Implementing probabilistic methods into casework and development of the LiRa software

Science
for a safer world

Initial internal decisions

- Assuming you do want to implement a probabilistic method into casework then you will be faced with a number of decisions/questions
  - Be self-sufficient or procure off the shelf?
  - What's out there?
  - What does it do?
  - Does it meet our needs?
  - Come to mention it, what are our needs!!!?
  - How much is this all going to cost us?
- The last question was provided courtesy of your lab management!

Learning points

- Firstly, the implementation of any probabilistic method will rely on the development of a bio-statistical model which in turn will require a considerable amount of effort to test, challenge and refine.
- Unless of course you get someone else to do that bit for you!

Learning points

- Secondly, the model to be implemented will depend on the experimental dataset on which it is based and consequently those data must accurately reflect the analytical process used to generate the data that is to be analysed.

Learning points

- Thirdly, the complexity of the model and the intensity of the subsequent computations will require the development of a mathematical algorithm to avoid computational errors.
- Inevitably this means the production of computer software.
- The corollary of this is that a project to develop and implement probabilistic statistical methods requires a multi-disciplinary, team-based approach , needing inputs from forensic biologists, forensic statisticians and computer programmers.

Learning points

- Fourthly, one also needs to consider the significant validation requirements and the fact that there is no accepted road map to validating and deploying expert software

Learning points

- Fifthly, one also needs to factor in practitioner training requirements
- and for those in an adversarial legal system, court support when the inevitable big challenge arises
- This will all cost time and lost output

Learning points

- Lastly, one should not ignore the cultural impact on practitioners themselves.
- Naturally, concerns have been raised by the practitioner community about the subjugation of their expertise in favour of the increasing reliance being placed on ever more complicated ‘expert’ software and the concomitant dangers of developing a ‘black box’ approach to mixture evaluation.
- The key to this is training, education and understanding which will make the black box transparent
- The software is a tool and not a replacement for human understanding and judgment – let the computer handle the math but that’s all.
- The GIGO principle still applies

Off–the –shelf solution? My advice...

- Form a project group with terms of reference
- Draw up your user specification
- Consider your technical specification
- Consider any jurisdictional requirements
- Look at delivery timelines?
- Look at the available software – obtain trial copies and evaluate it
- CapEx and OpEx?
- Validation?
- Training from provider?
- Casework rollout support from provider?
- Business continuity of provider?
- Court challenge?

Developing your own solution – My advice....

- Form project group to consider key decisions and appoint a project manager
- You will definitely need internal/external statistical expertise?
- Get a professional software developer/ programmer on-board
- Development timescales?
- Hardware?
- Developmental validation
- Casework validation
- Training plan
- Rollout plan
- CapEx?
- OpEx?

My user specification?

- Allowance for drop out
- Allowance for drop in
- Allowance for stutter (under and over? e.g D22)
- Allowance for uncertain alleles
- Multiple PCR replicates factored in
- Allele adjustment method (sampling correction)
- User specified θ correction
- Syntenic correction
- Nc= 1,2,3,4 and >4?
- User controlled proposition selection
- Allowance for known contributors (conditioning)
- Allowance for relatives

Technical specifications (hardware)

- Runs in a reasonable time frame on a PC
- no requirement to invest in servers and top of the range computers
- How realistic is this?
- Not very realistic
- You will encounter computing issues
- More on this later

Technical specification (software)

- Does not require purchase of an additional software package (e.g. R)
- Code is written in a professionally recognised language that can be debugged (e.g. C#)
- Does not require users to be conversant in that language
- Aesthetically pleasing and intuitively easy general user interface (GUI)

Technical specifications (others)

- Plasticity (futureproofing)
- Ability to add/change frequency databases and maintain legacy databases
- Ability to be configured to use different multiplex kits
- Ability to add new functionality
- Printable (and presentable) outputs for all loci

Which method to implement?
Discrete v Continuous models

- Discrete models use the presence/absence of peaks but do not take their heights into consideration (qualitative approach)
- Continuous models treat peak heights as continuous variables and take into account the amount of each allele (quantitative)
- Does it matter?
- Comparative performance data required?
- Have both available?

Discrete models

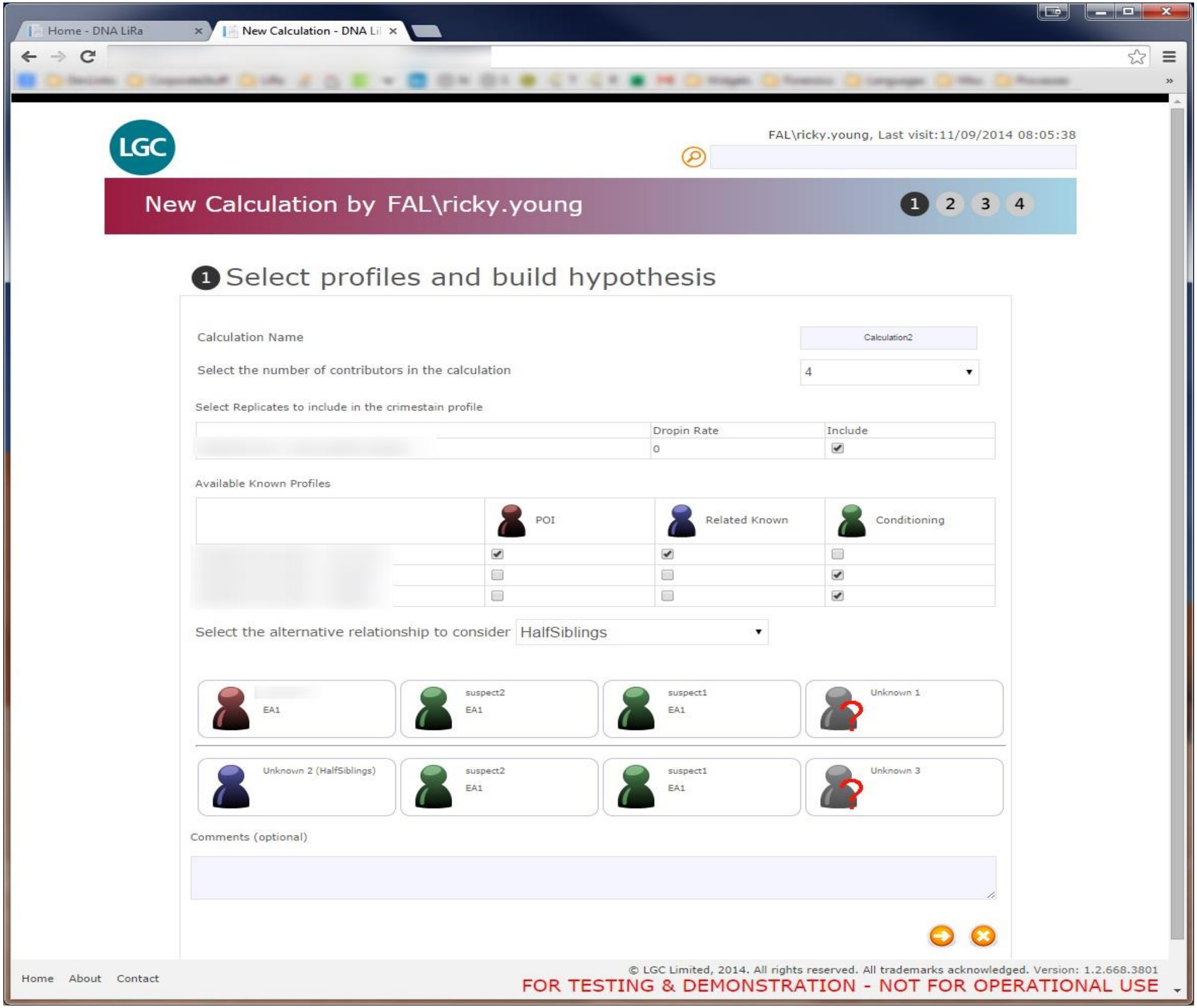
- Simpler
- Easier to implement?
- Less sensitive to variation in system?
- May perform better when peak heights are lower?
- Avoids extensive data collection and parameter estimation?

Continuous models

- Use all of the data including peak heights
- A more complete solution?
- More powerful?
- More sensitive to variation in system?
- Where MCMC simulation is used may not get the same numerical answer in two successive runs with the same data
- Will the CJS accept this?

LiRa

- Meets the user specification outlined above
- Has been professionally developed in C#
- User-friendly GUI
- Does not require R or any supporting software
- Will run on your PC
- LiRa suite has both a discrete and a continuous model implemented



FAL\ricky.young, Last visit:11/09/2014 08:05:38

New Calculation by FAL\ricky.young

1 Select profiles and build hypothesis

Calculation Name: Calculation2

Select the number of contributors in the calculation: 4

Select Replicates to include in the crimestain profile

	Dropin Rate	Include
[redacted]	0	☑

Available Known Profiles

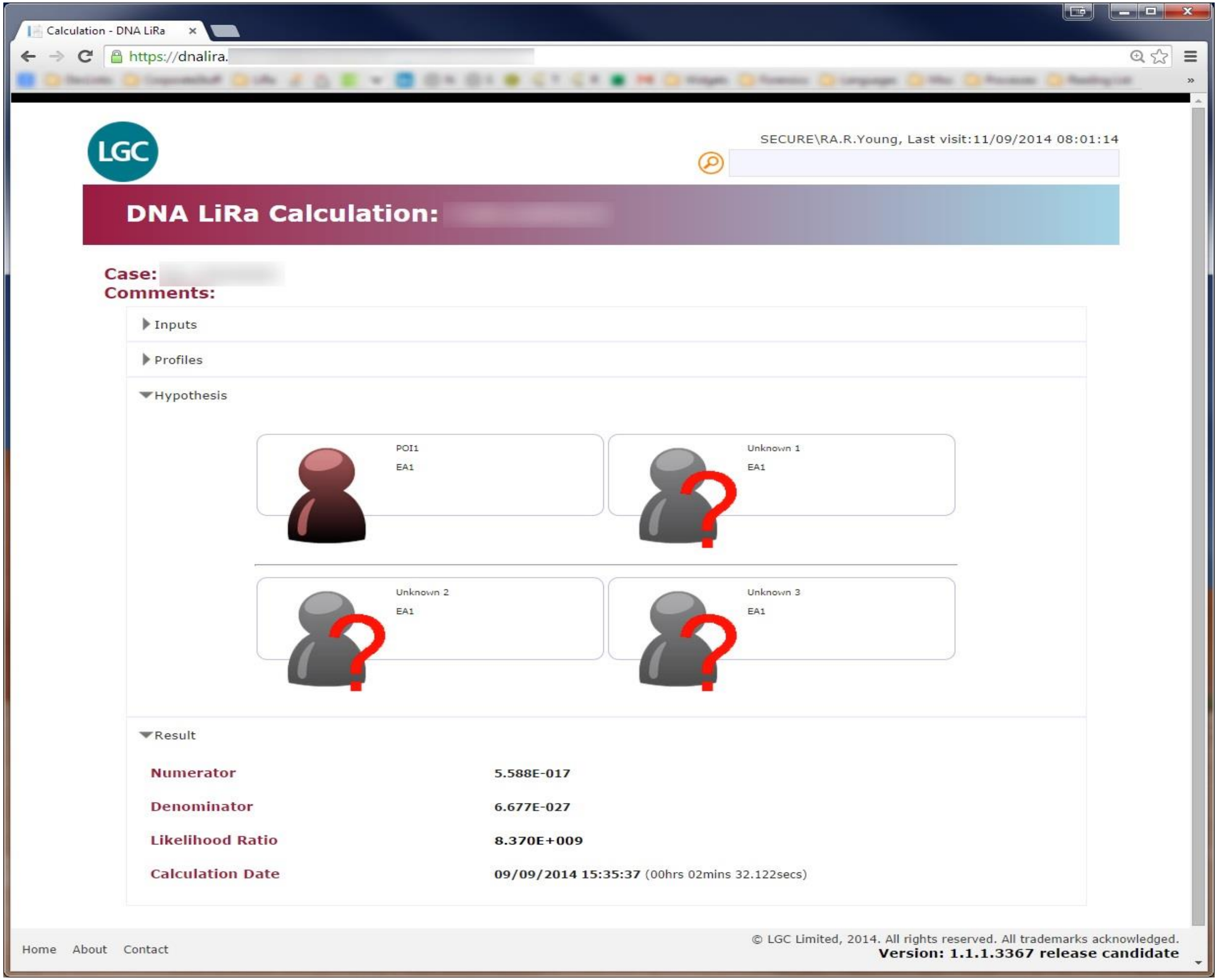
	POI	Related Known	Conditioning
[redacted]	☑	☑	□
[redacted]	□	□	☑
[redacted]	□	□	☑

Select the alternative relationship to consider: HalfSiblings

[redacted] EA1	suspect2 EA1	suspect1 EA1	Unknown 1
Unknown 2 (HalfSiblings)	suspect2 EA1	suspect1 EA1	Unknown 3

Comments (optional)

FOR TESTING & DEMONSTRATION - NOT FOR OPERATIONAL USE



SECURE\RA.R.Young, Last visit:11/09/2014 08:01:14

DNA LiRa Calculation: [redacted]

Case: [redacted]
Comments:

▸ Inputs

▸ Profiles

▾ Hypothesis

▾ Result

Numerator	5.588E-017
Denominator	6.677E-027
Likelihood Ratio	8.370E+009
Calculation Date	09/09/2014 15:35:37 (00hrs 02mins 32.122secs)

Version: 1.1.1.3367 release candidate

How do you validate probabilistic software?

- Generate a result from a test sample
- Import data into software, set the LR propositions and compute the LR
- Compare the calculated LR to true LR
- But here’s the conundrum.........
- If we knew the true LR then we would not need to create an algorithm to calculate it in the first place
- In other words there is **no ground truth** (i.e. no way of assessing whether your LR is ‘correct’)
- In fact the whole notion of there being a ‘correct’ answer is flawed – there is no fixed and definitive answer

Validating probabilistic software?

- There is no generally accepted method for validating probabilistic models
- However, in my view there are 3 separate components to consider:
  - The underlying statistical/mathematical model
  - The code (algorithm)
  - The forensic process (procedures, policies, inputs and outputs)
- All 3 require a different form of validation
- The last two are more straightforward to validate but the first is not
- How do you validate the bio-statistical model when there is no ground truth!
- Some suggestions...

1. Performance of the model (Behaviour testing)

- Model should behave in a predictable and logical way
- We can check that the model exhibits expected behaviours
- ADO should generally weaken the LR for a true contributor
- More complexity in the mixture should generally decrease LR for a true contributor
- More known genotypes in the mixture should constrain mixture and generally increase LR for a true contributor
- More known genotypes in the mixture should generally decrease the LR for a non-contributor
- Adding additional information in the form of more replicates should generally increase the LR
- The max LR should not be > 1/CMP (notwithstanding the effects of rounding and modelling approximations)

Use of BANS

- A vast array of possible LRs can be computed
- 10-e20 to 10+e20
- BAN is courtesy of war time code-breaker Alan Turing
- The power to which 10 would need to be raised to produce the observed LR
- E.g. LR of 10,000 is 1 x 10e4 which is BAN 4
- Concept of the BAN Quotient (BQ)
- BAN/BAN for inverse CMP
- If CMP is 1 x 10-e16 (based on full profile) then
- BQ = 4/16 = 0.25
- Use model to explore the effects of increasing uncertainty in the data (partiality and ambiguity)
- BQ should tend towards zero
- Or to increase information (e.g. through doing replicates, conditioning)
- BQ should tend towards one

Artificial test profiles

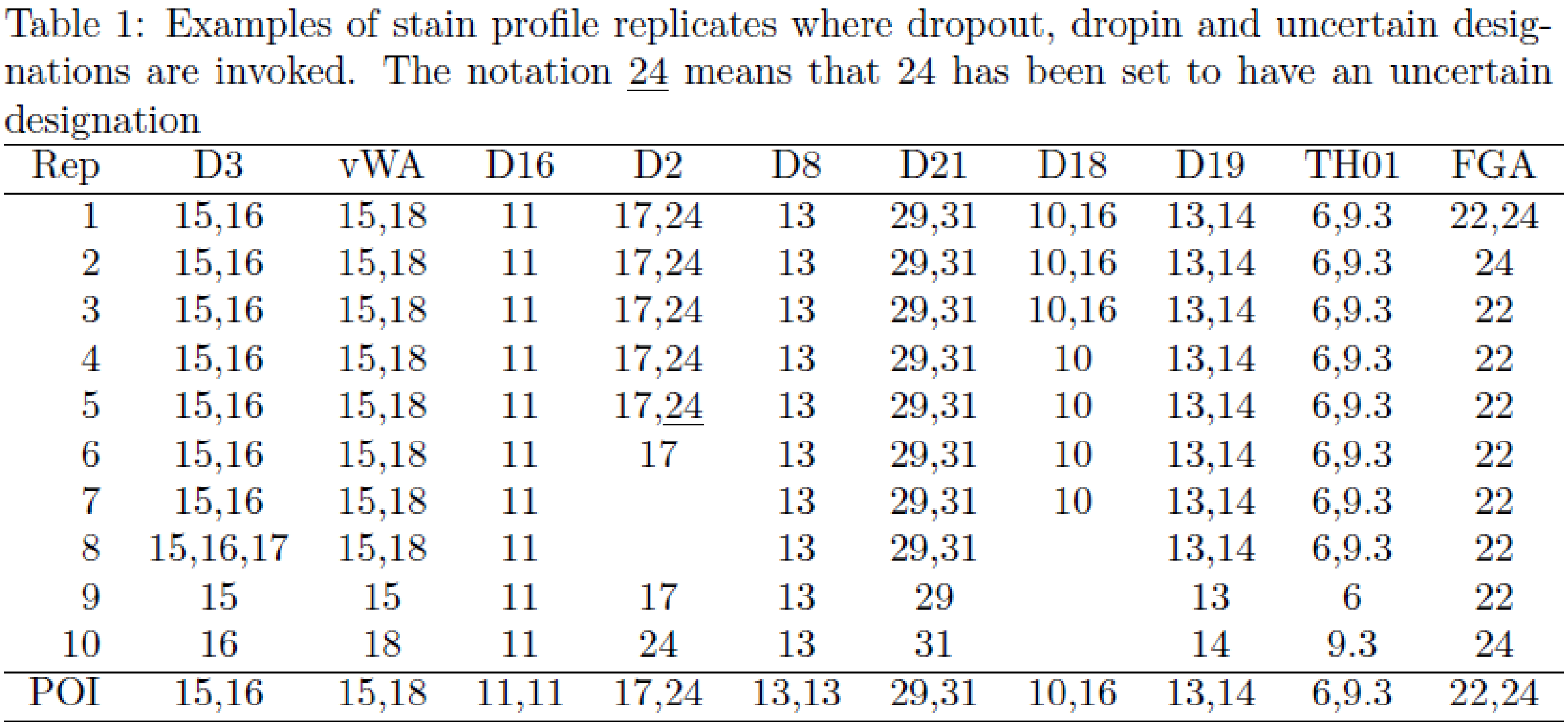


Table 1: Examples of stain profile replicates where dropout, dropin and uncertain designations are invoked. The notation 24 means that 24 has been set to have an uncertain designation

Rep	D3	vWA	D16	D2	D8	D21	D18	D19	TH01	FGA
1	15,16	15,18	11	17,24	13	29,31	10,16	13,14	6,9.3	22,24
2	15,16	15,18	11	17,24	13	29,31	10,16	13,14	6,9.3	24
3	15,16	15,18	11	17,24	13	29,31	10,16	13,14	6,9.3	22
4	15,16	15,18	11	17,24	13	29,31	10	13,14	6,9.3	22
5	15,16	15,18	11	17,24	13	29,31	10	13,14	6,9.3	22
6	15,16	15,18	11	17	13	29,31	10	13,14	6,9.3	22
7	15,16	15,18	11		13	29,31	10	13,14	6,9.3	22
8	15,16,17	15,18	11		13	29,31		13,14	6,9.3	22
9	15	15	11	17	13	29		13	6	22
10	16	18	11	24	13	31		14	9.3	24
POI	15,16	15,18	11,11	17,24	13,13	29,31	10,16	13,14	6,9.3	22,24

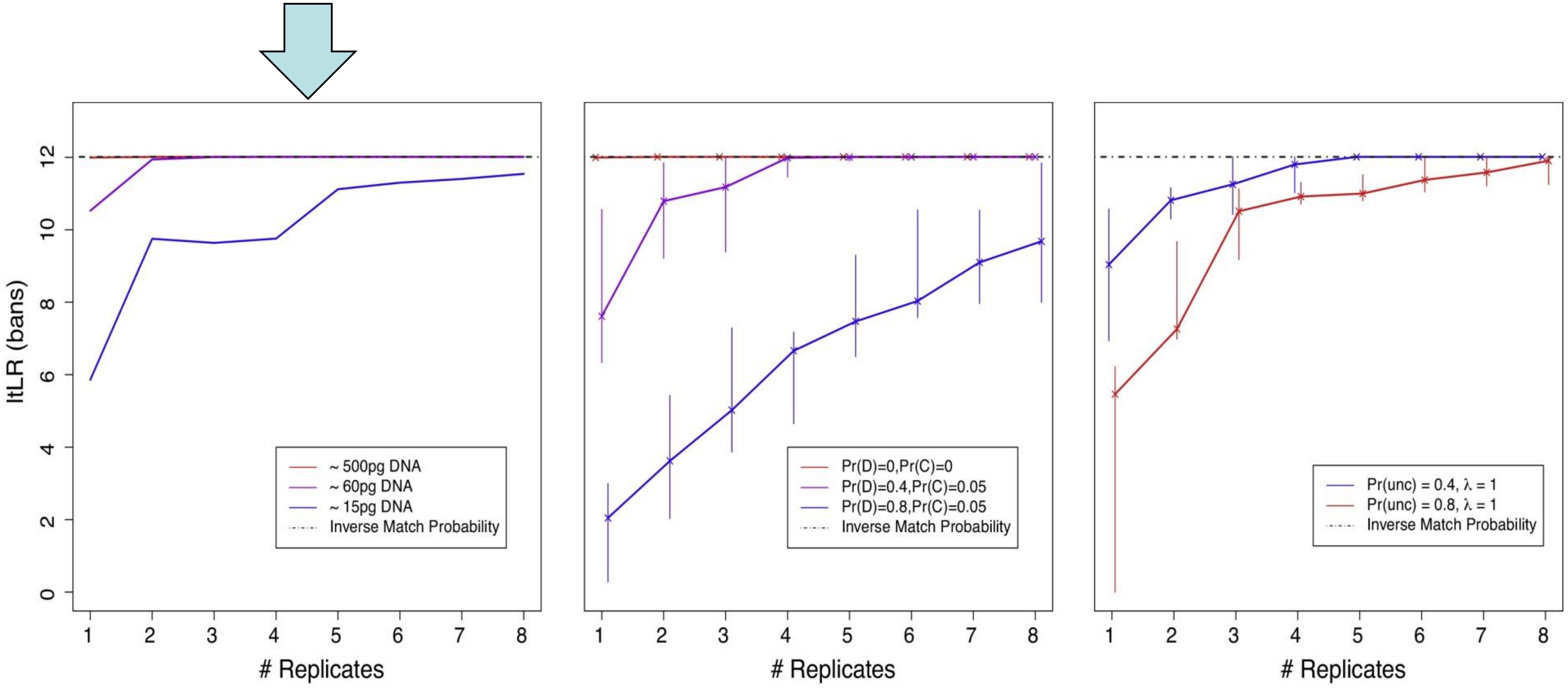
From : Puch-Solis, R and Clayton T (2014) FSI Genetics Vol 11 pg 220-228

From :
Puch-Solis, R and Clayton T (2014) FSI Genetics Vol 11 pg 220-228

No.	LiRa (*BQ*)
1	11.47 (1.00)
2	10.17 (0.89)
3	9.84 (0.86)
4	9.31 (0.81)
5	8.48 (0.74)
6	8.42 (0.73)
7	8.29 (0.72)
8	6.61 (0.58)
9	3.69 (0.32)
10	3.91 (0.34)

No.	Reps	LiRa (*BQ*)
1	1, 1	11.47 (1.00)
2	2, 3	11.45 (1.00)
3	7,9,10	10.62 (0.93)
4	9, 10	9.61 (0.84)
5	4, 7	8.88 (0.77)
6	5, 8	8.43 (0.73)
7	7, 8	8.22 (0.72)
8	5, 6	7.90 (0.69)

Or use ‘real’ replicates (diluted to show the required phenomena)



From :
Steele, C, Greenhalgh M and Balding D (2014) FSI Genetics Vol 13 pg 82-89

2. Performance of model (Comparative testing)

- Take another program with similar modelling (e.g. Compare a discrete model with another discrete model)
- Use same input data
- Set parameters to be same or as close as software will allow
- Compare outputs in BANS
- Similar BANS tends to suggest that, irrespective of modelling choices and the computer implementation, the weight of evidence is being estimated consistently
- Remember - there is no ground truth!

LiRa discrete -v- LikeLTD

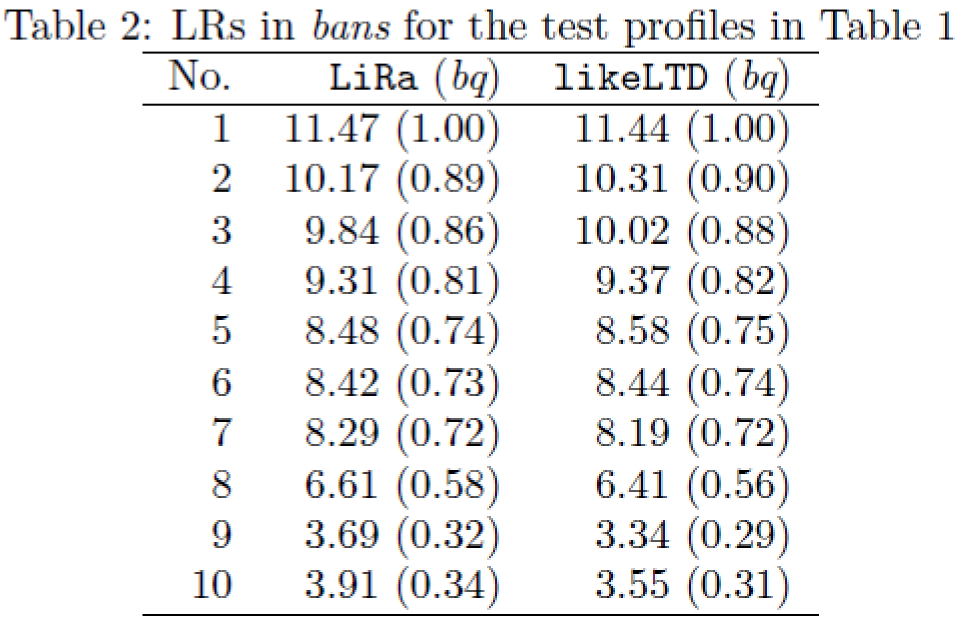


Table 2: LRs in *bans* for the test profiles in Table 1

No.	LiRa (bq)	likeLTD (bq)
1	11.47 (1.00)	11.44 (1.00)
2	10.17 (0.89)	10.31 (0.90)
3	9.84 (0.86)	10.02 (0.88)
4	9.31 (0.81)	9.37 (0.82)
5	8.48 (0.74)	8.58 (0.75)
6	8.42 (0.73)	8.44 (0.74)
7	8.29 (0.72)	8.19 (0.72)
8	6.61 (0.58)	6.41 (0.56)
9	3.69 (0.32)	3.34 (0.29)
10	3.91 (0.34)	3.55 (0.31)

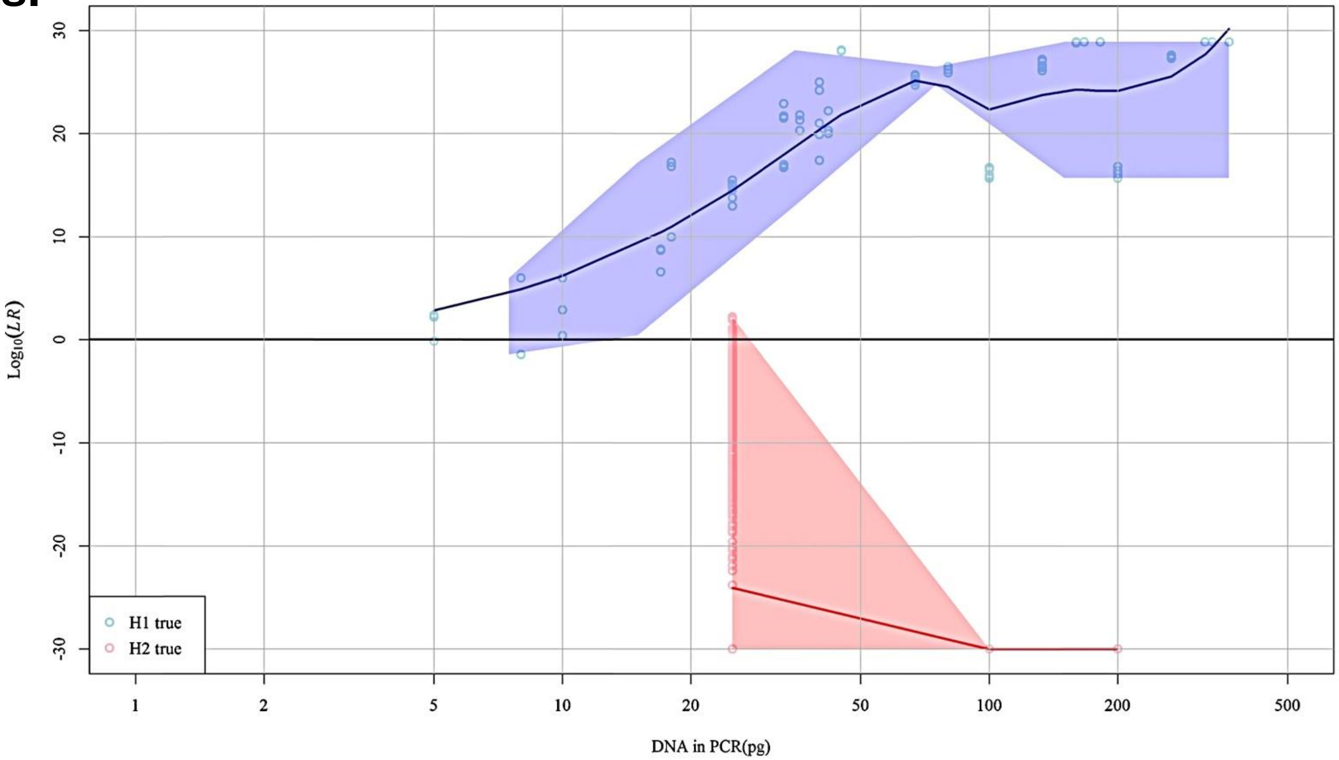
From :
Puch-Solis, R and Clayton T (2014) FSI Genetics Vol 11 pg 220-228

3. Performance of model (Empirical testing)

- Generate in vitro mixtures using extracted DNA from contributors with known genotypes
- These are the ‘true contributors’
- Generate a ‘panel’ of known non-contributors e.g. from randomly simulated genotypes from a frequency database
- *Nc* = 2,3 and 4 person
- Vary the mixing proportions
- Vary the input of total DNA into PCR
- Compute LR for true contributors (blue)
- Compute LRs for non-contributors (red)
- Plot out data graphically

STRmix

Fig. 1
Experiment 1 – LRs produced for two person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points.



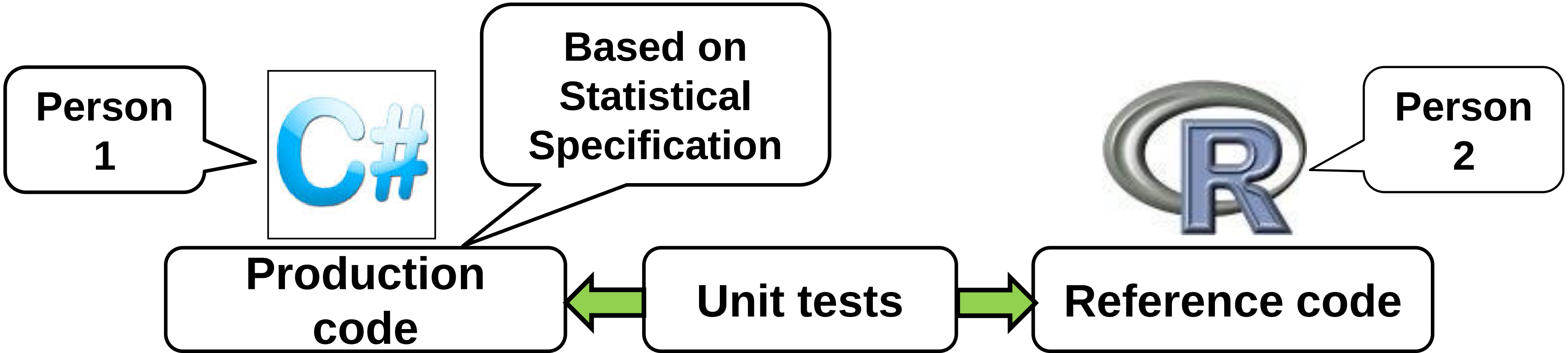
From : Taylor, D. (2014) FSI Genetics. Vol 11 pg 144-153

4. Computer implementation verification

Unit testing method

Unit: smallest testable piece of code

Bugzilla prevention !!



Person 1

Based on Statistical Specification

Person 2

Production code

Unit tests

Reference code

Computation times

- The calculations are thirsty!
- LiRa on a standard 4 core processor PC
- *Nc* = 1 seconds
- *Nc* = 2 minutes
- *Nc* = 3 depends upon number of unknowns
  - One unknown = minutes
  - Two unknowns = few hours
  - 3 unknowns = many hours
- *Nc* = 4 days

LiRa – run time on one PC

Hp : S + U + U
Hd : U + U + U

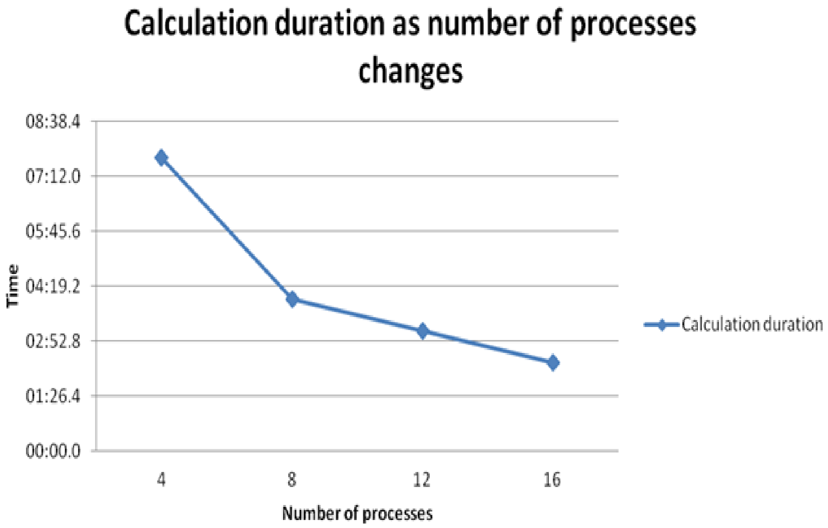
```
RESULTS
=======
Elapsed time:       1.09:40:42.2136823
Numerator:          1.16487995753364E-39
Denominator:        3.92360630592709E-46
LR:                 2968901.22684822
```

33hrs to run – computer crashed on first attempt!!

Long computation times (design options)

- Tie up a PC for as long as it takes
- PC clusters (hundred $)
- Purchase server(s) (thousand $)
- Export and purchase server time for time-consuming calculations (Cloudbursting)
- Use a cloud service (e.g. Azure)
  - Pay-as-you-go (tens of $ per calculation)
- To export a calculation you need to design the software so that the calculation can be broken up into pieces
- LiRa computation can be broken into pieces and can thus be exported

Adding processors by using a PC cluster array reduces computation time



Number of processes	Duration	% of baseline time
4	07:41.1	100
8	03:58.2	52
12	03:09.1	41
16	02:19.6	30

NB For processes read processors

Number of contributors and proposition selection

- Need the ability to user specify
  - *Nc*
  - Hp
  - Hd
  - Conditioning (known or assumed contributors)
  - Any relatedness between a POI and an unknown

Propositions for Nc = 2

Hp	Hd
POI & U	U1 & U2
POI1 & POI2	U1 & U2
K & POI	K & U1

Propositions for Nc = 3

Hp	Hd
POI & U1 & U2	U1 & U2 & U3
POI1 & POI2 & U1	U1 & U2 & U3
POI1 & POI2 & POI3	U1 & U2 & U3
K & POI & U	K & U1 & U2
K & POI 1 & POI2	K & U1 & U2
K & K & POI	K & K & U

Propositions for Nc = 4

Hp	**Hd**
POI & U1 & U2 & U3	U1 & U2 & U3 & U4
POI1 & POI2 & U1 & U2	U1 & U2 & U3 & U4
POI1 & POI2 & POI3 & U1	U1 & U2 & U3 & U4
POI1 & POI2 & POI3 & POI4	U1 & U2 & U3 & U4
K & POI1 & U1 & U2	K & U1 & U2 & U3
K & POI1 & POI2 & U	K & U1 & U2 & U3
K & POI1 & POI2 & POI3	K & U1 & U2 & U3
K1 & K2 & POI1 & U1	K1 & K2 & U1 & U2
K1 & K2 & POI1 & POI2	K1 & K2 & U1 & U2
K1 & K2 & K3 & POI	K1 & K2 & K3 & U

With thanks to the architects of LiRa ....

Marvellous Mexican Maths Maestro
Forensic statistican - Roberto Puch-Solis

Computer wizardry and software development - Ricky Young and Matt Baron

And finally......

“The Answer to the Great Question... Of Life, the Universe and Everything... Is... Forty-two,' said Deep Thought, with infinite majesty and calm.”

— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

“Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"
"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is.”
— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

Or.....
to understand the answer is to understand the problem

Thanks for listening and happy computing

Dr Roberto Puch-Solis
Statistician
LGC Forensics
Unit 3, Drayton Manor Business Park,
Tamworth, Staffs. B78 3GL. UK
Direct Line: +44 (0)1827 266994
Email: Roberto.Puch-Solis@lgcgroup.com

Dr Tim CLAYTON MBE
Forensic Biologist
LGC Forensics
Sir Alec Jeffries Building, Peel
Avenue, Calder Park, Wakefield, West
Yorks. WF2 7UA. UK
Direct Line: +44 (0)1924 241746
Email: Tim.Clayton@lgcgroup.com