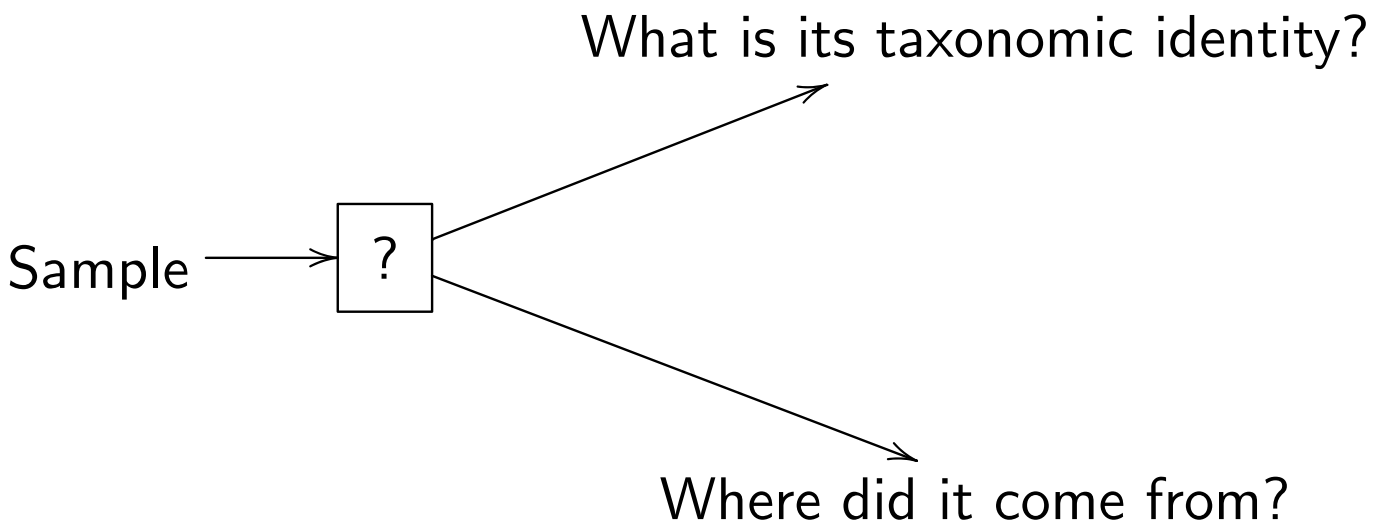
Identifying Samples and their Sources: Case Studies and Lessons Learned

Brook Milligan
Conservation Genomics Laboratory
Department of Biology
New Mexico State University
Las Cruces, New Mexico 88003 USA
brook@nmsu.edu

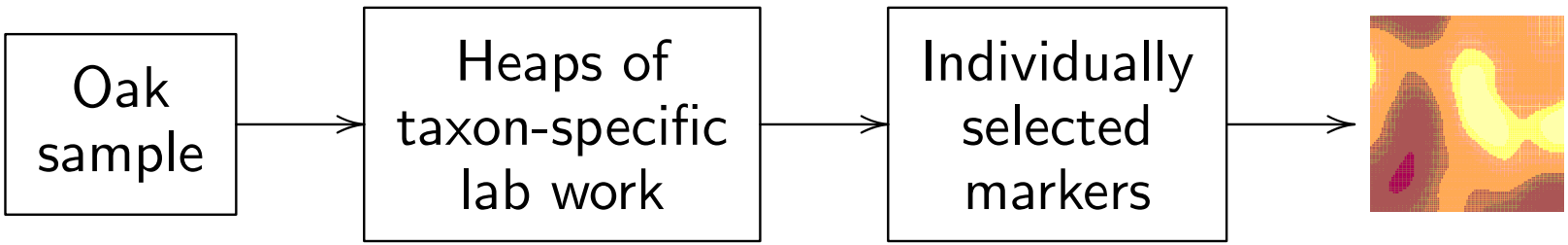
Development and Scaling of Innovative Technologies for Wood Identification
February 28, 2017

The questions we face

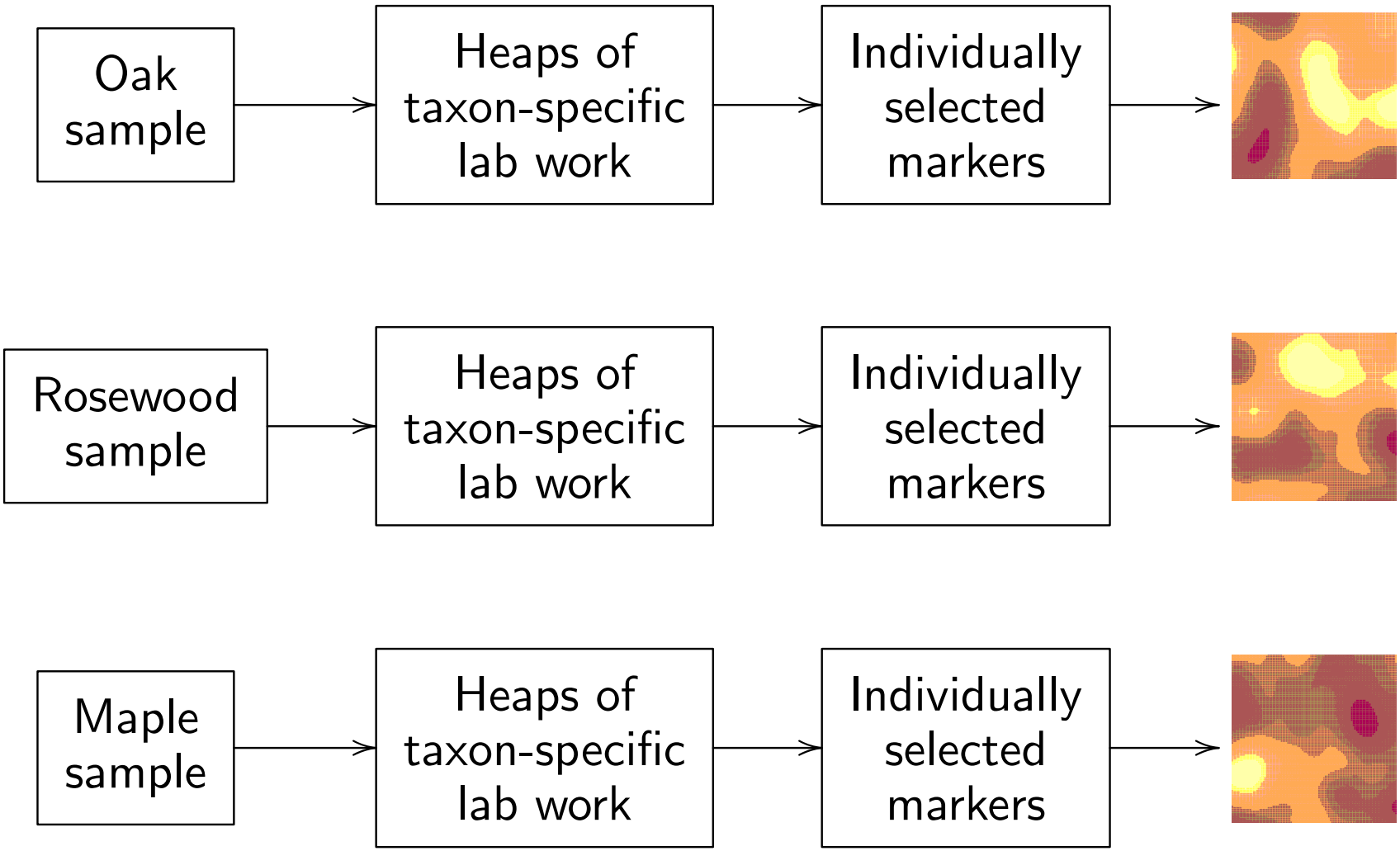


- Case studies
  - Taxonomic identification via direct comparison with a database
  - Taxonomic identification via inference
  - Geographic origin identification via inference
- Lessons learned
  - Direct comparison is of limited usefulness
  - Inference is essential for taxonomic and geographic origin identification
  - These lessons apply to all identification methods, not just DNA

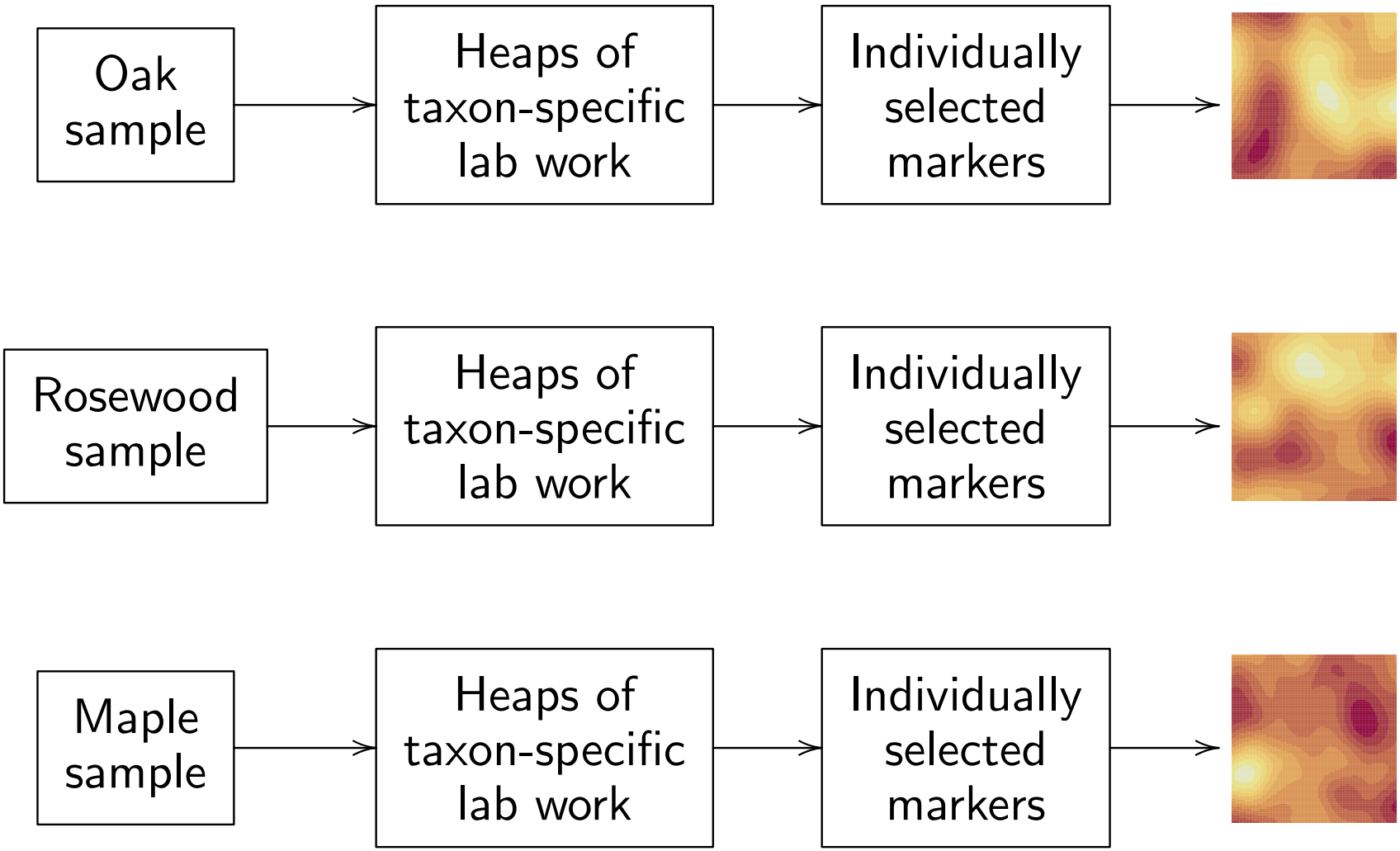
Traditional genetics: a cottage industy



Traditional genetics: an *inefficient* cottage industry



Traditional genetics: a *simplified* world view



Genomics: industrialization and economies of scale

Dividing a sequence into k-mers

```
5′- ...gacaccatcgaatggcgcaaaacctttcgc...  -3′
3′- ...ctgtggtagcttaccgcgttttggaaagcg...  -5′
```

Dividing a sequence into k-mers

5′- ...gacaccatcgaatggcgcaaaacctttcgc... -3′

⋮

ccatcgaatg

catcgaatgg

atcgaatggc

tcgaatggcg

cgaatggcgc

gaatggcgca

aatggcgcaa

atggcgcaaa

tggcgcaaaa

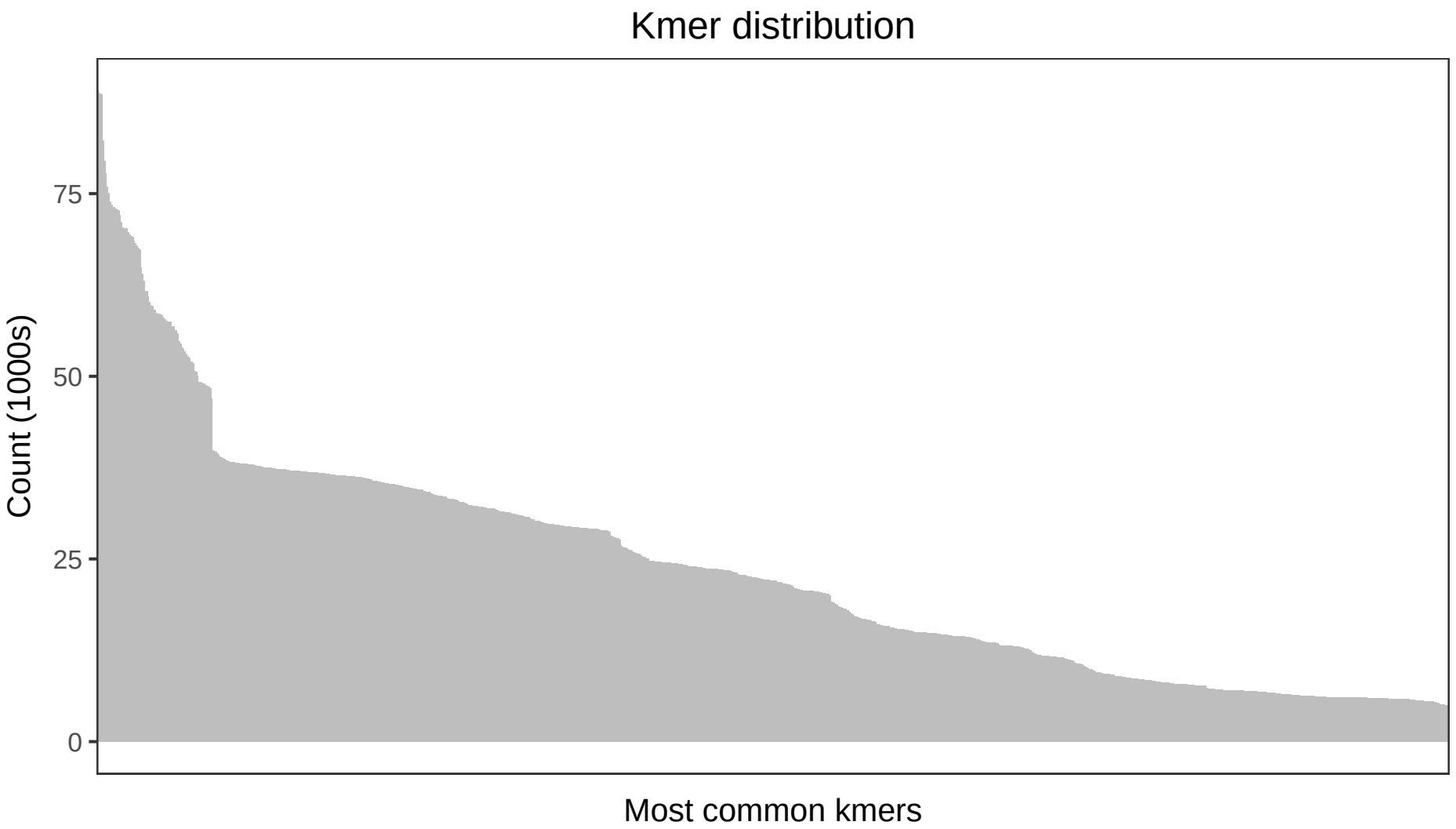
ggcgcaaaac

gcgcaaaacc

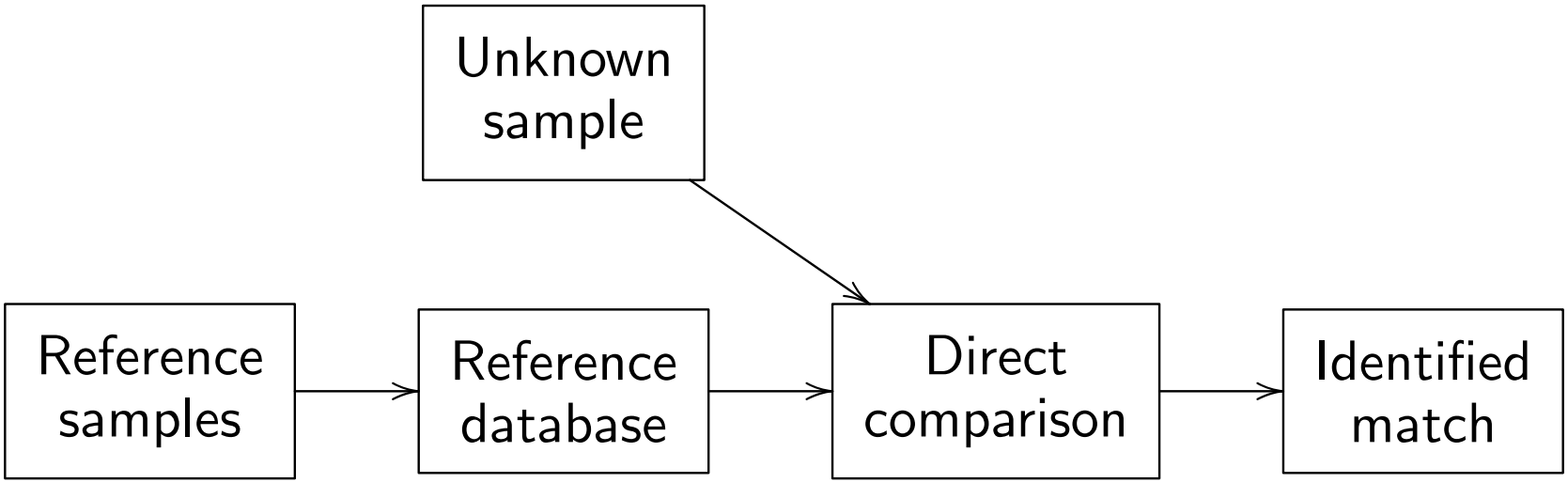
cgcaaaacct

⋮

Distribution from common to rare kmers



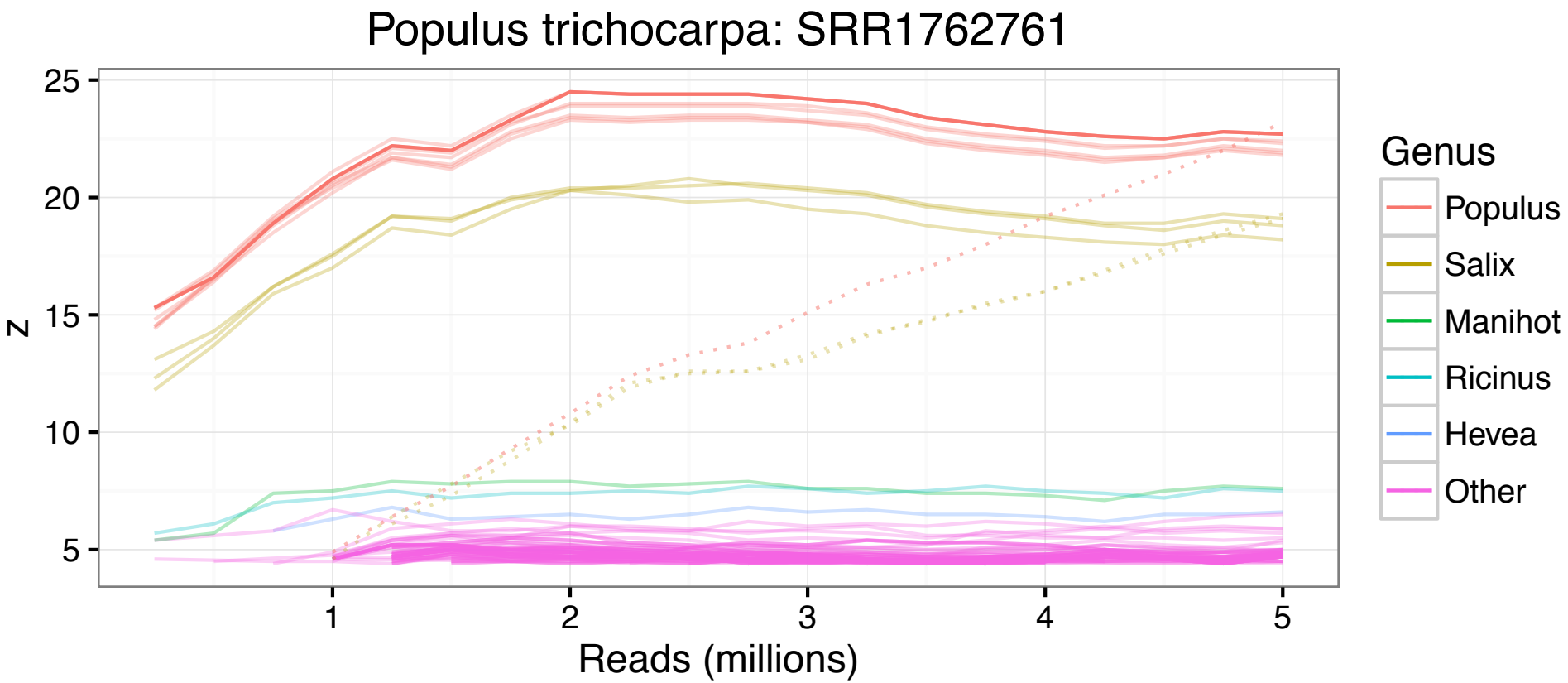
1. Identification by directly matching reference samples



Testing genomic identification with a diversity of plants

	Taxon	Genome size	Accessions
1	Cycas revoluta	13399 MB	1
2	Lactuca sativa	2592 MB	1
3	Magnolia yunnanensis	880-5844 MB (genus)	1
4	Oryza sativa	489 MB	3
5	Picea abies	19570 MB	2
6	Pinus taeda	21614 MB	3
7	Populus alba	509 MB	1
8	Populus balsamifera	440-528 MB (genus)	1
9	Populus tremula	440 MB	1
10	Populus trichocarpa	484 MB	18
11	Prunus armeniaca	293 MB	1
12	Prunus davidiana	303 MB	1
13	Prunus dulcis	323 MB	1
14	Prunus ferganensis	269-3570 MB (genus)	1
15	Prunus kansuensis	293 MB	1
16	Prunus mume	269-3570 MB (genus)	2
17	Prunus persica	269 MB	2
18	Prunus serotina	489 MB	1
19	Quercus mongolica	489-978 MB (genus)	1
20	Solanum lycopersicum	1002 MB	3

Match score


Populus trichocarpa: SRR1762761
25
20
15
10
5
N
1
2
3
4
5
Reads (millions)
Genus
Populus
Salix
Manihot
Ricinus
Hevea
Other

1. Identification by directly matching reference samples

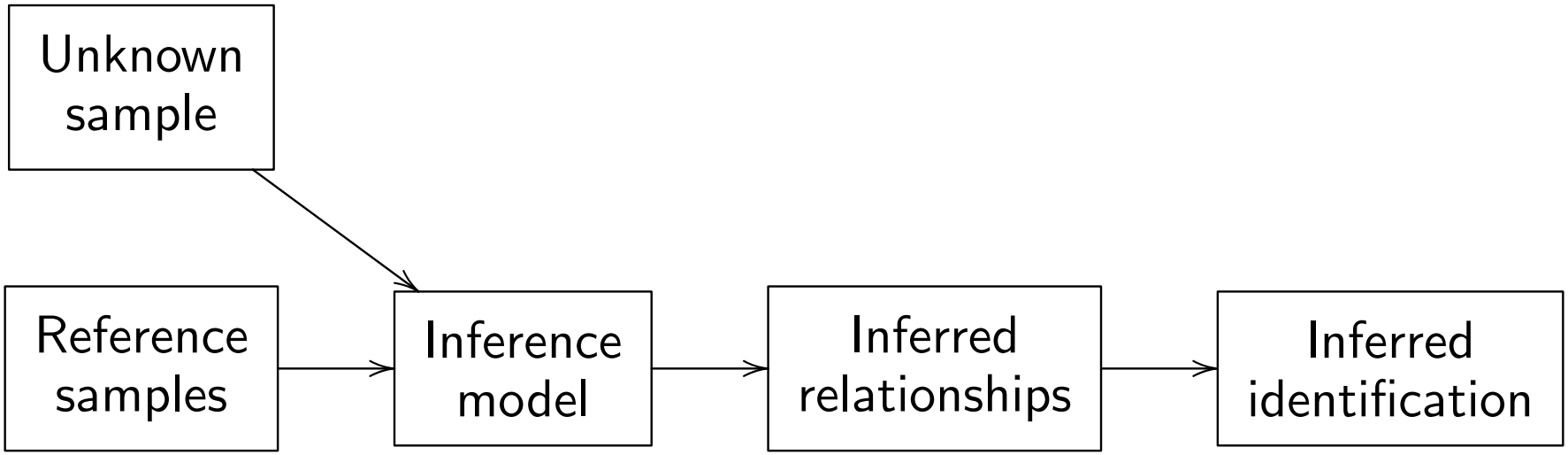
Highly reliable Consistently good performance across all comparisons irrespective of taxonomic distance

Very discriminatory Closely related taxa can be distinguished

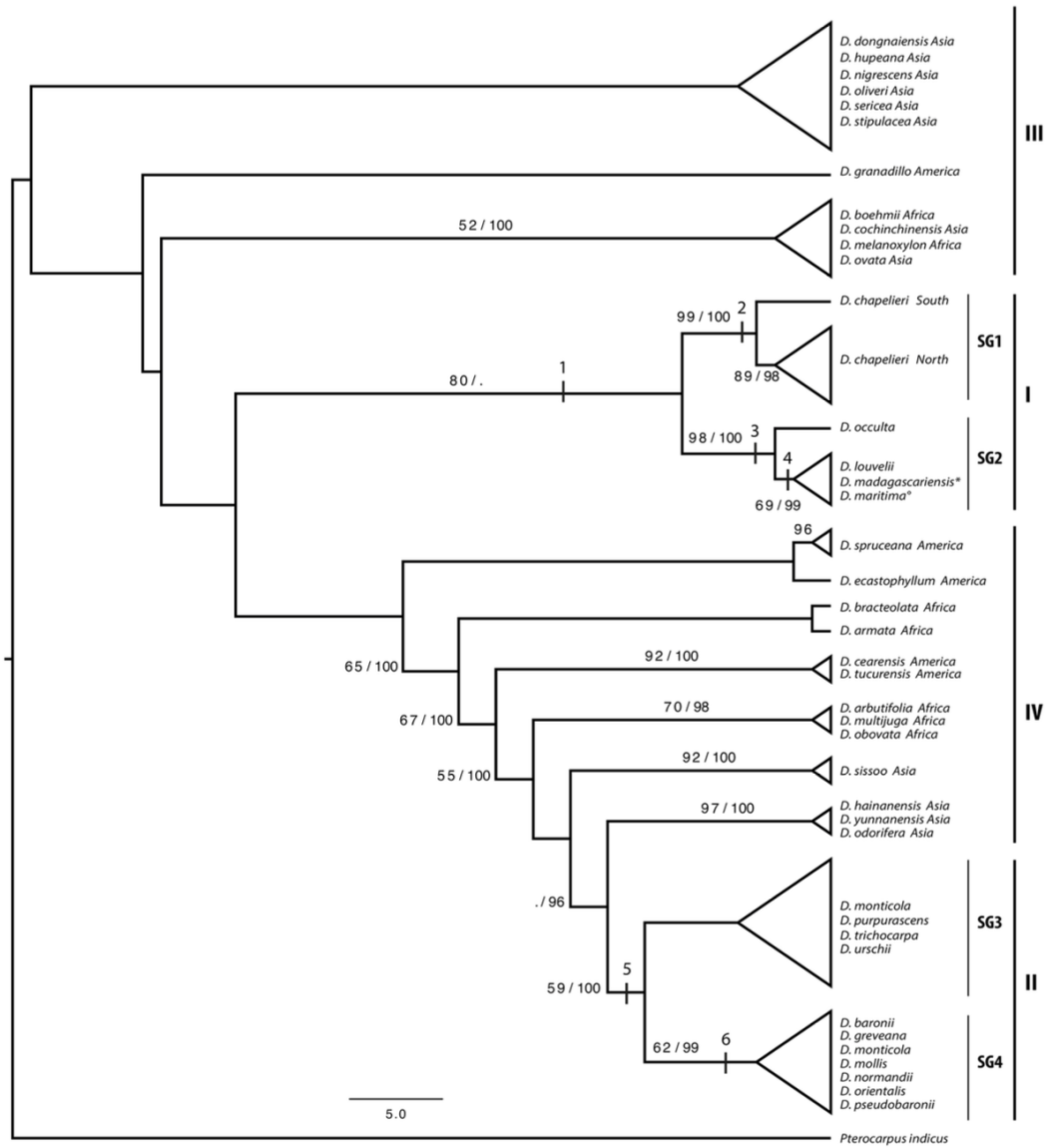
Efficient Relatively little data is required

Requirements A complete reference database for positive identification

2. Identification of taxa using formal inference

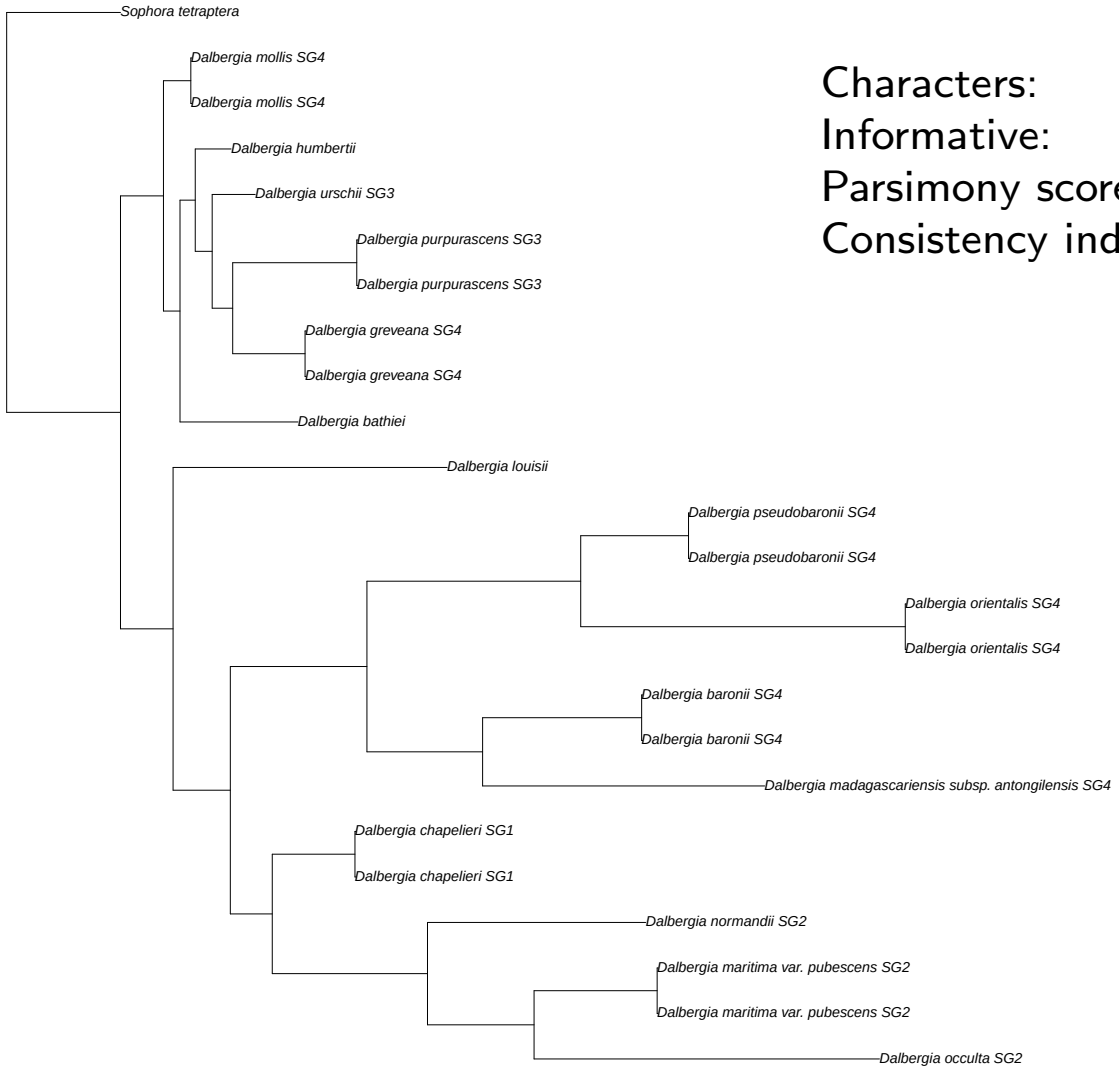


Phylogenetic relationships of Malagasy Dalbergia



Hassold et al. (2016), Figure 2

Phylogenetic relationships of Malagasy Dalbergia



Characters:	185,032
Informative:	185,018
Parsimony score:	226,052
Consistency index:	81.8%

2. Identification of taxa using formal inference

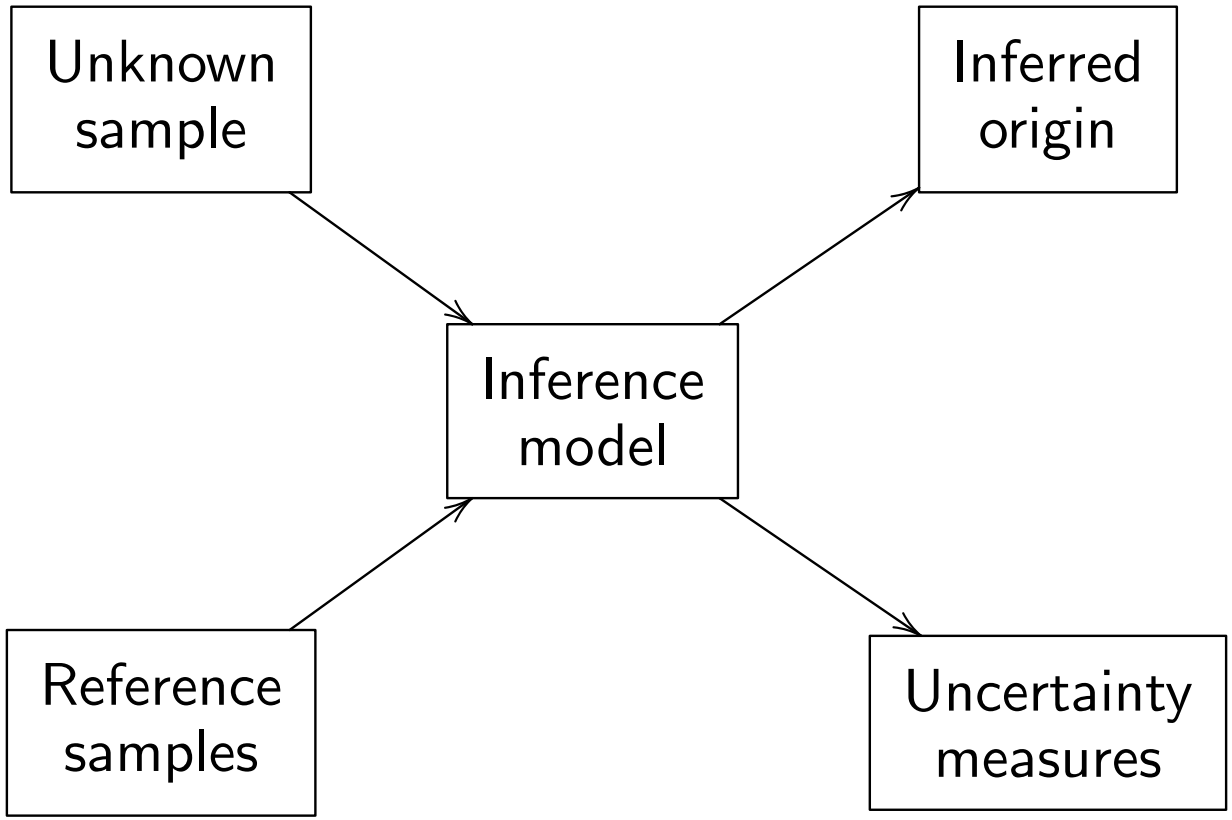
Phylogenetic relationships of Malagasy Dalbergia:

- Species-specific groups are recovered
- Chloroplast species groups 1 and 2 (SG1 and SG2) are identified
- Geographic regional groupings are identified
- Inconsistencies between genomic and chloroplast trees in less well-supported regions

General taxonomic identification:

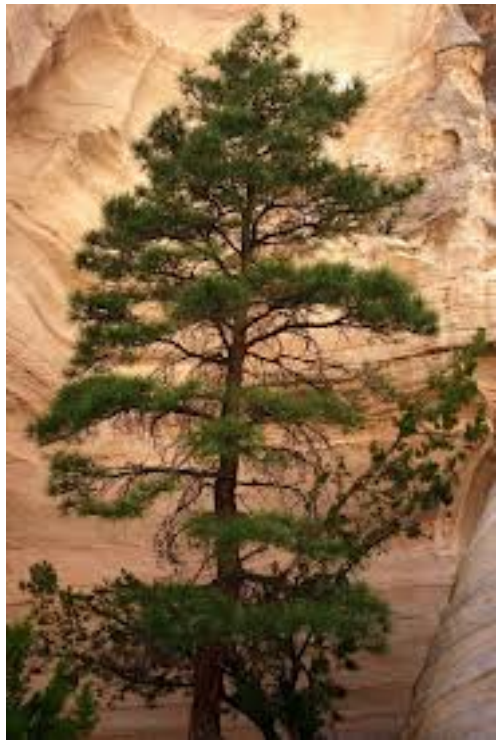
- Genomic data contain phylogenetically informative information
- Inference models are needed when a complete database does not exist, i.e., *always*

3. Identification of geographic origin using formal inference

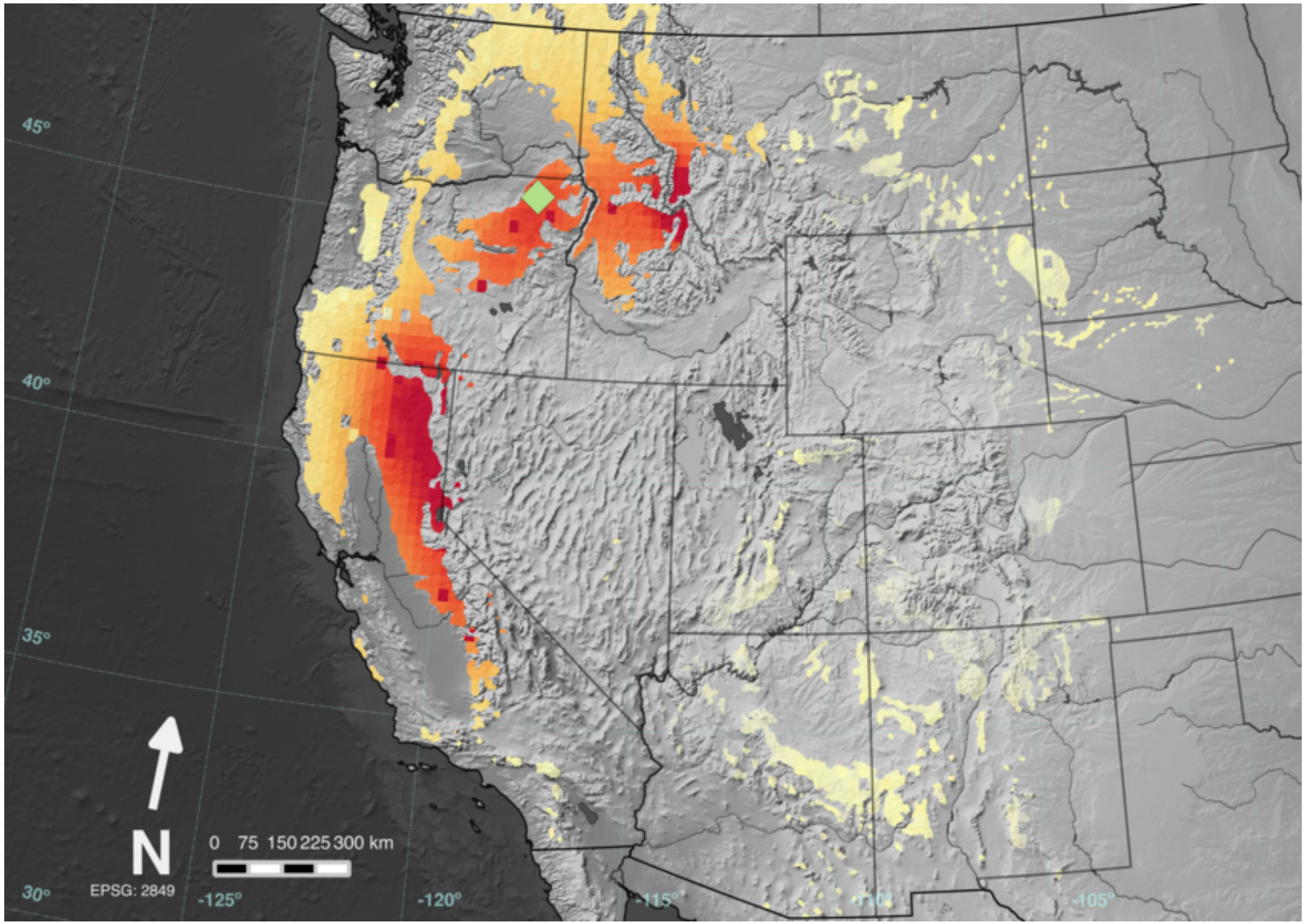


Identifying geographic origins: *Pinus ponderosa*

Third in production volume (4.5 million m^2); second in value (WWPA 2001)



Identifying geographic origins using models of differentiation: *Pinus ponderosa*



3. Identification of geographic origin using formal inference

Very discriminatory	Inferred origin can be reduced to a small region
Useful	Quantitative measures of uncertainty are available
Efficient	Relatively little data is required
Flexible	Multiple types of data, i.e., not just genetic, may be integrated to improve accuracy

Lessons learned: differences among identification strategies

Taxonomic identification:

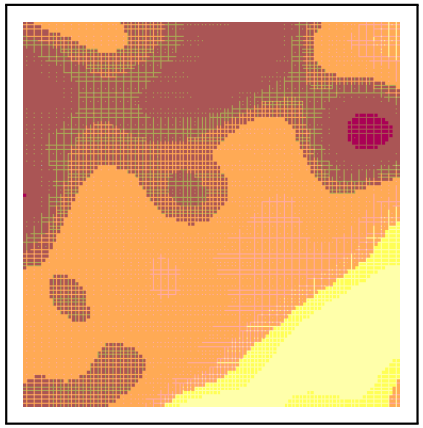
- Direct matching in a reference database
- Modeling evolutionary processes

Geographic origin:

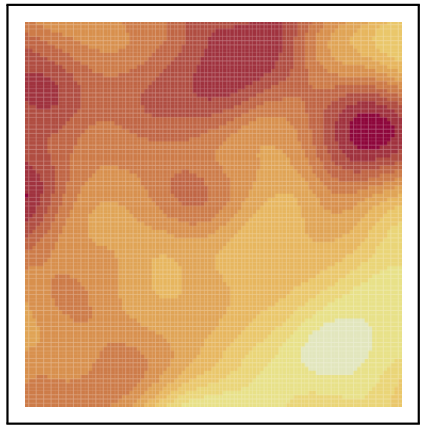
- Direct matching in a reference database
- Modeling spatial differentiation processes

Regardless of effort and expense, reference databases will always be incomplete relative to the questions that need answering. Therefore, gaps must be filled in with inference, which also yields the benefit of learning about uncertainty.

The real world is messy!



Clearly identifiable categories



Continuous gradation

To be useful, tools we develop must match the real world, not what we consider to be conceptually convenient.

The importance of inferential analysis transcends genomics and applies to all methods of identification.

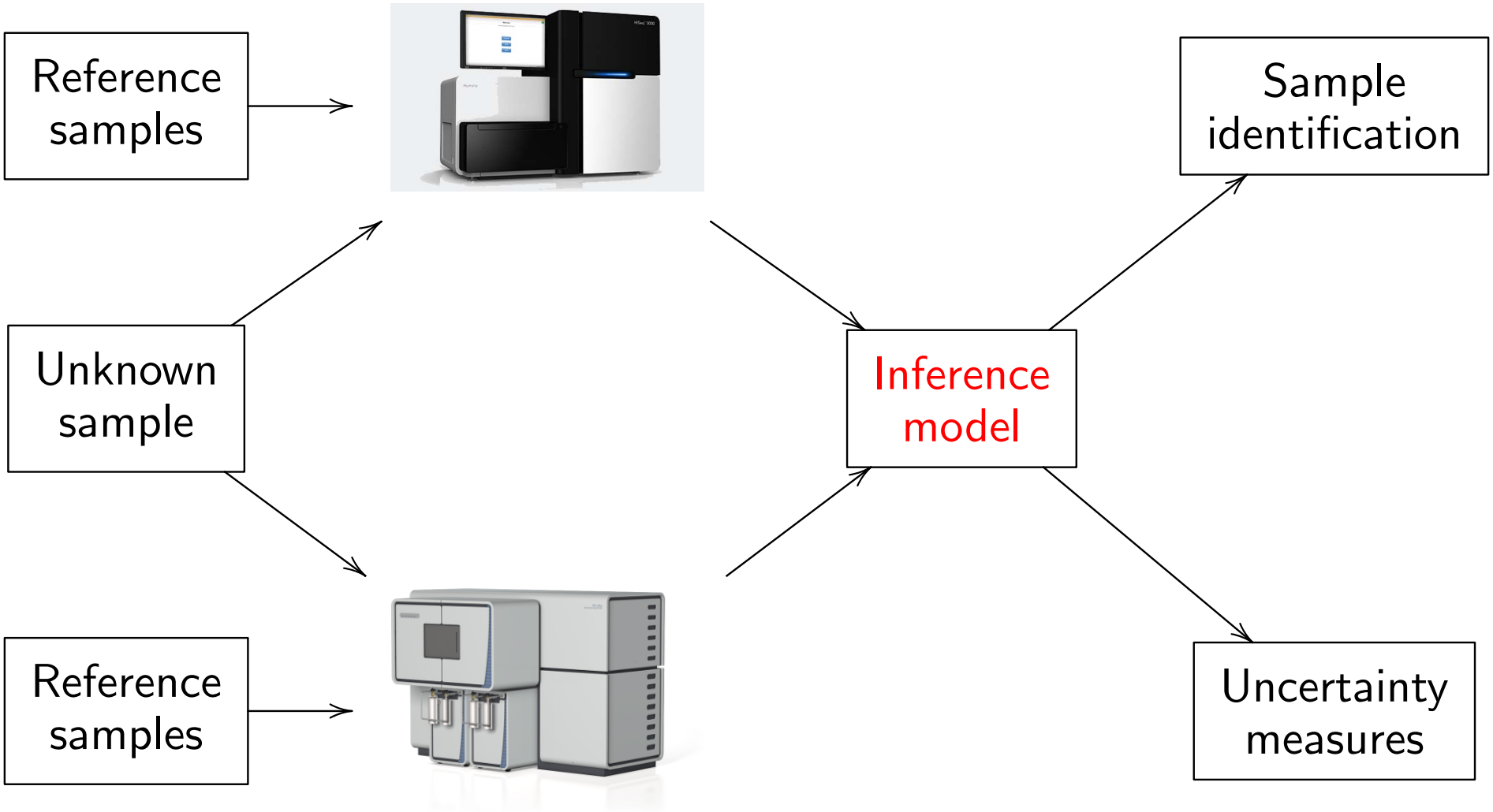
Validity and reliability must be assured

President's Council of Advisors on Science and Technology (2016):

- There is a need to evaluate specific forensics methods to determine their validity.
- Of six types of identification methods examined, only one was deemed valid and applied appropriately.
- Quantitative, inferential analysis is an appropriate methodology.

  All of the identification problems under consideration for wood would fall into the "difficult to validate" cases identified by PCAST (2016). Therefore, inferential analysis must play a central role, not just for genomics analysis, but for every method.

Identification: a general strategy of integrating information



Thanks to . . .

- Meaghan Parker-Forney and the World Resources Institute
- David Erickson, DNA4 Technologies
- Alex Widmer and Simon Crameri, ETH Zürich
- Valerie Hipkins, USFS National Forest Genetics Laboratory