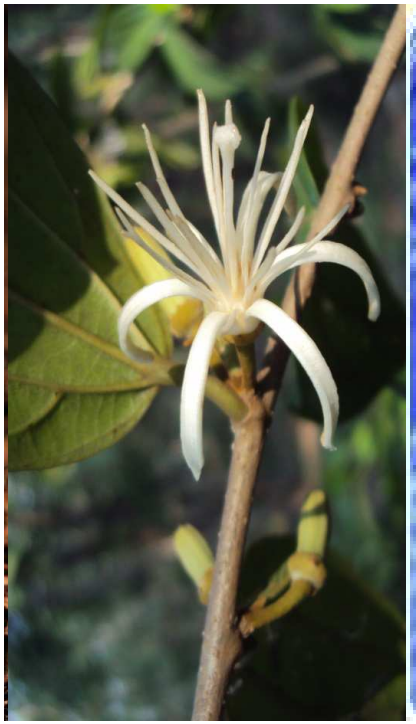
A Targeted Enrichment Strategy for Sequencing of Medicinal Species in the Indonesian Flora

Berenice Villegas-Ramirez,
Erasmus Mundus Master Programme in Evolutionary Biology (MEME)

Supervisors:
Dr. Sarah Mathews, Harvard University
Dr. Hugo de Boer, Uppsala University



Introduction

- Up to 70,000 plant species are used worldwide in traditional medicine.
- At least 20,000 plant taxa have recorded medicinal uses.

- Main commercial producers are in Asia: China, India, Indonesia, and Nepal.
- Indonesia has c. 7000 plant species of documented medicinal use.

- But……

Transmigration and Farming

Herbarium Specimens

- Plastid genes *rbcL* and *matK* have been be adopted as the official DNA barcodes for all land plants.
  - *rbcl* ~ 1428 bp
  - matK ~ 1500 bp

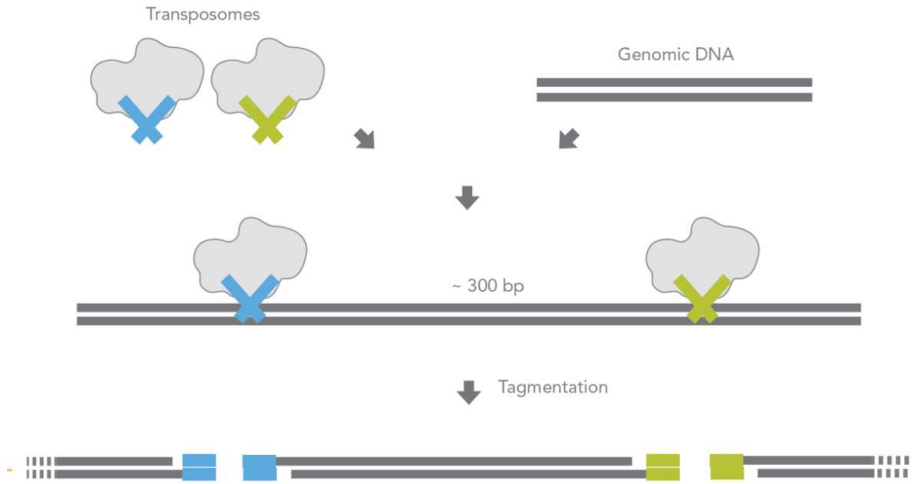
- Herbarium specimens often require more attempts at amplification with more primer combinations.
  - Higher possibility of obtaining incorrect sequences through increased chances of samples becoming mixed up or contaminated.

- Lower performance using herbarium material due to lower amplification success.
  - Caused by severe degradation of DNA into low molecular weight fragments.

- But fragmented DNA is not a curse!

Next-Generation Sequencing

- Fragmented DNA is less of a problem



- Only a few milligrams of material are necessary

Targeted Enrichment

- Defined regions in a genome are selectively captured from a DNA sample prior to sequencing.

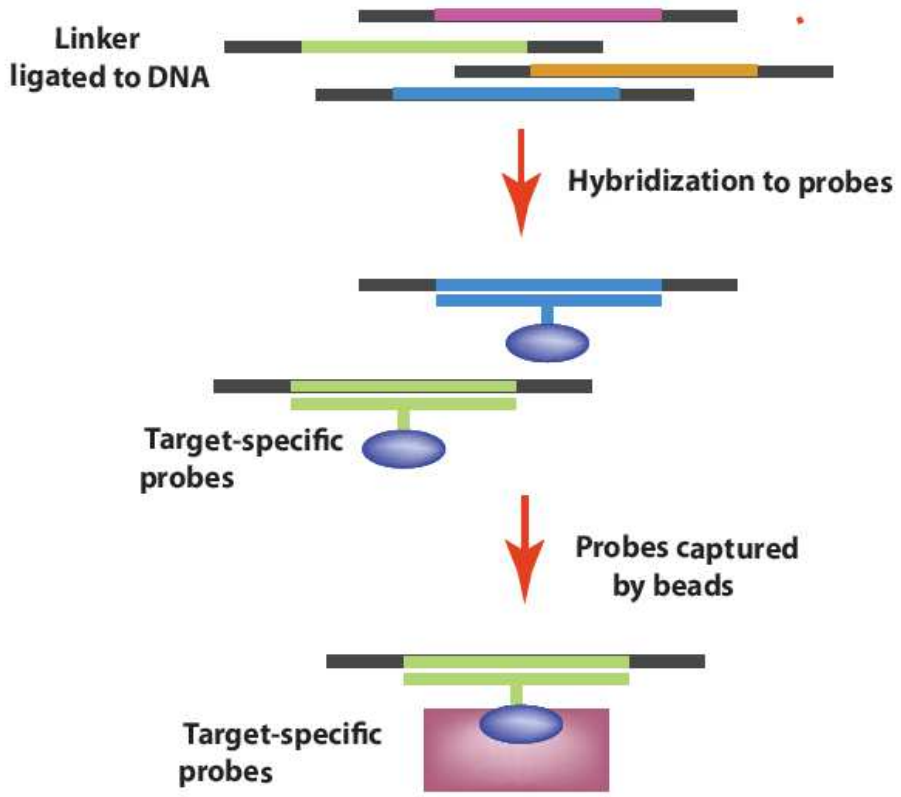
- The genomic complexity in a sample is reduced.

- More time- and cost-effective.

Hybrid Capture Targeted Enrichment

- Library DNA is hybridized to a probe.
  - Pre-prepared DNA or RNA fragments complementary to the targeted regions of interest.
- Non-specific hybrids are removed by washing.
- Targeted DNA is eluted.

Easy to use, utilizes a small amount of input DNA (<1-3 ug), and number of loci (target size) is large (1-50 Mb).



Purpose of Research

- Test the feasibility and utility of a hybridization-based target enrichment approach to sample herbarium specimens.

- Obtain genome-scale data appropriate for phylogenetic analyses of important medicinal plants in the Indonesia flora.

- Explore the utility of the probes for Hyb-seq to target conserved regions and therefore work across a group of interest (e.g. genus or family).

Alangium

- Member of Cornales family and sister to dogwood genus Cornus.
- Approximately 24 species, of which only 1 is not a tree or shrub.
- *A. chinense*, *A. salviifolium*, and *A. platanifolium* have a long history of medicinal use.
    - *A. chinense* is one of the 50 fundamental herbs in Chinese medicine.
- Well-resolved phylogeny with no major hybridization or species delimitation problems.
- But with some room for improvement.

Step 1: Probe design

- Reference transcriptome sequences downloaded from the 1 KP Project database.
  Alangium chinense
  Cornus florida
  Camptotheca acuminate
  Nyssa ogeche
  Dichroa febrifuga

- Identified putatively orthologous loci within each respective transcriptome.

  •Targets filtered to retain only those with a high enough similarity for the hybridization to work (12.5%) but a certain degree of mismatch.

  - **255 nuclear loci** ranging from 300 bp to 2, 229 bp (208, 785 bp total)

- Baits were then constructed by Mycroarray as an 80-mer and tiled across each locus with a 40-bp overlap between baits (2x tiling).

Step 2: Data Collection

- 75 specimens representing 23 species of Alangium were sampled.
  - 15 were borrowed from other herbaria in the U.S. and China.
  - Oldest specimen collected was Alangium begoniifolium dating from 1917.
- An additional 20 specimens belonging to the Dichroa and Camptotheca genera were sampled.

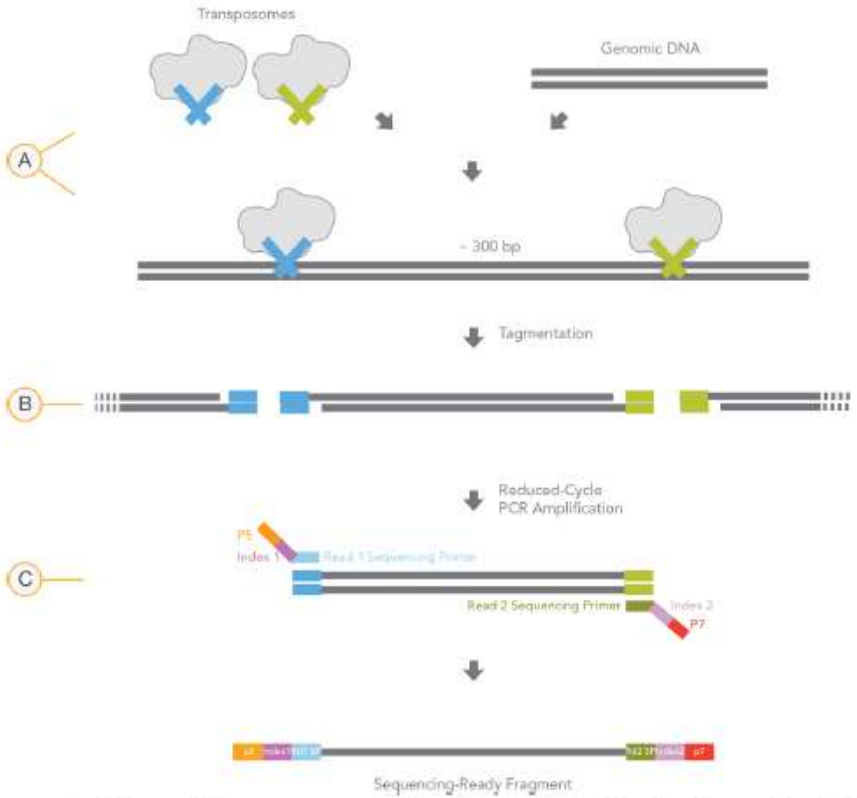
Step 3: DNA Extractions

- DNeasy Plant Mini Kit
- Incubation at 65°C for 24 hours.
- Extracted DNA samples were visualized on an agarose gel and quantified using the Qubit dsDNA BR assay.

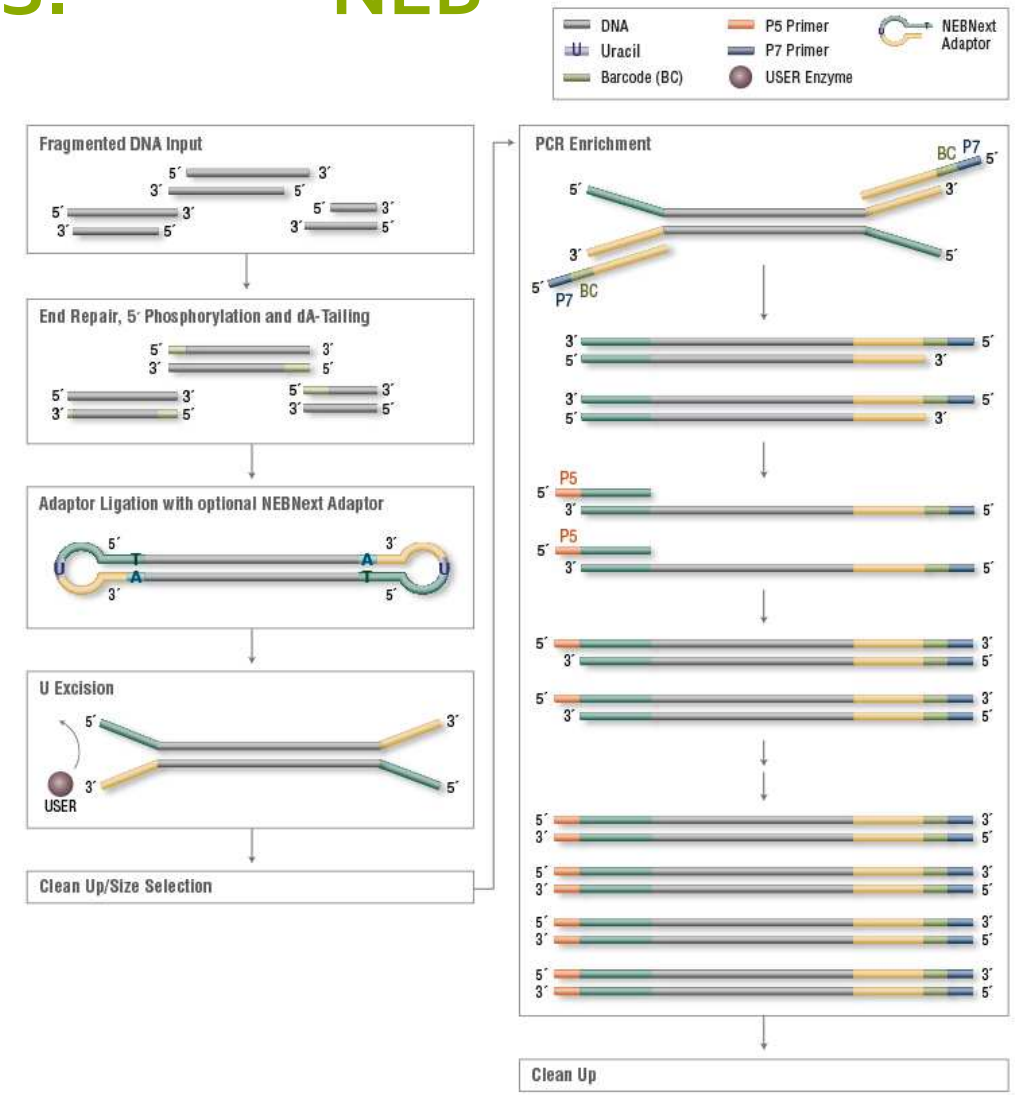
Step 4: DNA Libraries

Nextera VS. NEB

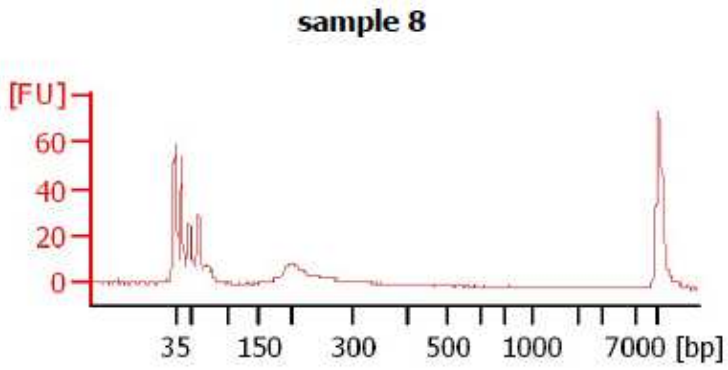
Nextera VS. NEB

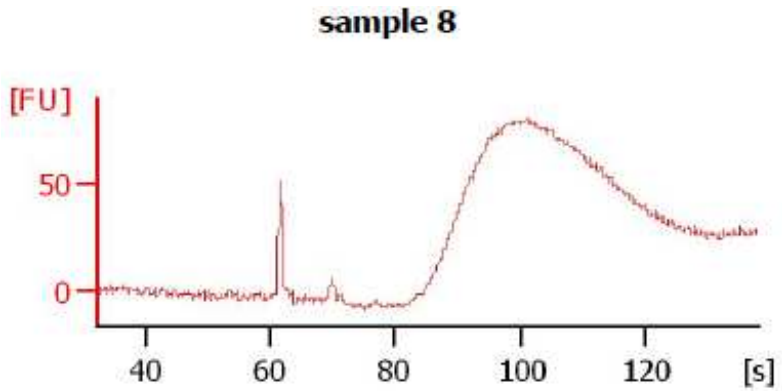


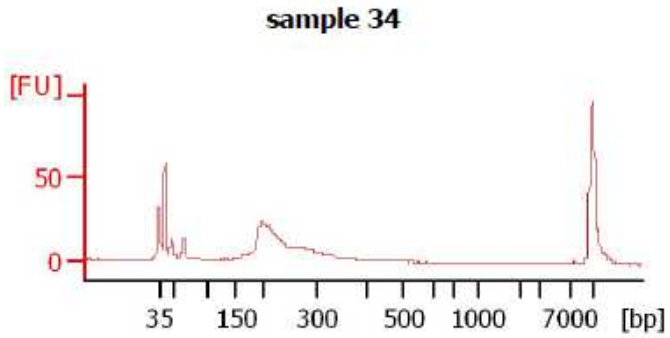
A Nextera XT transposome with adapters is combined with template DNA
B Tagmentation to fragment and add adapters
C Limited cycle PCR to add sequencing primer sequences and indicies

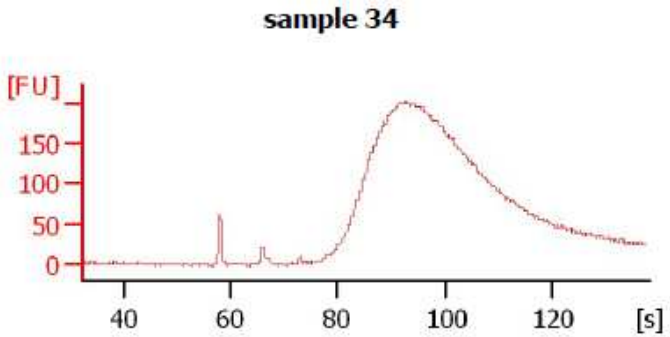


Nextera VS. NEB







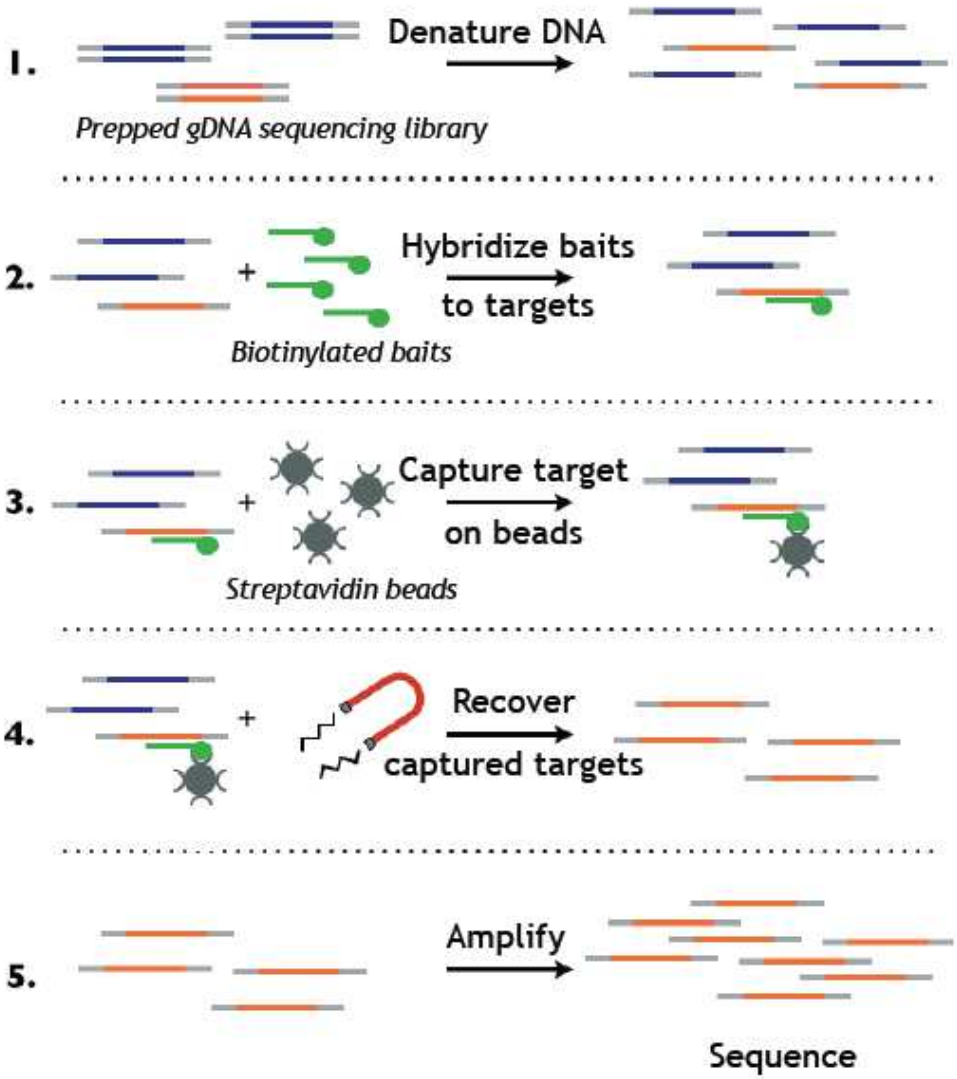


Step 4: DNA Libraries

- 41 libraries were made.
  - Input DNA concentration ranged from 1.19 to 115 ng/uL.
  - Output library concentration ranged from 0 to 41.5 ng/uL.

- 30 libraries divided into 3 sets for 3 hybridization reactions (H2, H3, and H4).
  - Libraries pooled together in 9- 10- or 11- plexes with equimolar ratios and concentrated using ethanol precipitation.
  - 450 ng total input DNA/hybridization reaction.

Step 5: Hybridization



65°C for 36 h

Enriched products are purified

PCR amplification of 30 cycles

Step 6: Sequencing

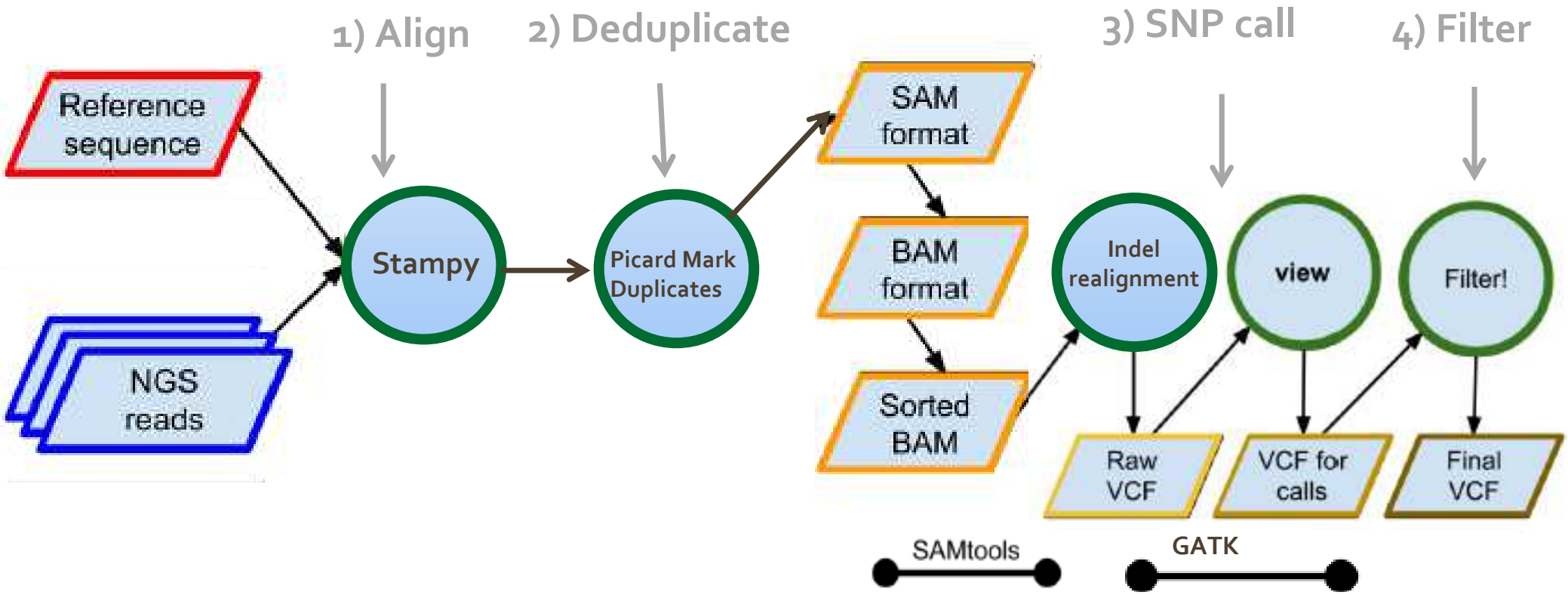
Illumina HiSeq2500

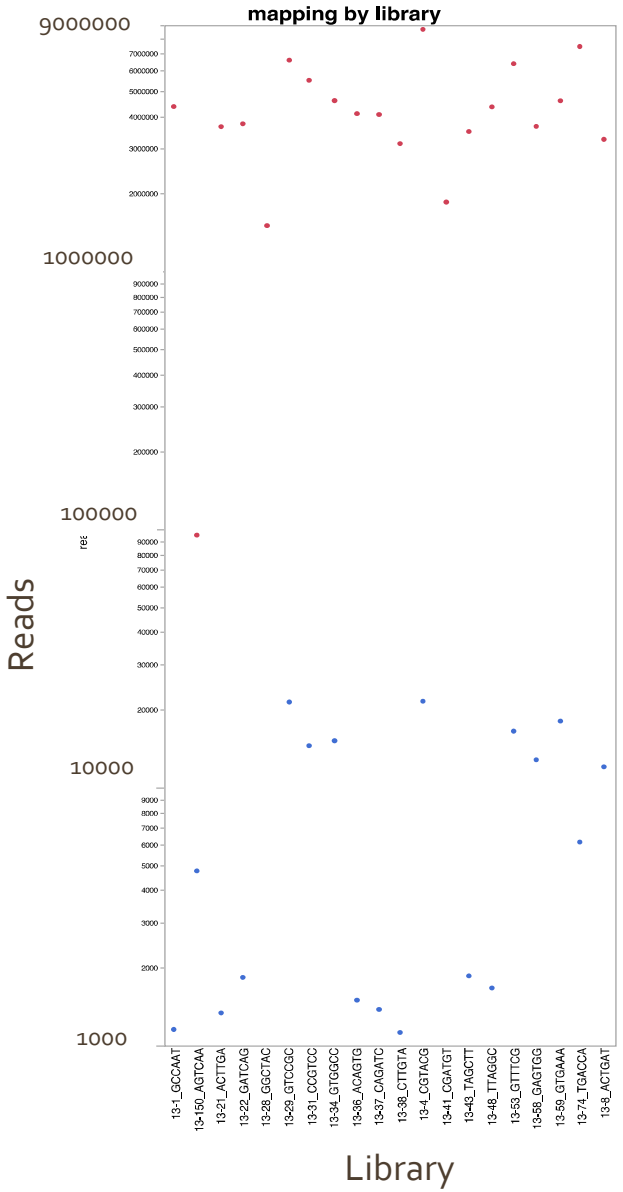
- H2 and H4 pooled together for sequencing.
  - 20 libraries with unique barcodes.

- Two sequencing runs of the same 20 libraries each time:
  - First 2x150 runs were not deep enough.
  - Added 1x50 as a follow-up to deepen sequence pool.

2X150 produced significantly better results, especially in respect to % of reads mapped!

Step 7: Data Processing

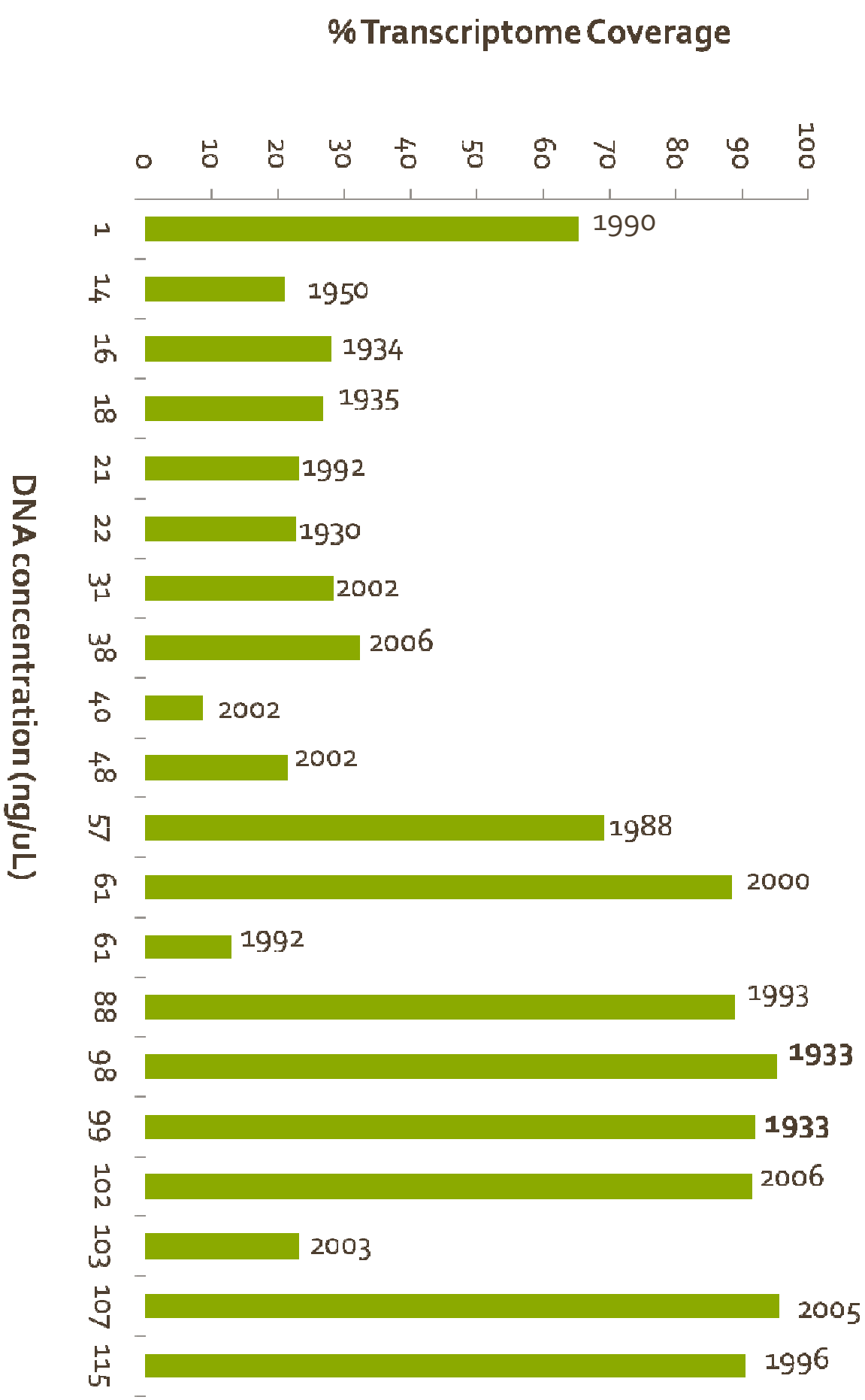




library	mapped	unmapped	pct
13-1_GCCAAT	1153	4356311	0.03
13-21_ACTTGA	1338	3642457	0.04
13-22_GATCAG	1837	3738446	0.05
13-28_GGCTAC	400	1505321	0.03
13-29_GTCCGC	21440	6591960	0.33
13-31_CCGTCC	14526	5513195	0.26
13-34_GTGGCC	15180	4595054	0.33
13-36_ACAGTG	1499	4089777	0.04
13-37_CAGATC	1379	4058104	0.03
13-38_CTTGTA	1123	3131085	0.04
13-41_CGATGT	564	1857364	0.03
13-43_TAGCTT	1860	3487108	0.05
13-48_TTAGGC	1671	4346012	0.04
13-4_CGTACG	21581	8683774	0.25
13-53_GTTTCG	16527	6385387	0.26
13-58_GAGTGG	12801	3653489	0.35
13-59_GTGAAA	18085	4588074	0.39
13-74_TGACCA	6140	7447651	0.08
13-8_ACTGAT	12019	3251624	0.37
13-150_AGTCAA	4746	95027	4.99

Preliminary Results

Species	DNA concentration (ng/µL)	Mapped reads	Unmapped reads	% reads on target	% Transcriptome coverage
A. hirsutum	**107**	**21440**	6591960	**0.33**	**95.60**
A. begoniaefolium	**98.1**	**21581**	8683774	**0.25**	**95.12**
A. salviifolium	**98.6**	**16527**	6385387	**0.26**	**91.84**
A. kurzii	**102**	**15180**	4595054	**0.33**	**91.54**
A. javanicum	**115**	**14526**	5513195	**0.26**	**90.42**
A. villosum subps. tomentosum	**87.6**	**12801**	3653489	**0.35**	**88.85**
A. chinense	**60.9**	**12019**	3251624	**0.37**	**88.38**
A. sinicum	56.8	6140	7447651	**0.08**	69.09
A. grisolleoides	1.19	4746	95027	**4.99**	65.22
A. plantanifolia	37.5	1860	3487108	**0.05**	32.34
A. griffithi	31.3	1837	3738446	**0.05**	28.41
A. rotandifolium	15.7	1671	4346012	**0.04**	28.16
A. kwangiense	17.7	1499	4089777	**0.04**	26.70
A. alpinum	21.4	1153	4356311	**0.03**	23.11
A. villosum subsp. polyosmoides	103	18085	4588074	**0.39**	23.11
A. faberi	22.2	1338	3642457	**0.04**	22.60
A. longiflorum	47.8	1123	3131085	**0.04**	21.34
A. lamarckii	13.9	1379	4058104	**0.03**	20.94
A. nobile	61.3	564	1857364	**0.03**	12.95
A. havilandii	39.7	400	1505321	**0.03**	8.65


% Transcriptome Coverage
0
10
20
30
40
50
60
70
80
90
100
DNA concentration (ng/uL)
1
14
16
18
21
22
31
38
40
48
57
61
61
88
98
99
102
103
107
115
1990
1950
1934
1935
1992
1930
2002
2006
2002
2002
1988
2000
1992
1993
1933
1933
2006
2003
2005
1996

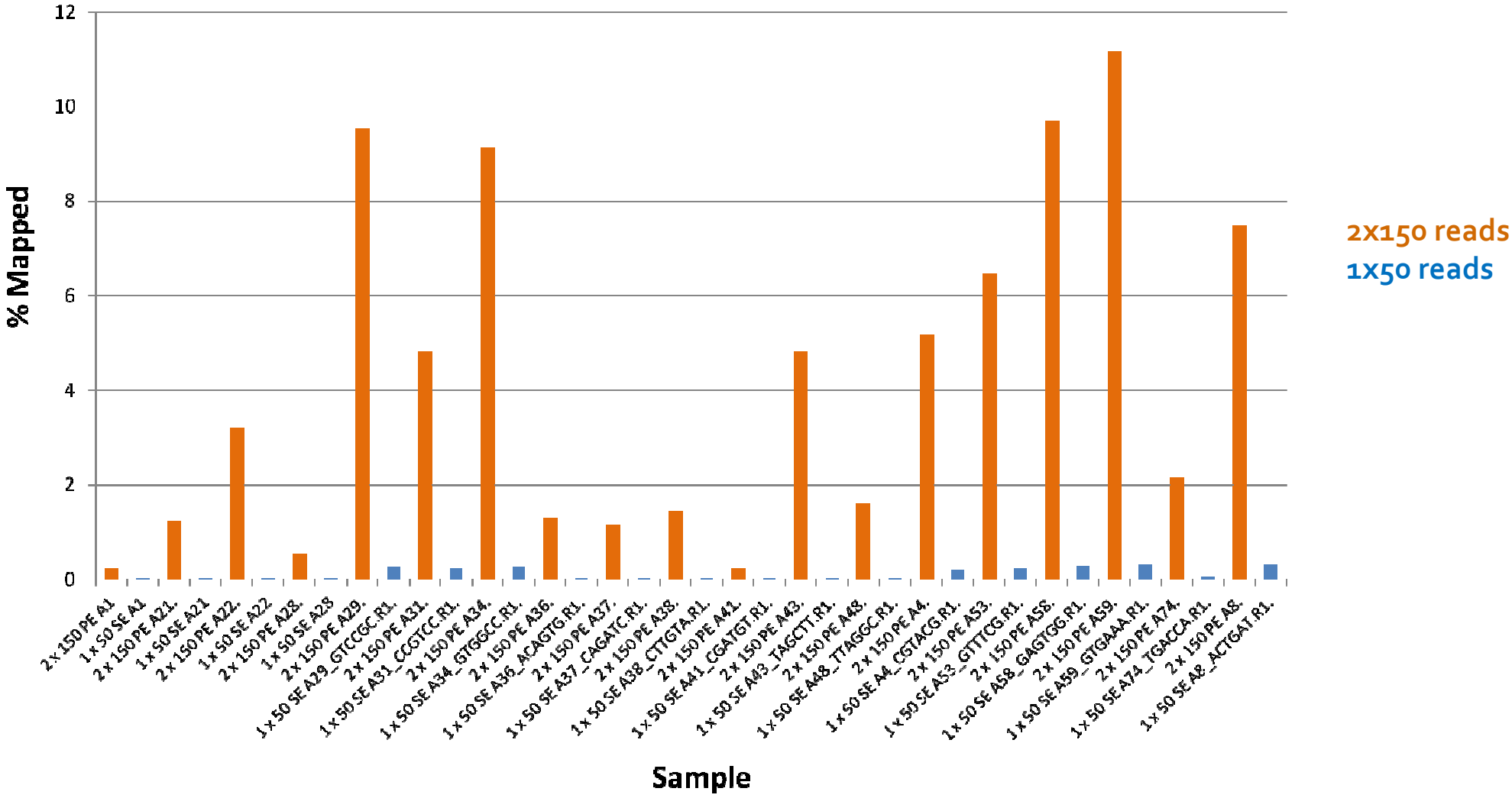
Preliminary Results

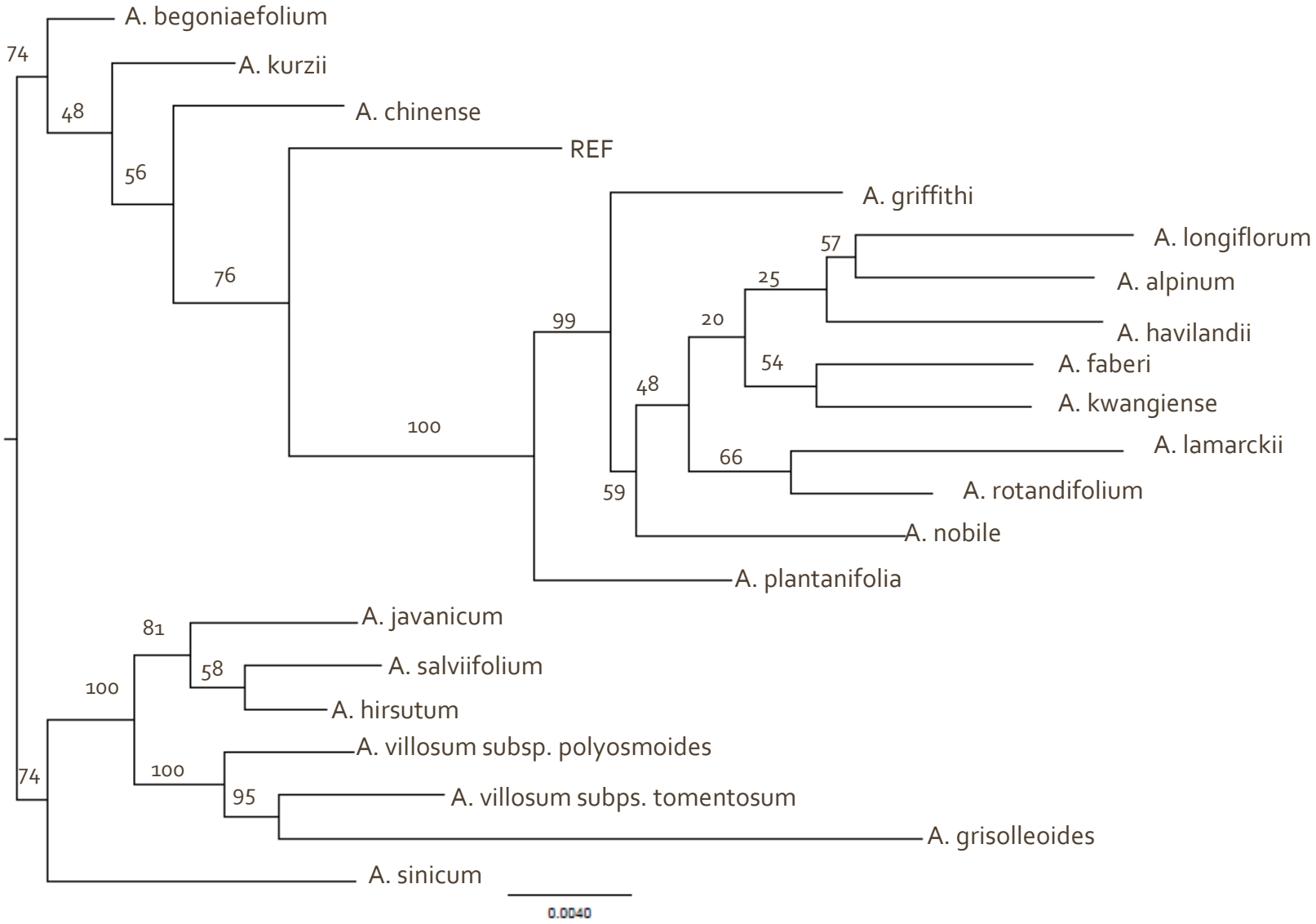
raw reads/run

2 X 150 PE run: 0.67 million

1 x 50 SE run: 102 million

it is *all* about read length




A. begoniaefolium
A. kurzii
A. chinense
REF
A. griffithi
A. longiflorum
A. alpinum
A. havilandii
A. faberi
A. kwangiense
A. lamarckii
A. rotandifolium
A. nobile
A. plantanifolia
A. javanicum
A. salviifolium
A. hirsutum
A. villosum subsp. polyosmoides
A. villosum subps. tomentosum
A. grisolleoides
A. sinicum
74
48
56
76
99
20
57
25
54
48
100
66
59
81
58
100
74
100
95
0.0040

Conclusions

- Hyb-seq approach recovered sequences that are hundreds of base pairs in length from hundreds of loci.
  - Even with modest levels of variation data might be appropriate for addressing questions at low taxonomic level.

- A large amount of data from the low-copy nuclear genome in a herbarium specimen can be generated.
  - Data could be used for phylogenomic analyses in plants.

- Availability of previously inaccessible genetic information from old type specimens, traded and processed material.
  - Crucial for resolving taxonomic uncertainties and for providing DNA barcodes.
    - Reduction in the need to rely on short molecular markers.

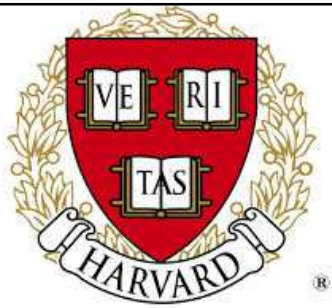
- Material otherwise not available or costly to obtain *is in reach* for comparative genome analyses.

Future Steps

- Sequencing of H3.
- Comparing alignment pipelines for short-read data of enriched regions with others.
  - Also for pipelines to pull the plastid genes, etc.
- Conduct final phylogenetic analyses.
- Enriching fresh specimens for comparison.

UPPSALA
UNIVERSITET

Acknowledgements



- Sarah Matthews
- Hugo de Boer
- Levi Yant
- Cam Webb
- The Harvard University Herbaria

THANK YOU!