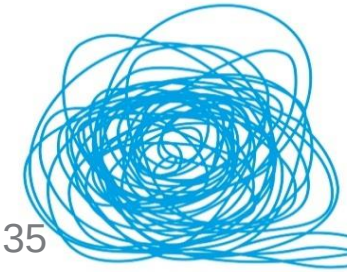


Likelihood ratios

- By default considering unknown individual's, common relatives
- (optional) Unified – unrelated and related under Hd
- (optional) Stratified across sub populations
- Lower bound taking into account allele frequency, MCMC, theta variability
- Database search
  - Standard and familial propositions

Run settings

- STRmix runs on a standard desktop or laptop
- No server required
- Standalone and server licenses available



Hardware requirements

Low spec[1]	Medium spec[2]	High spec[3]
Intel core 2 (Quad core) processor	Intel i5, i7 2.6 GHz	Intel 2.4 GHz Xeon CPU E5[4]
4 GB RAM	16 GB RAM	128 GB RAM
Minimum of 2 cores	Minimum of 4 cores	Minimum of 8 cores
300MB free HDD space[5]	300MB free HDD space[5]	300MB free HDD space[5]
Windows 7 or above	Windows 7 Professional 64 bit or above	Windows 7 Professional 64 bit or above

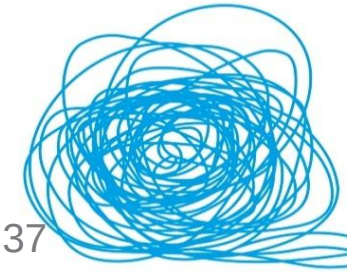
1-3p mixtures
Some 4p

Most 1-4p mixtures

Most 1-5p mixtures

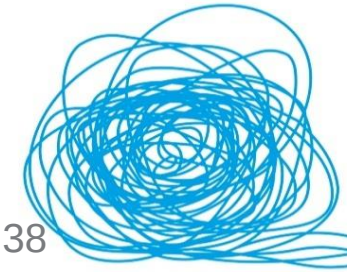
Approximate run times

- (Depends on hardware settings)
- Single source < 1 minute
- Two person mixture < 5 minutes
- Three person mixtures < 15 minutes
- Four person mixtures < 1 hour
- Five person mixture < 12 hours



Outputs

- Fully configurable report
- Contains settings, inputs, genotype combinations, weights
- Diagnostics for run performance



Science for Communities

PRIME MINISTER'S SCIENCE PRIZE WINNER 2018

STRmix. RESOLVE MORE DNA MIXTURES.

Interpretation Report

DETAILS

STRMIX VERSION:	STRmix 2.6.1
USER:	jbright
RUN DATE:	24 Feb 2019 21:48:34
TOTAL RUN TIME:	12 minutes, 0 seconds
REPORT VERSION:	STRmix 2.6.2
REPORT RUN:	01 May 2019 20:40:53

RUN PARAMETERS

CONTRIBUTORS:	3
PROFILING KIT:	Investigator_24plex_3500
SAMPLE FILE:	CMIX_2-2_D06_3500 Instrument_Investigator 24_plex.hid_EV.csv

CASE NUMBER:	QIAGEN
SAMPLE NAME:	CMIX_2-2_D06
COMMENTS:	
SEED:	401301

SUMMARY OF CONTRIBUTORS

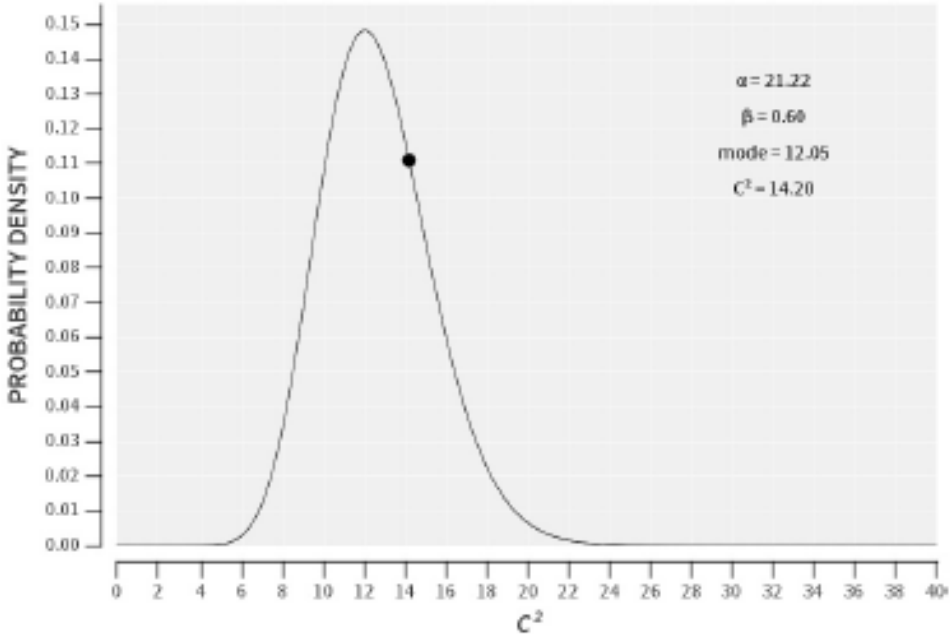
CONTRIBUTORS	1	2	3
Template (rfu)	3461	2059	784
Mixture Proportion	55%	33%	12%
Degradation starts at 88bp			
Degradation linear approximation (rfu/bp)	5.944	6.502	1.406
Degradation exponential curve	2.0964E-3	4.9769E-3	2.2186E-3

POST BURN-IN SUMMARY

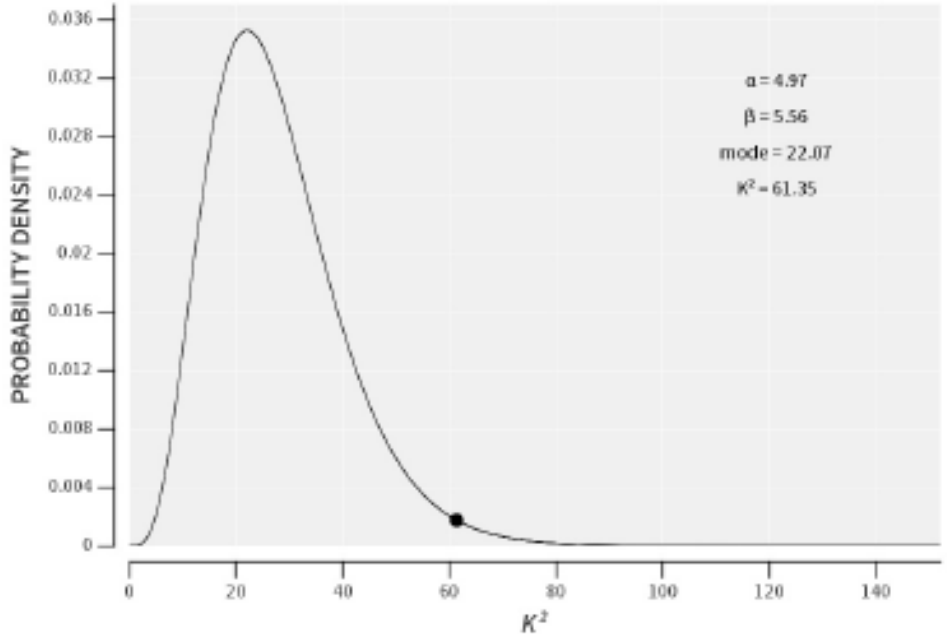
Total iterations	6,643,577	Acceptance rate	1 in 16.61
Effective sample size	20,052.49	log(likelihood)	38.58
Gelman-Rubin convergence diagnostic	1.16		
Allele variance (mode = 12.046)	14.204	Back Stutter variance (mode = 22.073)	61.352
Forward Stutter variance (mode = 12.046)	12.735		

VARIANCE CHARTS

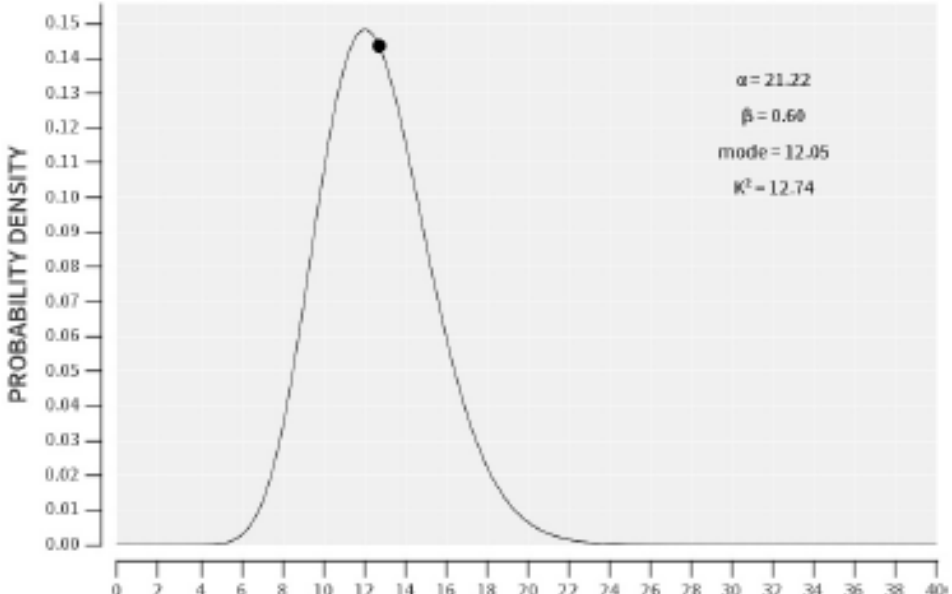
ALLELE VARIANCE



BACK STUTTER VARIANCE



FORWARD STUTTER VARIANCE



WEIGHTS

LOCUS	CONTRIBUTORS			WEIGHT
	1 (55%)	2 (33%)	3 (12%)	(HIGHLIGHT ≥ 0.99)
TH01	8, 9	8, 9.3	8, 9.3	2.8179E-1
	8, 8	9, 9.3	9, 9.3	1.0048E-1
	8, 9	8, 9.3	9.3, 9.3	6.6817E-2
	8, 9.3	8, 9	8, 9	6.5073E-2
	8, 9	8, 9.3	8, 9	5.9530E-2
	8, 9	8, 9.3	9, 9.3	5.6507E-2
	8, 8	9, 9.3	9, 9	5.2356E-2
	9, 9.3	8, 8	8, 8	5.1979E-2
	8, 9	8, 9.3	8, 8	4.8350E-2
	8, 9	9.3, 9.3	8, 8	4.1835E-2
	8, 8	9, 9.3	8, 9	3.7797E-2
	9, 9.3	8, 8	8, 9	2.5354E-2
	8, 9.3	8, 9	9, 9	2.5102E-2
	8, 9	8, 8	9.3, 9.3	1.7948E-2
	8, 8	9, 9	9.3, 9.3	1.2717E-2
	8, 9.3	9, 9	8, 8	9.8235E-3
	8, 9.3	8, 9	8, 8	9.3541E-3
	8, 8	9, 9.3	8, 9.3	9.2422E-3
	8, 9	8, 9.3	9, 9	5.2756E-3
	8, 8	9, 9.3	9.3, 9.3	3.8238E-3
	8, 8	9, 9.3	8, 8	3.4232E-3
	8, 9.3	8, 8	9, 9	2.8474E-3
	8, 8	9.3, 9.3	9, 9	2.3043E-3
	8, 9	9.3, 9.3	8, 9	1.8051E-3
	9, 9.3	8, 8	8, 9.3	1.3789E-3
	8, 9.3	8, 9	9, 9.3	1.1608E-3
	8, 9	9.3, 9.3	8, 9.3	9.9449E-4
	9, 9.3	8, 8	9, 9	8.1853E-4

COMPONENT INTERPRETATION

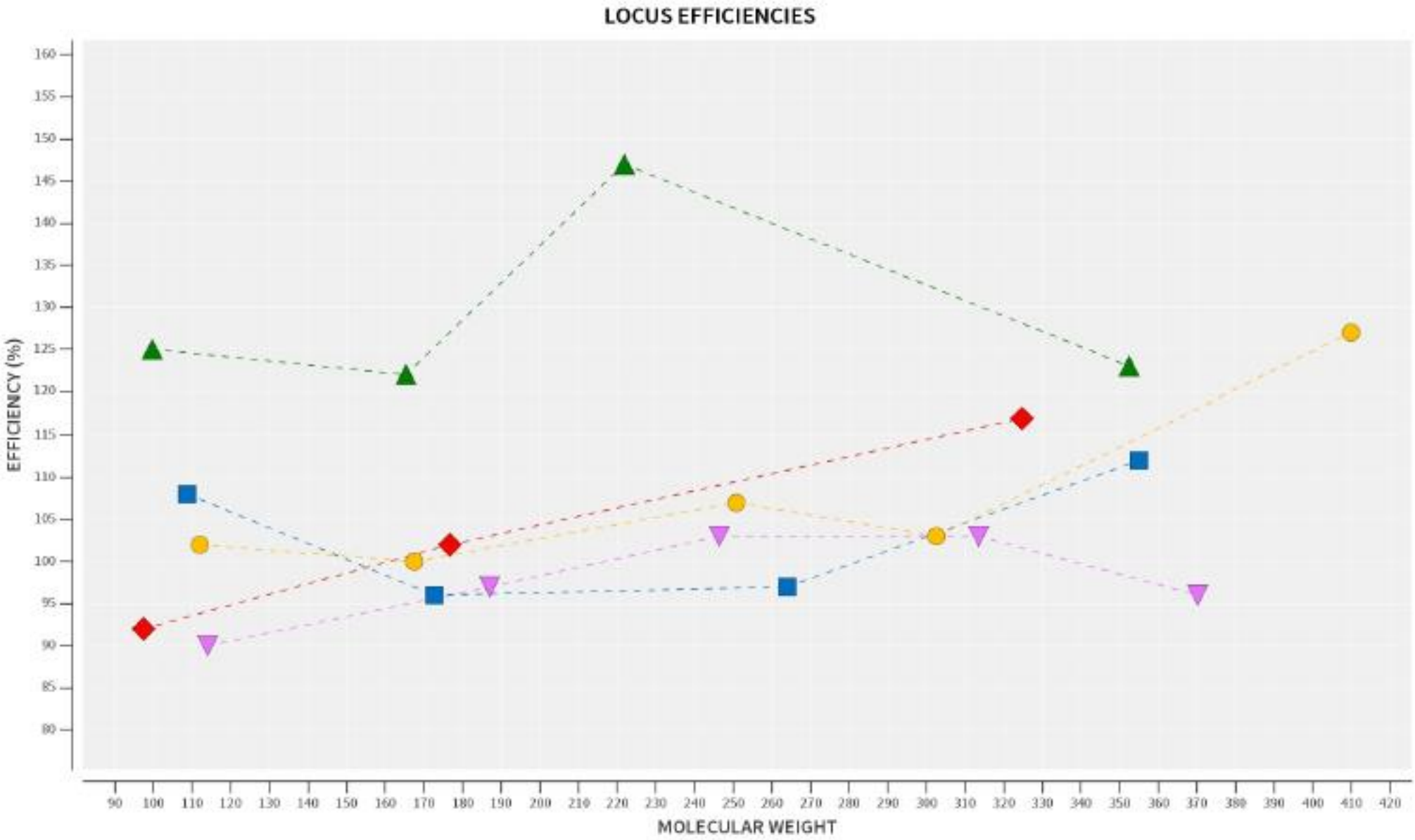
CONTRIBUTOR 1 (55%)

Questioned contributor

LOCUS	GENOTYPE	WEIGHT	COMPONENT ≥ 99%
TH01	8, 9	58.24%	F, F
	8, 8	22.29%	
	8, 9.3	11.48%	
	9, 9.3	7.99%	
D3S1358	14, 18	45.95%	F, F
	16, 18	31.30%	
	14, 16	22.68%	
	18, 18	0.07%	
	14, 14	0.00%	
vWA	16, 17	58.05%	F, F
	16, 18	27.84%	
	17, 18	13.72%	
	16, 16	0.37%	
	17, 17	0.01%	
	18, 18	0.00%	
D21S11	29, 32.2	98.59%	29, F
	29, 29	1.16%	
	28, 29	0.11%	
	28, 32.2	0.09%	
	29, 30	0.03%	
	30, 32.2	0.01%	
	32.2, 32.2	0.00%	
TPOX	8, 8	100.00%	8, 8
DYS391			

COMPONENT INTERPRETATION SUMMARY ≥ 99%

LOCUS	CONTRIBUTORS		
	1 (55%)	2 (33%)	3 (12%)
TH01	F, F	F, F	F, F
D3S1358	F, F	F, F	F, F
vWA	F, F	F, F	F, F
D21S11	29, F	F, F	F, F
TPOX	8, 8	8, F	11, F
DYS391			
D1S1656	F, F	F, F	15, F
D12S391	F, F	F, F	F, F
SE33	17, 19	F, F	F, F
D10S1248	F, F	F, F	F, F
D22S1045	15, 15	F, F	F, F
D19S433	15, F	F, F	F, F
D8S1179	11, F	F, F	F, F
D2S1338	19, 23	F, F	F, F
D2S441	F, F	F, F	F, F
D18S51	16, 17	F, F	F, F
FGA	20.2, 25	F, F	F, F
D16S539	F, F	F, F	F, F
CSF1PO	F, F	F, F	F, F
D13S317	F, F	F, F	F, F
D5S818	13, F	F, F	F, F
D7S820	F, F	F, F	F, F


LOCUS EFFICIENCIES
EFFICIENCY (%)
160
155
150
145
140
135
130
125
120
115
110
105
100
95
90
85
80
90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420
MOLECULAR WEIGHT

EVIDENCE INPUT FILES

CMIX_2-2_D06_3500 INSTRUMENT_INVESTIGATOR 24_PLEX.HID_EV.CSV

LOCUS	ALLELE	HEIGHT	SIZE
AM	X	6549	76
	Y	5618	79
TH01	7	216	103
	8	6756	107
	9	3559	111
	9.3	2735	114
D3S1358	13	234	162
	14	2630	166
	15	369	171
	16	2378	175
	17	974	179
	18	3048	183
vWA	15	232	258
	16	3071	262
	17	2384	266
	18	1797	270
D21S11	28	1200	346
	29	3270	350
	30	819	355
	31.2	178	360
	32.2	1861	364
TPOX	7	2026	93
	8	14040	97
	11	584	109

SETTINGS

CASE SETTINGS

Case number	QIAGEN
Sample ID	CMIX_2-2_D06
Comments	
Seed	401301
Extended output	N

MCMC SETTINGS

Number of contributors	3
Use Mx priors	N
Number of chains	8
Burn-in accepts per chain	100,000
Post burn-in accepts per chain	50,000
Random walk SD	0.005
Post burn-in shortlist	9.0
Auto-continue on Gelman-Rubin	N

KIT SETTINGS

Ignored loci	DYS391	
Detection thresholds	**TH01**	**80**
	D3S1358	**80**
	vWA	**80**

Thanks from team STRmix