RANDOM FORESTS OF FORESTS:
integrating data sources to combat illegal logging

Rich Cronn [1], Kristen Finch [2], Ed Espinoza [3], Andy Jones [2]

[1] Pacific Northwest Research Station, US Forest Service, Corvallis, OR

[2] Botany and Plant Pathology, Oregon State University, Corvallis, OR

[3] US Fish and Wildlife Forensics Laboratory, Ashland, OR

RANDOM FORESTS: AN OVERVIEW

integrating data sources to combat illegal logging

1. A moment to state the obvious
2. Random Forests
3. Integrating data
4. Concluding remarks

1. A MOMENT TO STATE THE OBVIOUS…

identifying the taxonomic and geographic source of wood is challenging

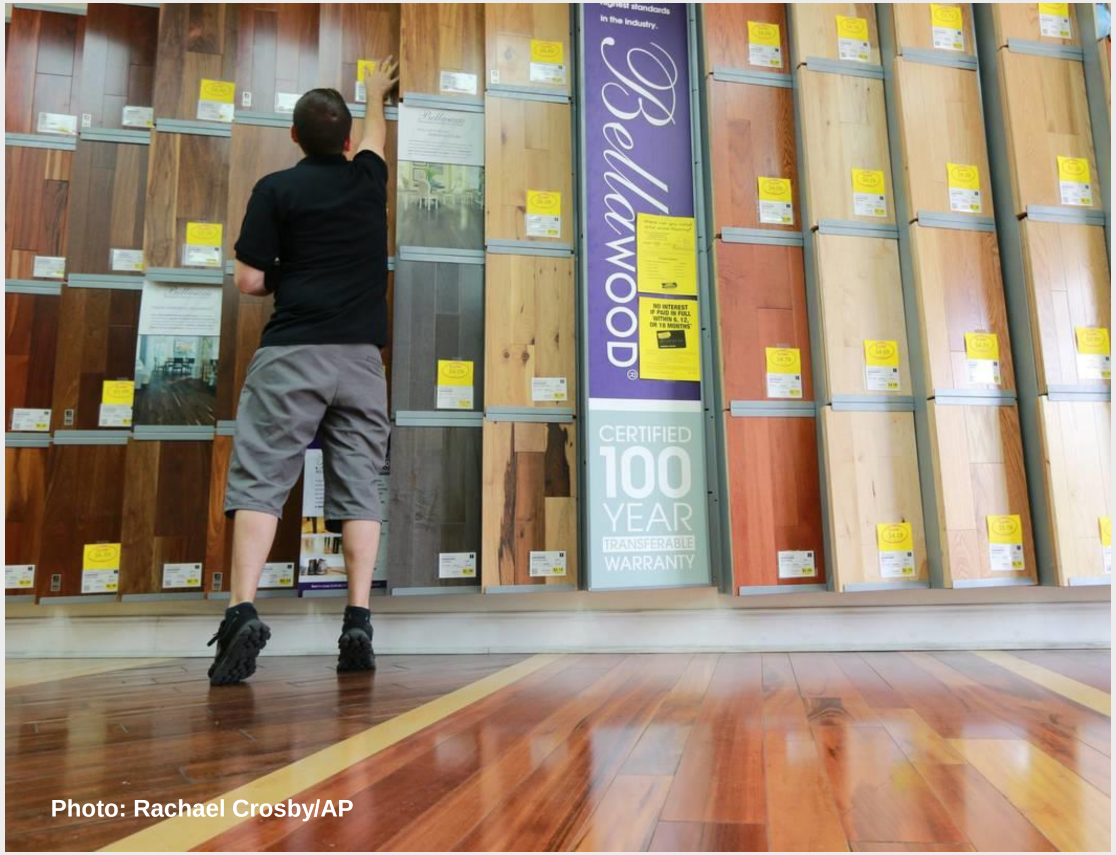
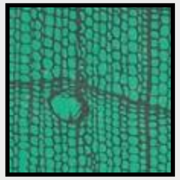


Photo: Rachael Crosby/AP

- Trees are genetically complex
- Trees are long lived, and have overlapping generations
- Trees share genetic information over long temporal and geographic spans
- Genetic complexity influences metabolic and anatomic traits, and these influence taxonomic complexity

ADDRESSING THE CHALLENGE

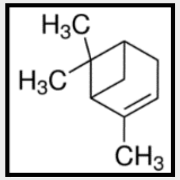


- Anatomy

  Cellular (rays, vessels); color; scent

- Genetics

  Organelle genomes; nuclear SNPs



- Chemicals

  Metabolites; spectra

$\delta^{14}N$

- Stable isotopes

ELSEVIER

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/bioc

Discussion

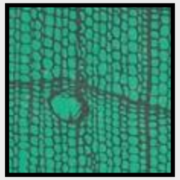
Forensic timber identification: It's time to integrate disciplines to combat illegal logging

CrossMark

Eleanor E. Dormontt[a], Markus Boner[b], Birgit Braun[c], Gerhard Breulmann[d], Bernd Degen[e], Edgard Espinoza[f], Shelley Gardner[g], Phil Guillery[h], John C. Hermanson[i], Gerald Koch[j], Soon Leong Lee[k], Milton Kanashiro[l], Anto Rimbawanto[m], Darren Thomas[n], Alex C. Wiedenhoeft[o], Yafang Yin[p], Johannes Zahnen[q], Andrew J. Lowe[a,*]

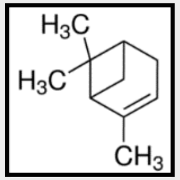
DATA INTEGRATION (?)

multiple methods = yes



Machine vision; classification trees; phylogenetic trees

Classification trees; 'barcode' phylogenetic trees; spatial-genetic interpolation

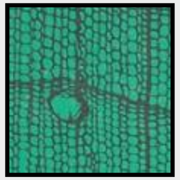


Least squares; discriminant analysis (linear; kernel; quadratic)

$\delta^{14}N$

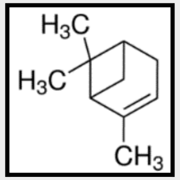
Least squares; discriminant analysis; k-nearest neighbor

integration = ?



+

+



+

$\delta^{14}N$

→ Data mining, machine learning, neural nets

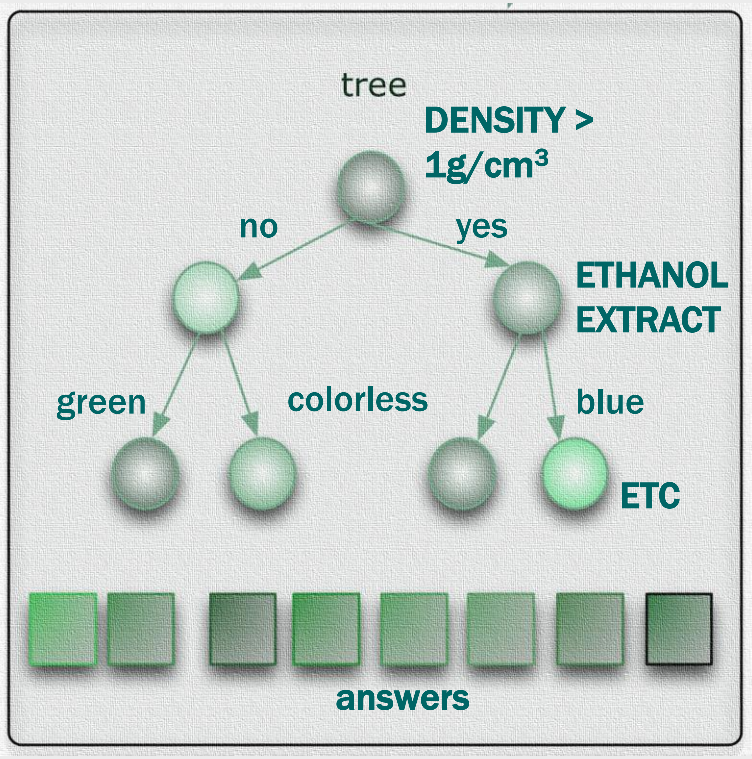
2. RANDOM FORESTS

a.k.a., “the best ‘black box’ method ever invented…”

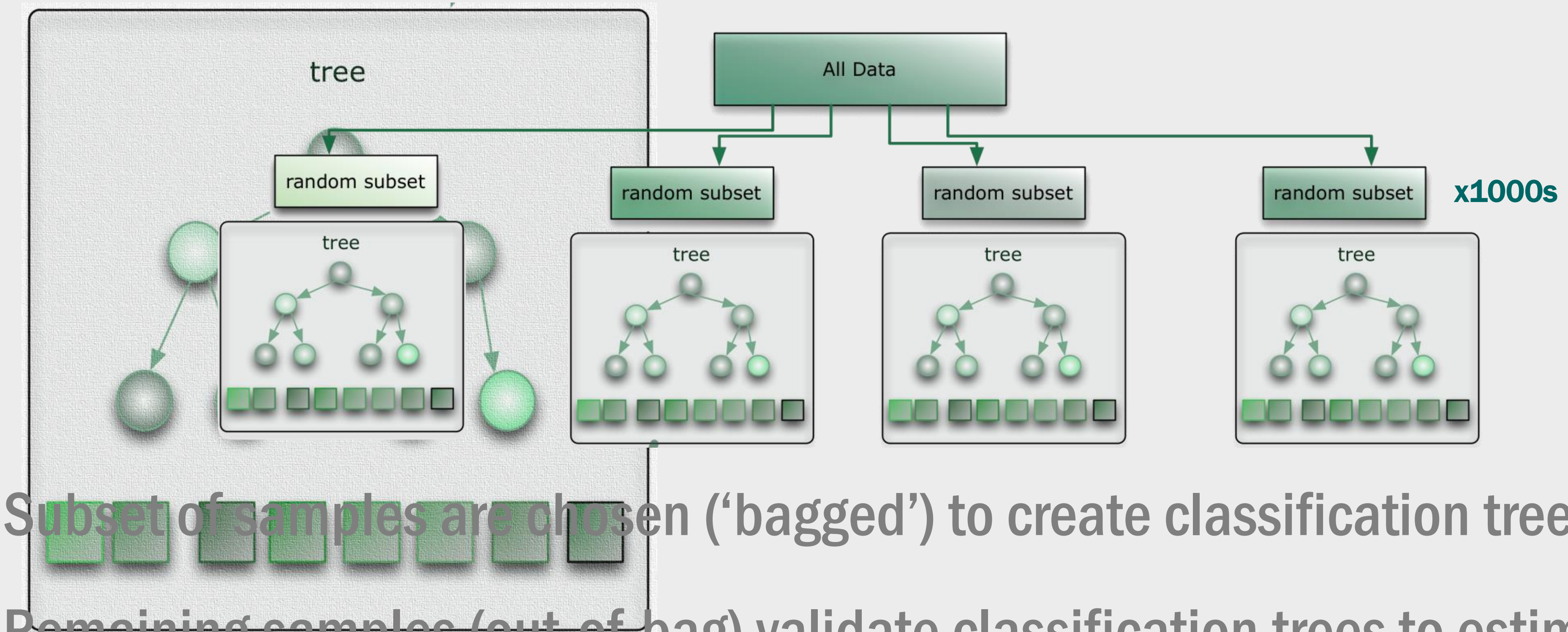
A versatile ensemble method – combines many models into one

- Can be used for simple or complex classification problems
- Handles large data sets, missing data, nearly any kind of data
- Directly identify features important in classification prediction

ONE CLASSIFICATION TREE



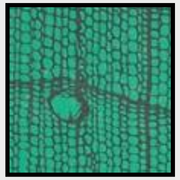
MANY RANDOM TREES = A ‘FOREST’

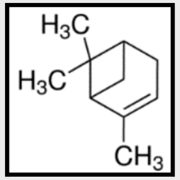


- Subset of samples are chosen (‘bagged’) to create classification trees
- Remaining samples (out-of-bag) validate classification trees to estimate error
- Classification model determined by ‘voting’ from all trees in the forest
- BONUS! Classification variables are ranked by ‘importance’ to the model

3. INTEGRATING DATA: DOUGLAS-FIR

what species can we choose?

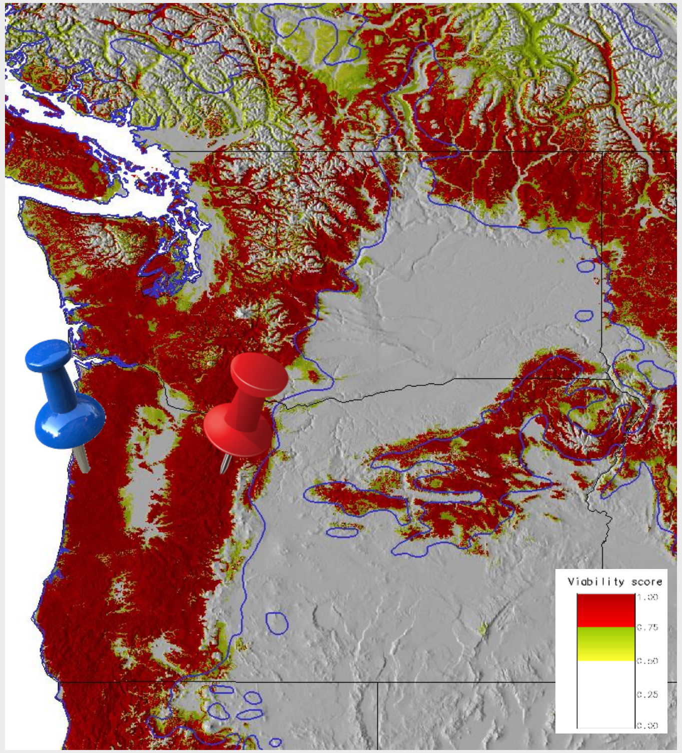




$\delta^{14}N$

- Easy to obtain
- Large geographic, climatic range, with continuous and patchy distributions
- Wealth of knowledge on D-fir

INTEGRATING DATA: PILOT STUDY

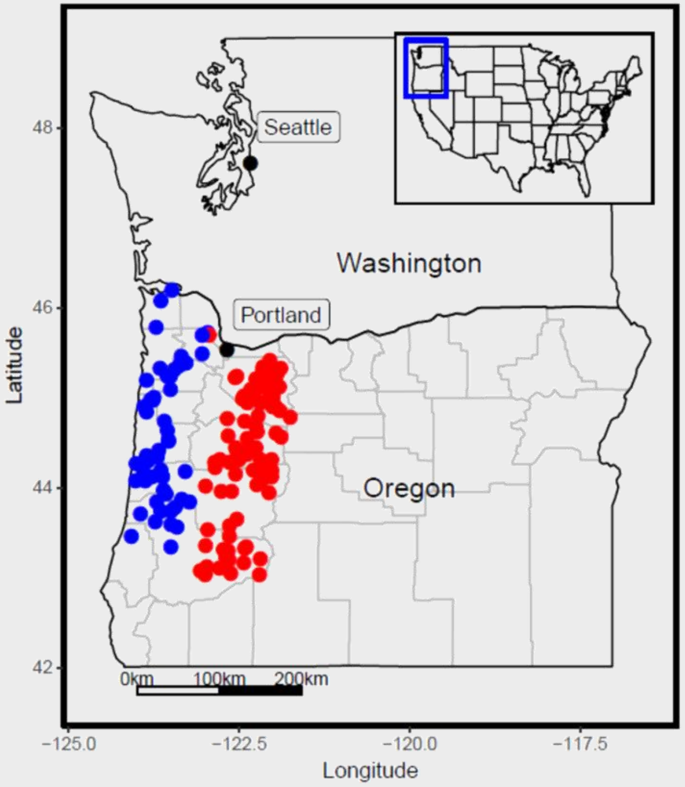


PNW REGION D-FIR

Q: can we identify tree source as coast v. cascade?

- Genetics
- Metabolomics
- Anatomy
- Isotopes

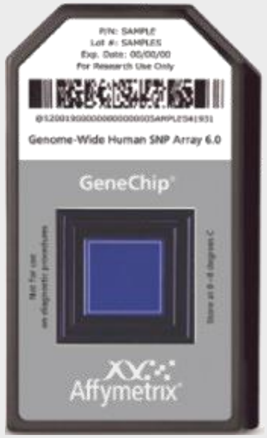
INTEGRATING DATA: GENETICS



PNW GENETICS STUDY

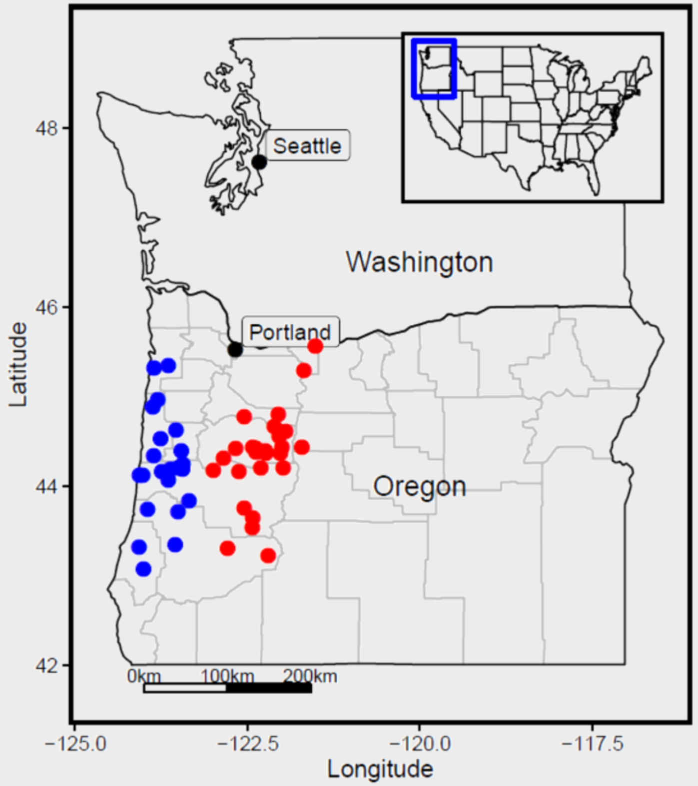
384 individuals; 141 in Oregon

51 coast 90 cascade



- Needle DNA assayed for nuclear genetic variation at 25,000 genes
- 16,467 usable Single Nucleotide Polymorphisms (SNPs)
- SNPs ranked by spatial signal; 500 'top Fst' SNPs selected
- Random Forest classification performed using 500 SNPs

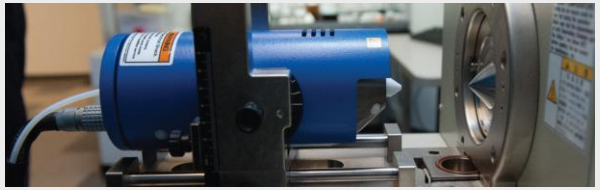
INTEGRATING DATA: METABOLOMICS



METABOLOMIC STUDY

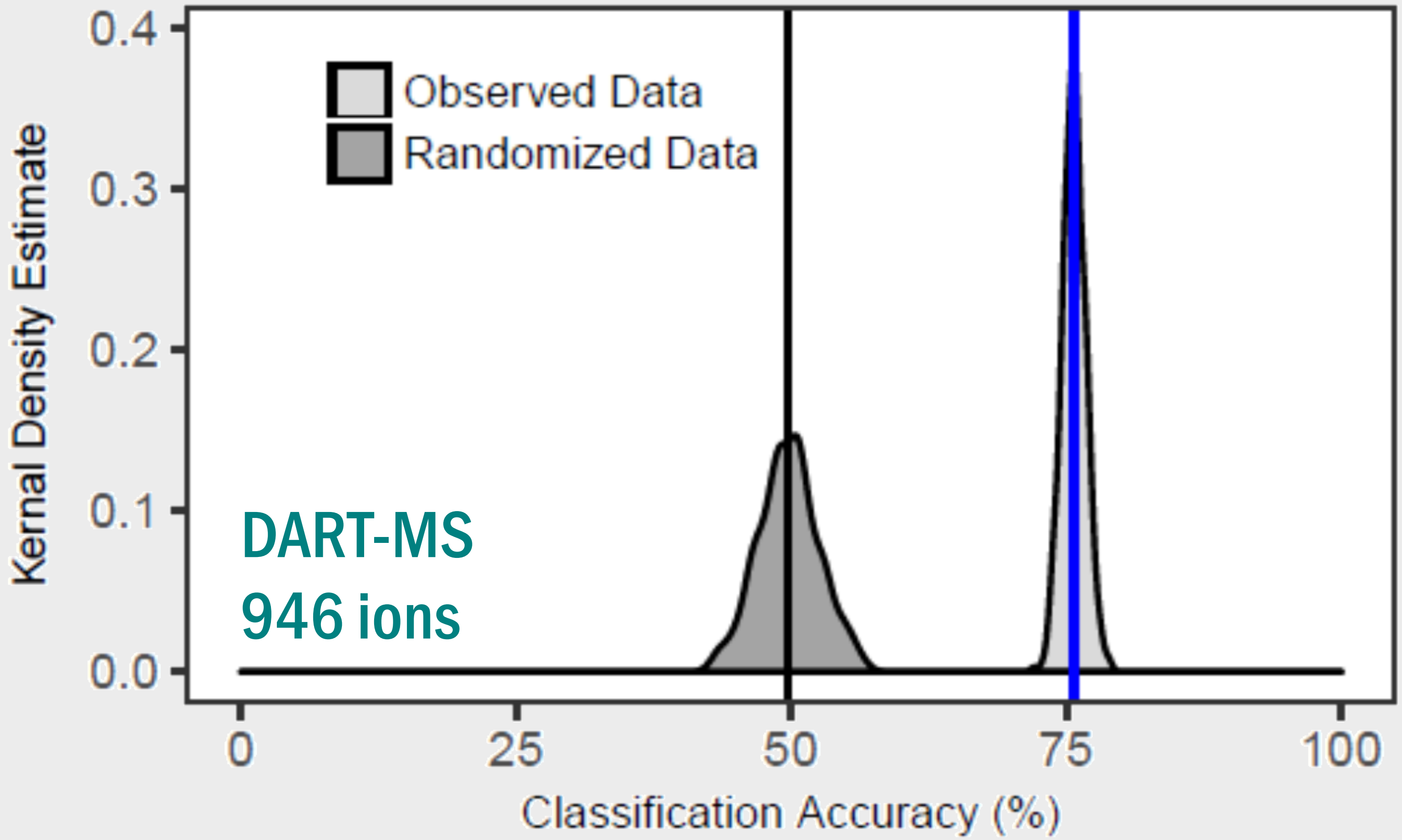
188 individuals; Oregon

86 coast 102 cascade



- Cores extracted from trees, dried
- Heartwood (yrs 27-29) profiled by DART-MS
- Ion presence, abundance estimated by Mass Mountaineer™; 946 ions
- Mean profiles estimated (n=3)
- Random Forest classification performed using 946 ions

RESULTS: RF CLASSIFICATION



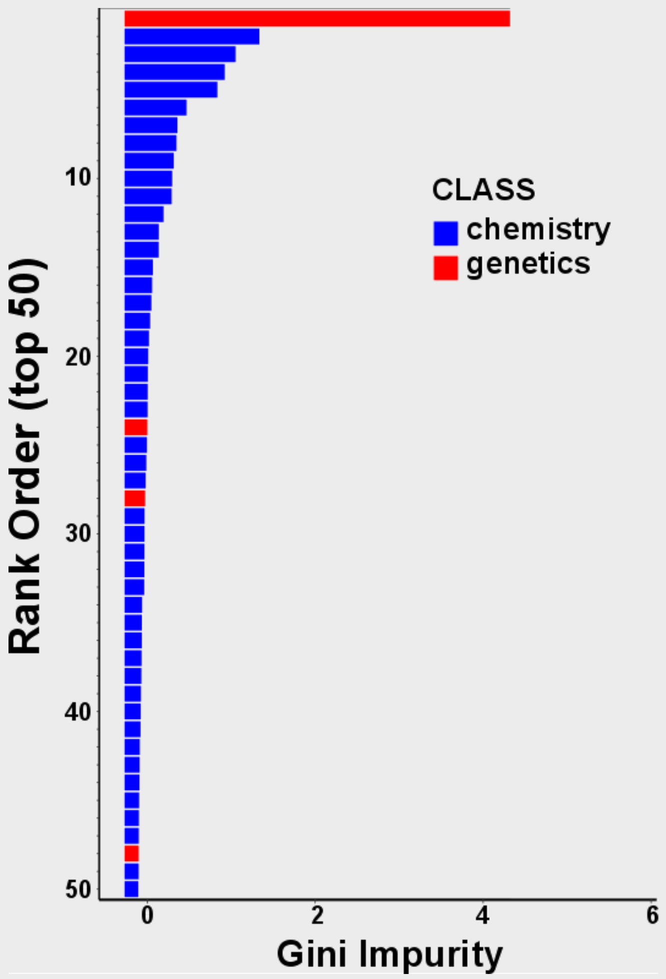
- Sanity check: randomized data accurately classified 50% of time…
- Observed classifications estimated from 500 replicates
  - *Example: DART-MS accurate 75.7% of time*

RF CLASSIFICATION ACCURACY

MODEL	INPUTS	ACCURACY
GENETICS MODEL	500 SNPs	83.4%
METABOLITE MODEL	Metabolites: 946 ions	75.7%
FULL MODEL	Genet+Metab: 500 SNPs + 946 ions	83.6 %

GENETIC & METABOLOMIC 'IMPORTANCE'

SNPs+METAB

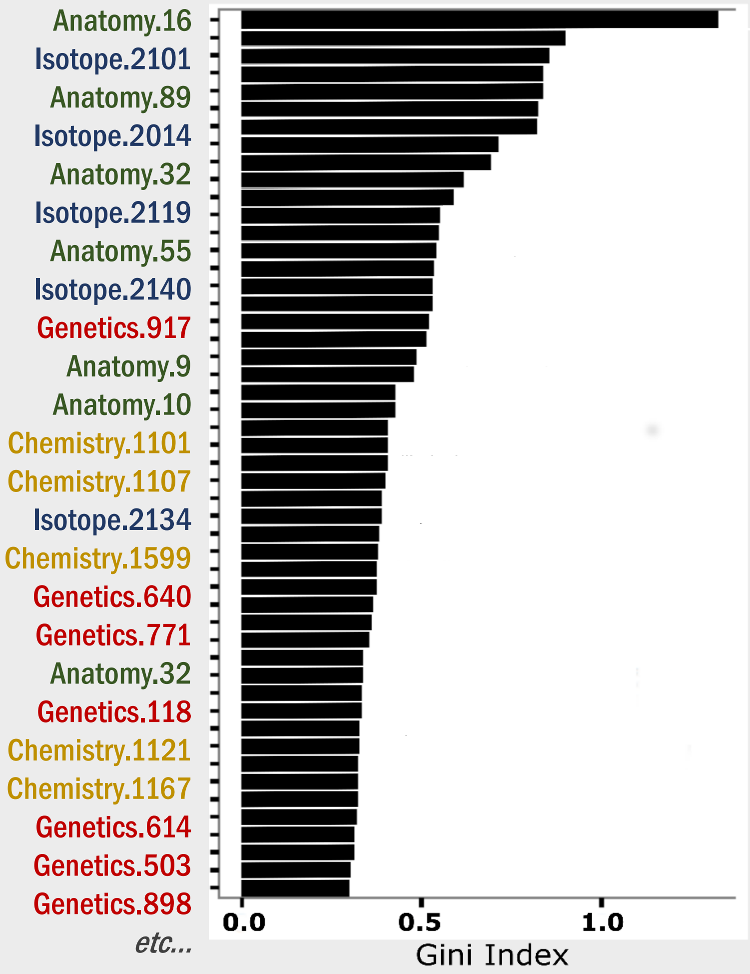


What can we learn from integrated analysis?

- Integration DOESN'T measurably improve classification accuracy (in this case)
- Integration DOES reveal contribution of genetics, metabolomics to the classification model
- Integration allows us to examine classifier 'importance' – what drives the classification?

GENETIC & METABOLOMIC 'IMPORTANCE'

SNPs+METAB+ANAT+ISO

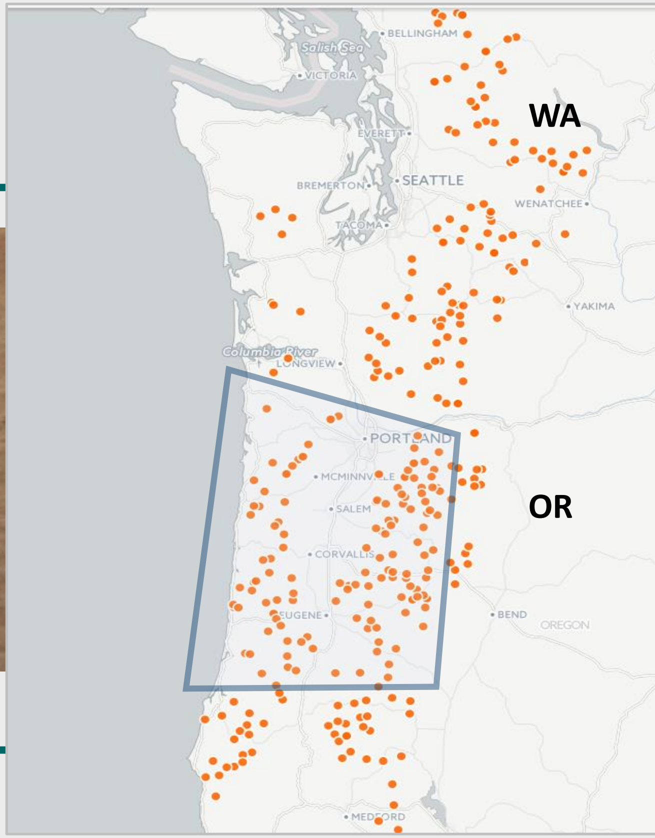


What can we learn from integrated analysis?

- Integration doesn't measurably improve classification accuracy (in this case)
- Integration reveals contribution of genetics, metabolomics to the classification model
- Integration allows us to examine classifier 'importance' – what drives the classification?
- *Imagine if you had a rich data set ....*

GENETICS + METABOLOMICS +

MODEL	INPUTS	VARIABLES
GENETIC MODEL		
METABOLITE MODEL		
ANATOMY MODEL		
ISOTOPE MODEL		
FULL MODEL		



PNW GENETIC STUDY

340 families (locations)

4. CONCLUDING REMARKS

- Integrated classification models from multiple data sources possible with Random Forests (and other algorithms)
- Gain insights into:
    - Factors responsible for classification
    - Methodological, variable importance
- Develops robust classification models

Contents lists available at ScienceDirect

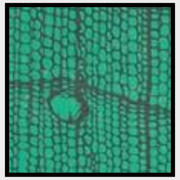
Biological Conservation

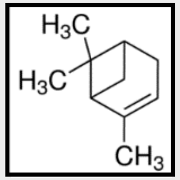
journal homepage: www.elsevier.com/locate/bioc

Discussion

Forensic timber identification: It's time to integrate disciplines to combat illegal logging

Eleanor E. Dormontt[a], Markus Boner[b], Birgit Braun[c], Gerhard Breulmann[d], Bernd Degen[e], Edgard Espinoza[f], Shelley Gardner[g], Phil Guillery[h], John C. Hermanson[i], Gerald Koch[j], Soon Leong Lee[k], Milton Kanashiro[l], Anto Rimbawanto[m], Darren Thomas[n], Alex C. Wiedenhoeft[o], Yafang Yin[p], Johannes Zahnen[q], Andrew J. Lowe[a,*]



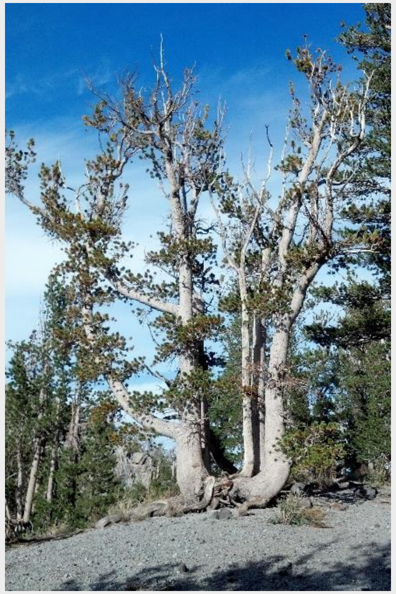


$\delta^{14}N$

CONCLUDING REMARKS

- "Field of Dreams" hypothesis: *Build it…*
- Temperate zone trees can help simulate…
  - Spatial classification
  - Taxonomic classification (e.g., White Oaks, Pines)
  - Spatial + Taxonomic classification

https://en.wikipedia.org/wiki/Oak#/media/File:Quercus_robur.jpg



Core collectors

Tara Jenings, Zolton Bair, Keaton Boeder, Whitney Meier (Oregon State Univ)

Shelley Stephan, Patric Krabacher (USFS-PNW)

Allan Braun, Devin Ashcraft, Nancy Shadomy (USFS-R6)

Support

US Forest Service Pacific Northwest Research Station

Oregon State University Department of Botany and Plant Pathology

US Fish and Wildlife Service Forensics Laboratory

US Forest Service International Programs

THANKS!

‘WE’ - DEFINED

Kristen Finch
OSU PhD candidate

Ed Espinoza
USFWS Forensics Lab

Andy Jones
OSU Botany and Plant Path

Rich Cronn
USFS PNW