Data Hygiene

Remco R. Bouckaert
remco@cs.auckland.ac.nz

Centre of Computational Evolution, University of Auckland
Max Planck Institute for the Science of Human History
Computer Science Department, Waikato University

Indo European – quality control

- Ascertainment correction bug
  - found after publication
- Calibrations wrongly scaled
  - found by visualising prior in DensiTree
- IELex export broken
  - found after innovations were generated
- Languages set changed, constraints broken
  - found after setting up different analysis

Open questions:

- Are priors correctly coded (tree priors, tip & clade calibrations)?
- Who has the time to check? Unrewarding, tedious work + fatigue (I stopped checking after the nth IELex export)
- Automating sanity checks to the rescue?

Automated sanity check 1 – cognate patterns

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

cognate list

language	hand	mano	mother	papa	father	mama	...
English	1	0	1	0	1	0	...
Dutch	1	0	1	0	1	0	...
German	1	0	1	0	1	0	...
French	0	1	1	0	1	0	...
Spanish	0	1	1	0	1	0	...
Dhudhuroa	?	?	0	1	0	1	...

Multiple language covered by duplicate pattern

1 stab (Bengali Hindi Oriya) 11 12

2 stone (Breton Cornish Welsh) 6 12

3 there (Ancient Greek Greek Greek Lesbos) 5 12

4 tree (Friulian Romansh) 18 19

5 vomit (Nepali Sindhi) 12 14

6 woods (Tocharian A Tocharian B) 15 16

Single language covered by duplicate pattern

1 all (Assamese) 27 28

...

6 at (Armenian) 36 41 42 45

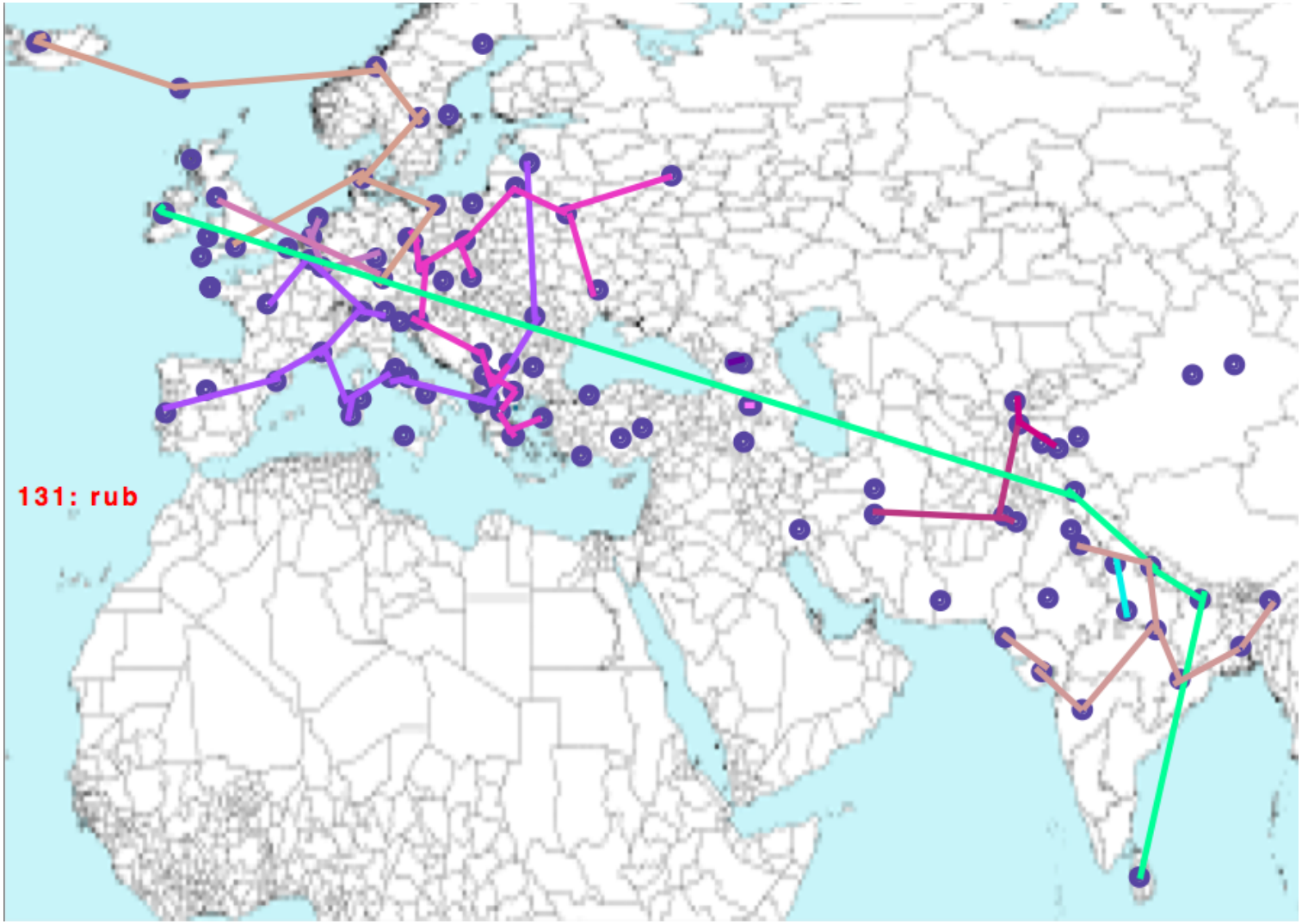
...

34 breast (Shughni) 13 16 17 18

...

234 woman (Shughni) 24 25

Automated sanity check 2 – Spanning trees



Automated sanity check 2 – Spanning trees

Spanning tree of cognate having branch $\geq$ 6000km

31 dirty 1391 6203 2 (Tocharian_B) (French, Provencal, Old_Irish, Irish, Scottish_Gaelic)

34 dry 2492 6607 2 (Tocharian_A, Tocharian_B) (Catalan, Spanish)

37 ear 2255 6261 2 (Tocharian_A, Tocharian_B) (Old_Irish, Welsh, Irish, Scottish_Gaelic)

57 fly 2504 6610 2 (Tocharian_A, Tocharian_B) (Old_Irish)

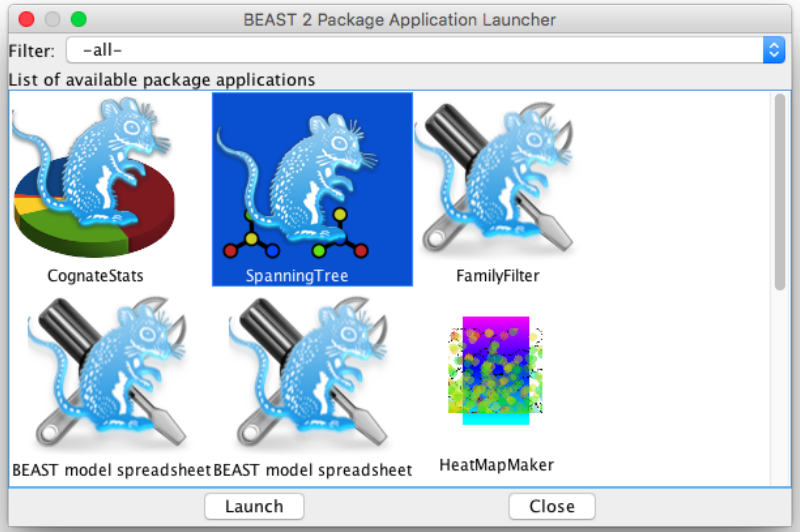
131 rub 1738 6595 2 (Nepali, Bihari, Kashmiri, Sinhalese) (Old_Irish, Irish)

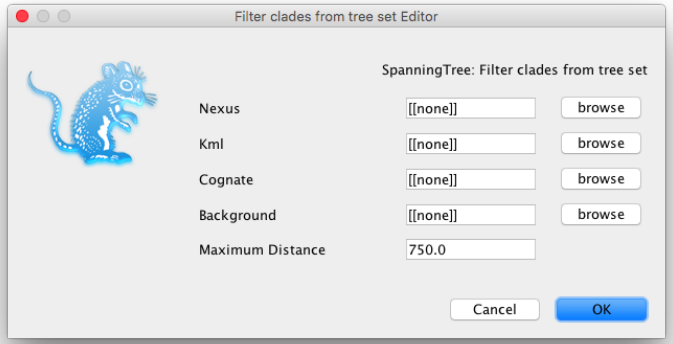
155 split 2956 6471 2 (Assamese, Oriya) (Old_Prussian, Frisian, German)

164 sun 2299 6261 2 (Tocharian_A, Tocharian_B) (Old_Irish, Irish, Scottish_Gaelic)

Use BEAST 2 Babel package, SpanningTree app

1. BEAUti menu File/Launch App
2. Fill in the form





Alternative: from a terminal

```
/path/to/beast/bin/appstore SpanningTree -nexus file.nex
    -cognate labels.txt  -kml languages.kml -background world-map.png
    -maximumDistance 4000
```

Results are reported in the terminal used to launch SpanninTree

File format

Cognate file:

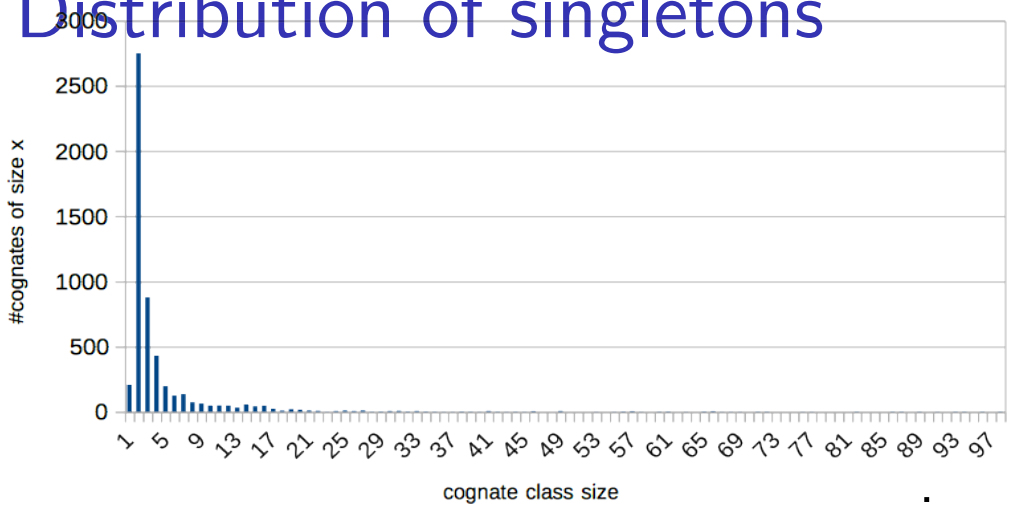
```
1 I_group,
2 I_cognate_62,
3 I_cognate_4343,
4 I_lexeme_28404,
5 I_lexeme_28780,
6 all_group,
7 all_cognate_175,
8 all_cognate_351,
9 all_cognate_543,
10 all_cognate_566,
11 all_cognate_682,
12 all_cognate_748,
```

Nexus file

```
...
matrix
[ ...some comments]
[ ...more comments]
'Hittite'  010001101010...
'Luvian'  001101010000...
'Lycian'  001101010000...
...
'Ancient_Greek'   010011000000.
   ;
end;
```

Automated sanity checks 3 – Distribution of singletons

Total number of 1s = 20494
Singletons = 2750 = 13.4% of data

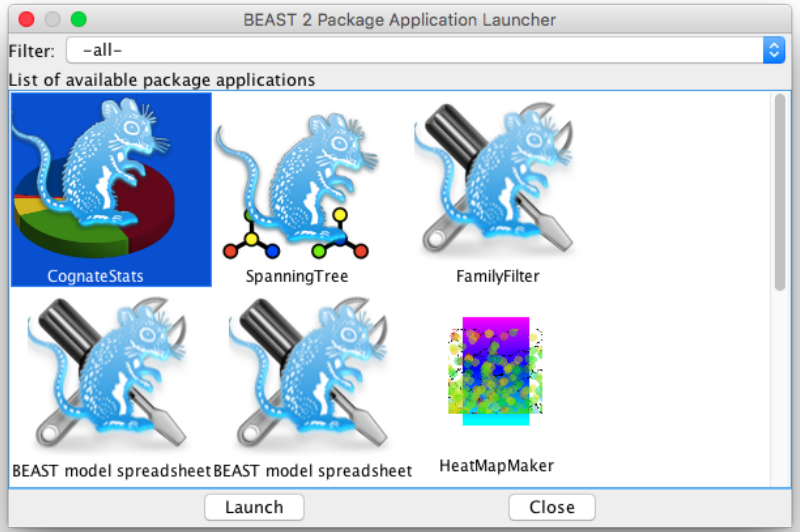


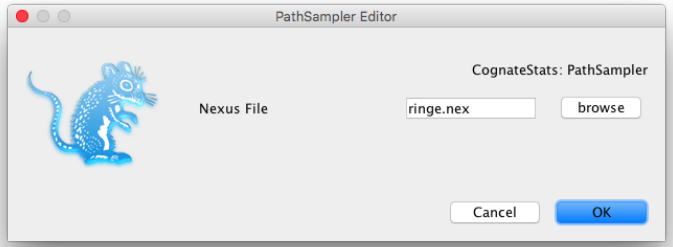
	total #singletons	#languages	average #singletons	max #singletons	min #singleto
Germanic	268	17	15.76	47	3
Slavik	367	16	22.94	68	2
Gaelic	138	6	23	30	17
Romance	347	15	23.13	48	6
Indo-Iranian	1053	29	36.31	111	2
Greek	577	15	38.46	81	7
All	2750	98	28.06	111	2

Greek & Indo-Iranian needs most attention

Babel package CognateStats app

1. BEAUti menu File/Launch App
2. Fill in the form



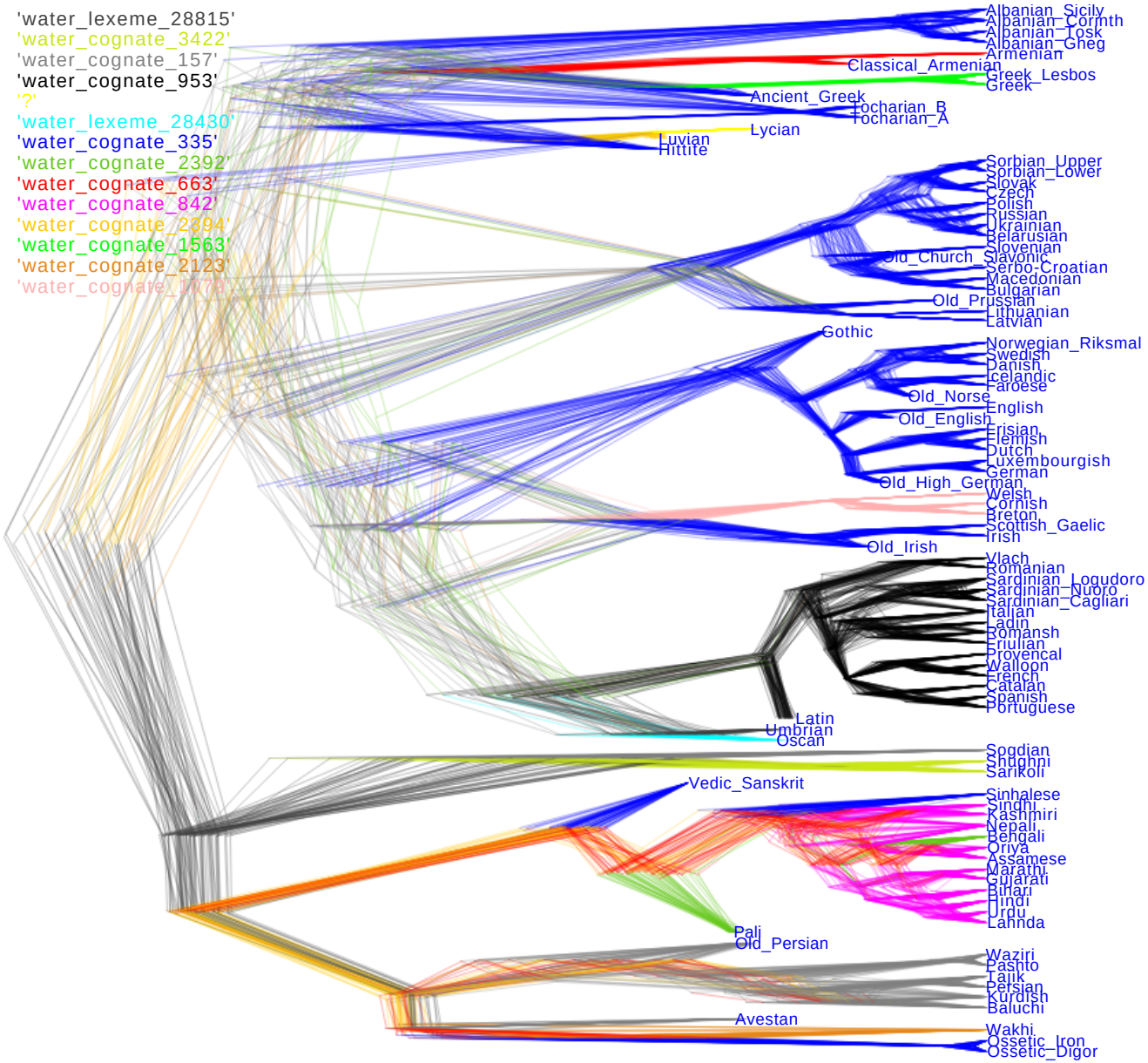



```
/path/to/beast/bin/appstore CognateStats -nexusFile file.nex
```

Output;

- a list of columns that are singletons
- a distribution of cognate patterns
- a list of duplicate cognate patterns

Sanity check 4 – (approx) most likely cognate on branches



Babel package

Set up analysis in BEAUti
Add logger by editing XML

```
<logger fileName='water.trees' id='treelog.water' logEvery='10000' mode='tree'>
   <log id='ancestral.water' spec='AncestralStateLogger2' tree='@Tree.t:ringe'
         useAmbiguities='true' branchRateModel='@RelaxedClock.c:ringe'
         data='@orgdata.water' siteModel='@SiteModel.s:water'>
          water_group water_cognate_157 water_cognate_335 water_cognate_663
          water_cognate_842 water_cognate_953 water_cognate_1079 water_cognate_1563
          water_cognate_2123 water_cognate_2392 water_cognate_2394
          water_cognate_3422 water_lexeme_28430 water_lexeme_28815
     </log>
</logger>
```

Use `DensiTree` to visualise the tree.
Interactive, or batch CLI:

```
java -jar DensiTree.jar -linecolortag site -linecolorlegend
     -b 10 -geo 1024x1024 -asPDF water.pdf water.trees
```

Sanity check 5 – alternative analyses

Disclaimer: All with wrong calibrations due to newly introduced languages

- grid of hyperpriors on birth death rate bounds
  - posterior root height always higher than prior
- Remove 486 duplicate columns (486 out of 20494 = 2.3% removed)
  treeHeight 8.16 [6.95 9.65]
- Remove 486 duplicate columns + randomly knock out 1011 1s (1497 out of 20494 = 7.3% removed)
  treeHeight 8.09 [6.70 9.51]
- Remove 486 duplicate columns + randomly knock out 2022 1s (2508 out of 20494 = 12.3% removed)
  treeHeight 8.06 [6.79 9.35]
- Remove 486 duplicate columns + randomly knock out 4044 1s (4530 out of 20494 = 22.2% removed)
  treeHeight 7.59 [6.46 8.86]
- Remove all singletons analysis (2750 out of 20494 = 13.4% removed)
  treeHeight 8.09 [6.70 9.69]
- Pseudo Dollo Covarion fits better than Covarion
  treeHeight 8.7 [7.7, 9.7]
- Phylogegraphy analysis points to Anatolian on summary tree
  - spherical diffusion
  - spherical diffusion + low rate through Caucasus
  - random walk on graph + Caucasus blocked

Any of these tend to bring out unexpected problems (e.g. newly added languages)

Summary

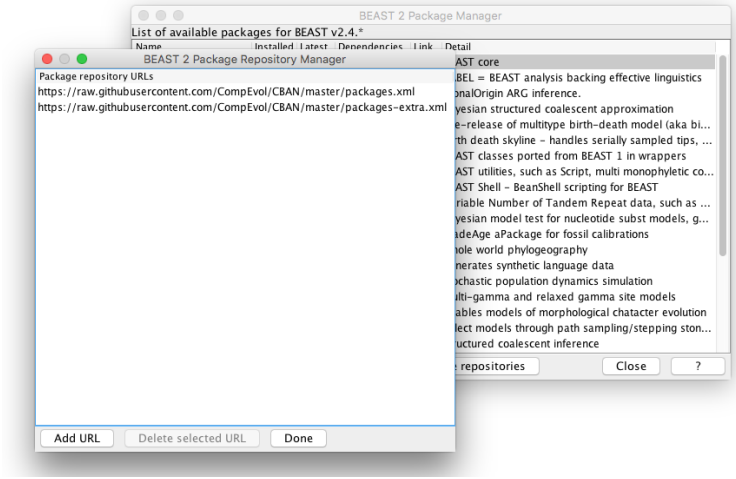
- Quality control is hard, since nobody has complete overview (not even over IE data alone!)
- Quality control is tedious, unrewarding + danger of analysis fatigue
- Unit tests for any software are mandatory, but not sufficient
- Automated sanity checks can help

Open questions:

- What else can we do to assure we do what we say we do?
- What about sanity checks for non-lexical data?
- What can be done to ensure tools are used (and make it worth developing them more)?

Installing BEAST packages

- BEAUti menu 'File/Manage packages', Press 'Package repositories'
- Add repository
  `https://raw.githubusercontent.com/CompEvol/CBAN/master/packages-extra.xml`



- Select package & press install button

BEAST package Babel – use to set up Change et al.

Prepare nexus

Specify partitions `begin assumptions;`

```
charset I = 1-5;
charset all = 6-39;
charset and = 40-75;
charset animal = 76-101;
...
end;
```

Specify calibrations

```
#Define monophyletic clades
begin sets;
taxset germanic = oldnorse oldhighgerman oldprussian oldenglish;
taxset tocharian = tocharian_a tocharian_b;
taxset anatolian = hittite lycian luvian;
end;
# Define time calibrations on tips, and clades.
begin assumptions;
calibrate oldnorse = normal(775,40)
calibrate avestan = normal(2500,50)
calibrate germanic = normal(1875,67)
end;
```

BEAST package Babel – use to set up Change et al.

In BEAUti select template

- BinaryCTMC
- BinaryCovarion
- SDollo (maybe not a good idea)

then import alignment.
Make sure ascertainment columns (first column of each partition) are in the alignment.

BEAST package LanguageSeqGen

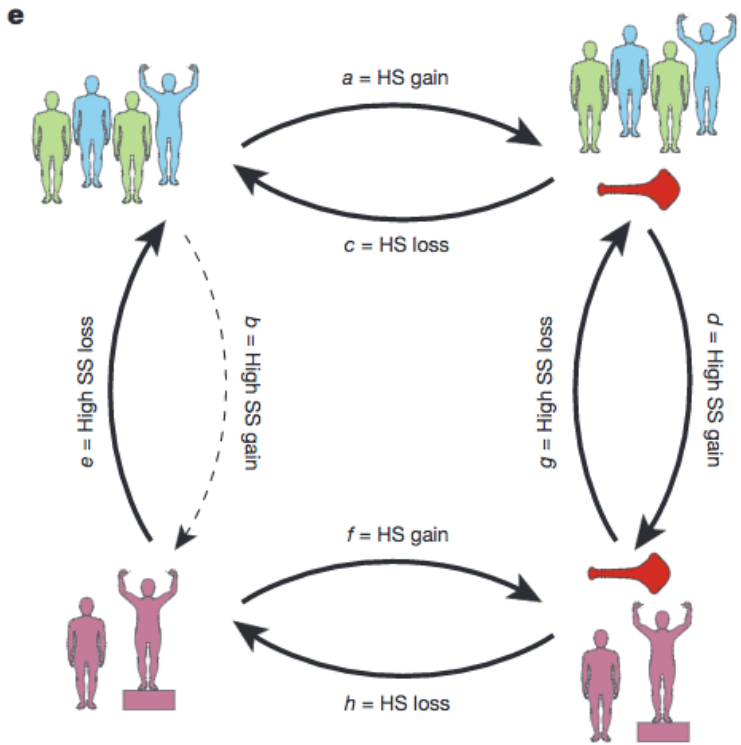
Simulate language data along a tree

- CTMC
- BinaryCovarion
- SDollo

Allows borrowing

BEAST package correlatedcharacters

Substitution model that tells whether two (binary) characters are independent or not



Watts et al. Nature (2016)

Questions?