PLANNING YOUR HYBSEQ PROJECT: WHAT YOU NEED AND WHAT IT COSTS

Elliot Gardner
PhD Candidate
Plant Biology and Conservation
Northwestern University and Chicago Botanic Garden

OBJECTIVES

1. Recognize potential sources for HybSeq data.

2. Identify the starter data, equipment, and consumables required for HybSeq.

3. Describe the relative advantages and disadvantages, and estimate the costs, of genome skimming and transcriptome sequencing for market development.

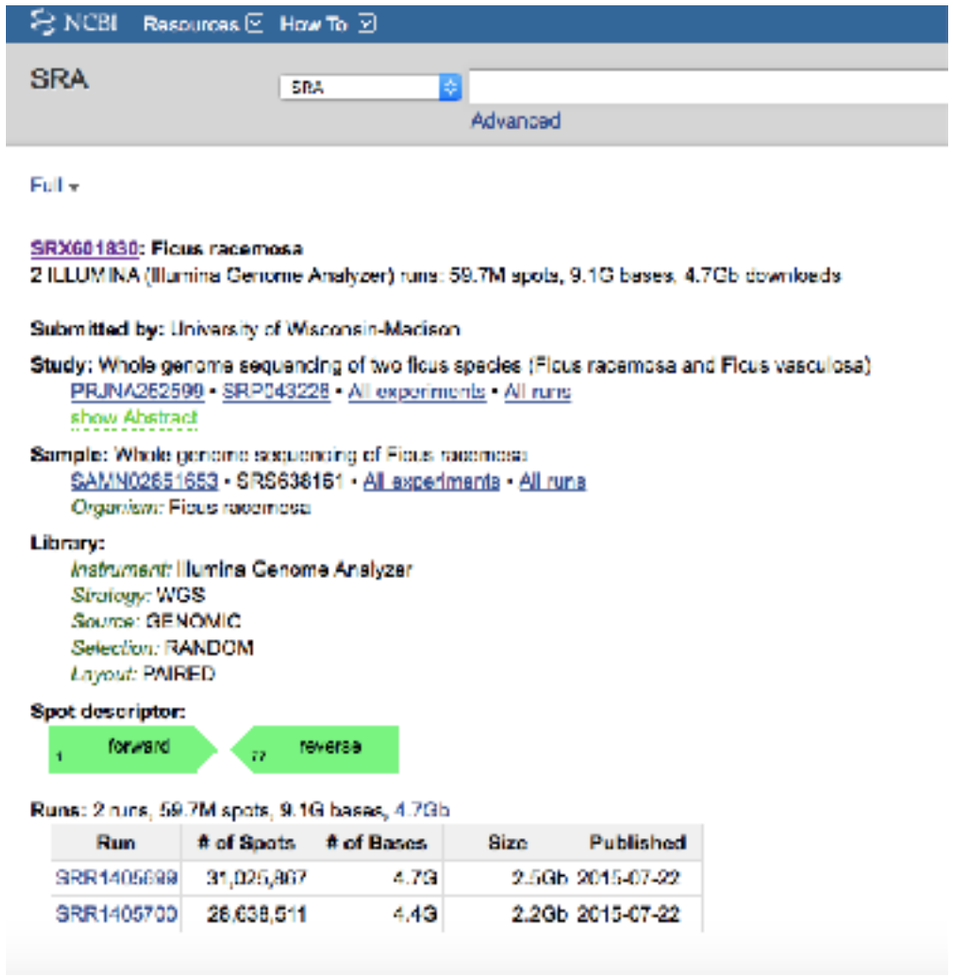
4. Design a HybSeq study that fits within a given timeframe and budget.

POTENTIAL SOURCES OF HYBSEQ DATA

- Publicly-available NGS reads from genomes or transcriptomes (e.g. NCBI SRA) (free!)
- Preserved material (e.g. silica-dried)
- Herbarium material
- Environmental and archaeological material

POTENTIAL SOURCES OF HYBSEQ DATA

- Publicly-available NGS reads from genomes or transcriptomes (e.g. NCBI SRA) (free!)



Ficus racemosa L. (Moraceae)

POTENTIAL SOURCES OF HYBSEQ DATA

- Preserved material (e.g. silica-dried)
  - Most flexible, longest potential “splash zone”

HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment

Author(s): Matthew G. Johnson, Elliot M. Gardner, Yang Liu, Rafael Medina, Bernard Goffinet, A. Jonathan Shaw, Nyree J. C. Zerega, and Norman J. Wickett
Source: Applications in Plant Sciences, 4(7)
Published By: Botanical Society of America
DOI: http://dx.doi.org/10.3732/apps.1600016
URL: http://www.bioone.org/doi/full/10.3732/apps.1600016

Artocarpus sericicarpus F.M. Jarrett (Moraceae)

POTENTIAL SOURCES OF HYBSEQ DATA

- Herbarium material
  - Shorter “splash zone” but works very well even with very old (>100y) material

TAXON 65 (5) • October 2016: 1081–1092

METHODS AND TECHNIQUES

Retrieval of hundreds of nuclear loci from herbarium specimens

Michelle L. Hart,[1] Laura L. Forrest,[1] James A. Nicholls[1,2] & Catherine A. Kidner[1,3]

1 *Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, U.K.*
2 *Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, U.K.*
3 *Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh EH9 3BF, U.K.*
Author for correspondence: *Catherine Kidner, c.kidner@rbge.ac.uk*
ORCID MLH, http://orcid.org/0000-0001-9503-7786; LLF, http://orcid.org/0000-0002-0235-9506; JAN, http://orcid.org/0000-0002-9325-563X; CAK, http://orcid.org/0000-0001-6426-3000

DOI http://dx.doi.org/10.12705/655.9

Inga vera Willd. (Fabaceae)

POTENTIAL SOURCES OF HYBSEQ DATA

- Environmental and archaeological material
  - Suitable for high-copy DNA (e.g. organelles)

MOLECULAR ECOLOGY
RESOURCES
Explore this journal >

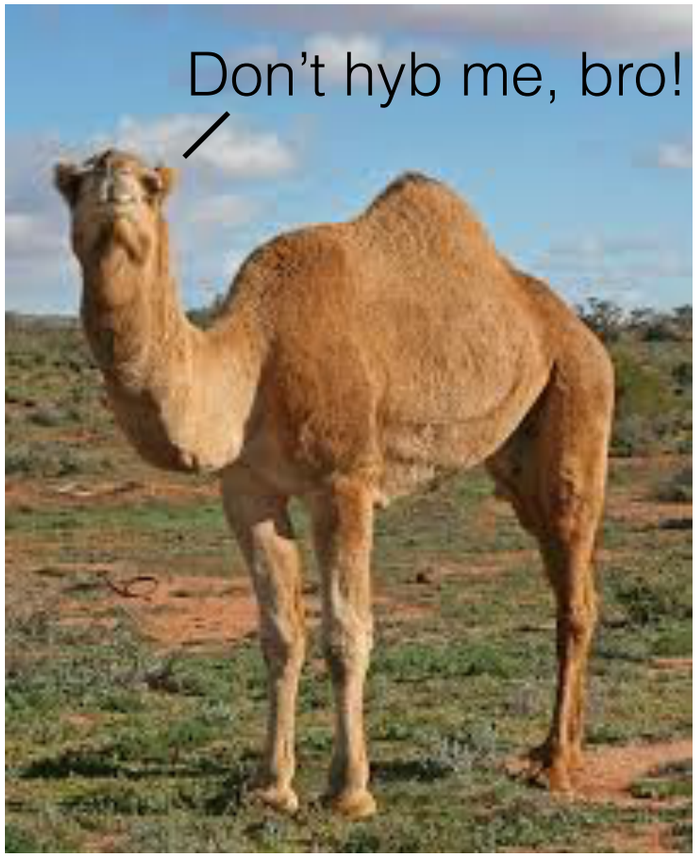
Resource Article

Combined hybridization capture and shotgun sequencing for ancient DNA analysis of extinct wild and domestic dromedary camel

Elmira Mohandesan, Camilla F. Speller, Joris Peters, Hans-Peter Uerpmann, Margarethe Uerpmann, Bea De Cupere, Michael Hofreiter, Pamela A. Burger

First published: 1 August 2016 Full publication history

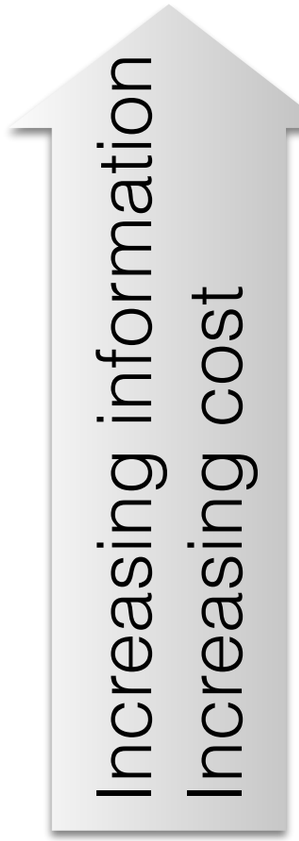
DOI: 10.1111/1755-0998.12551 View/save citation



REQUIRED STARTER DATA FOR MARKER DEVELOPMENT

<u>Sequences for target genes</u> spanning the phylogenetic scope of the project.

Sources of gene sequences



- Finished genomes — e.g. Phytozome, Genbank, 1KP
- Draft genomes
- Low-coverage genome assemblies
- Transcriptomes

REQUIRED STARTER DATA FOR MARKER DEVELOPMENT

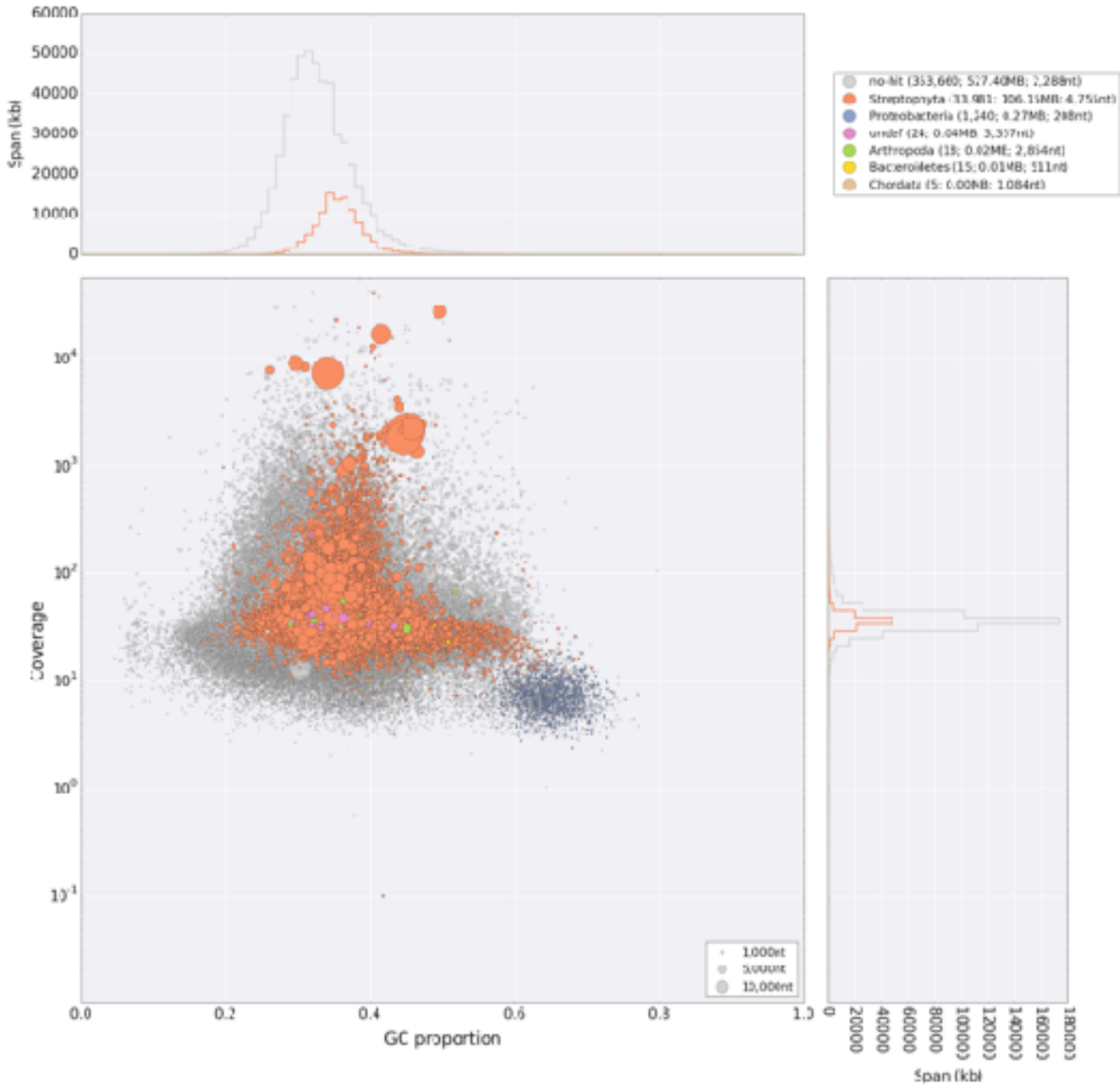
<u>Sequences for target genes</u> spanning the phylogenetic scope of the project.

Estimated costs to generate gene sequences:

- HiSeq 2000 (1 lane, ~60 Gbp): ca. £2,500
  - 1 Low-coverage genome (~15-20x for a ~700 Mbp genome)
  - 8 transcriptomes
- MiSeq (600 cycle, 1 run, ~15 Gbp): ca. £1,500
  - 1 Low-coverage genome (~15-20x for a ~175 Mbp genome)
  - 2 transcriptomes

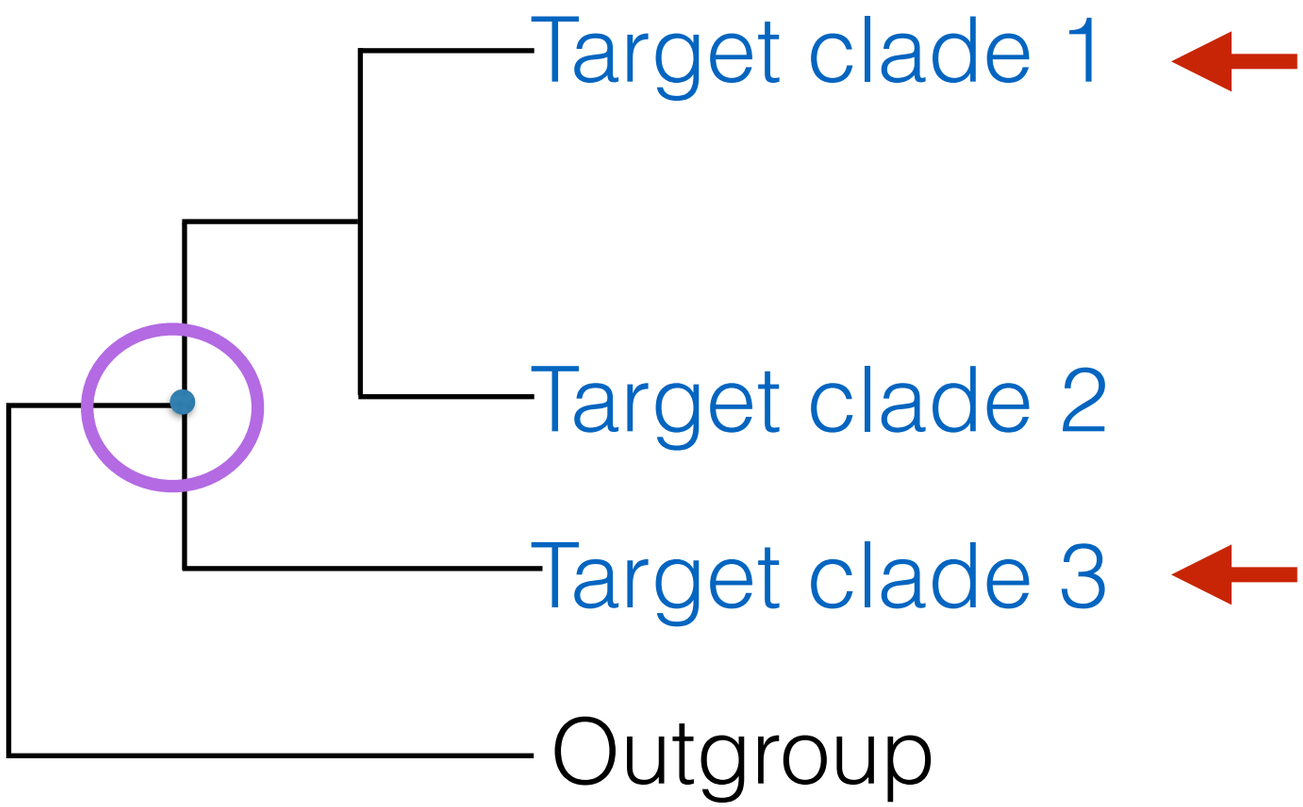
REQUIRED STARTER DATA

Remember: calculated depth differs from actual depth



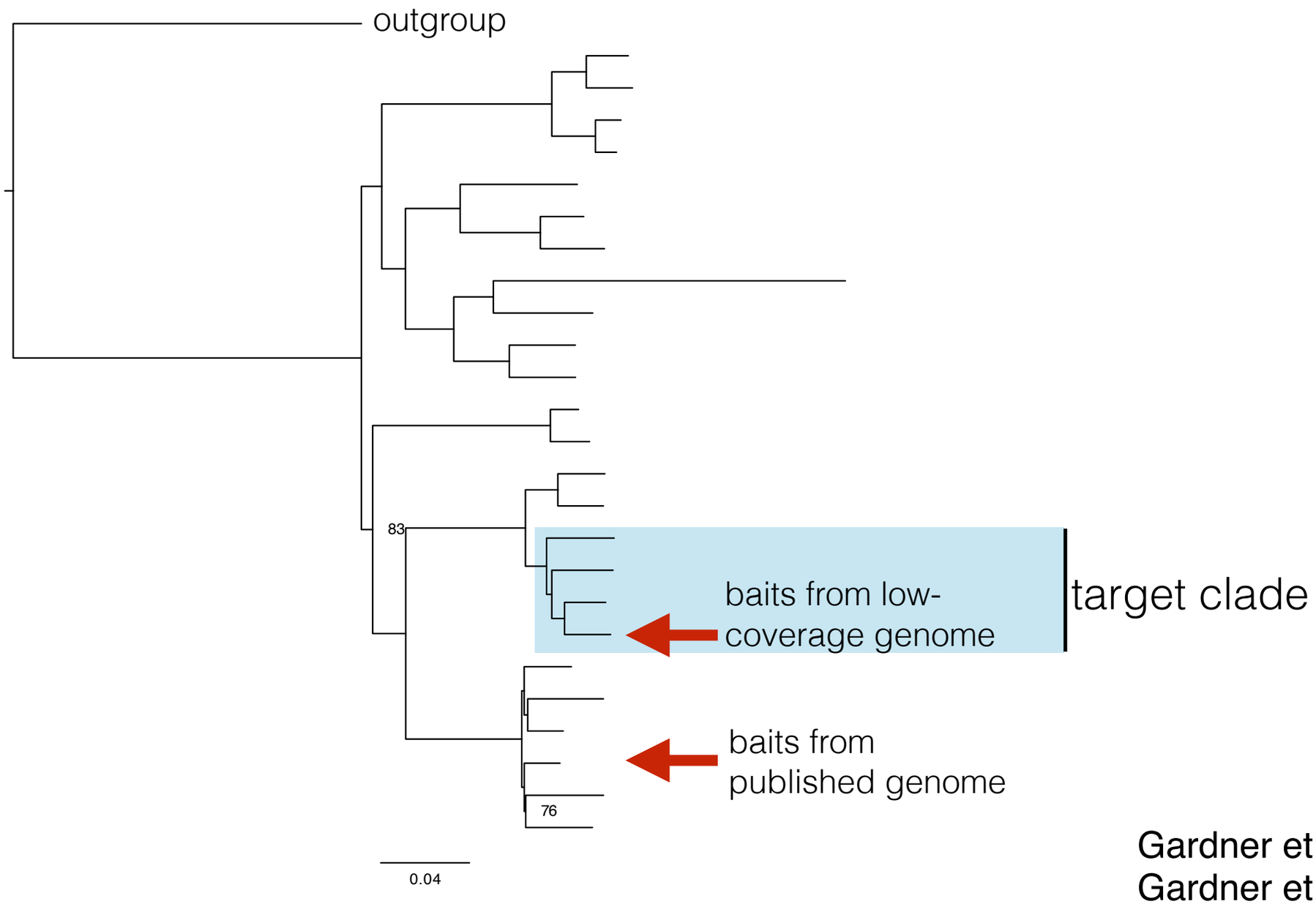
REQUIRED STARTER DATA

Sequences for target genes <u>spanning the phylogenetic scope of the project</u>.



REQUIRED STARTER DATA

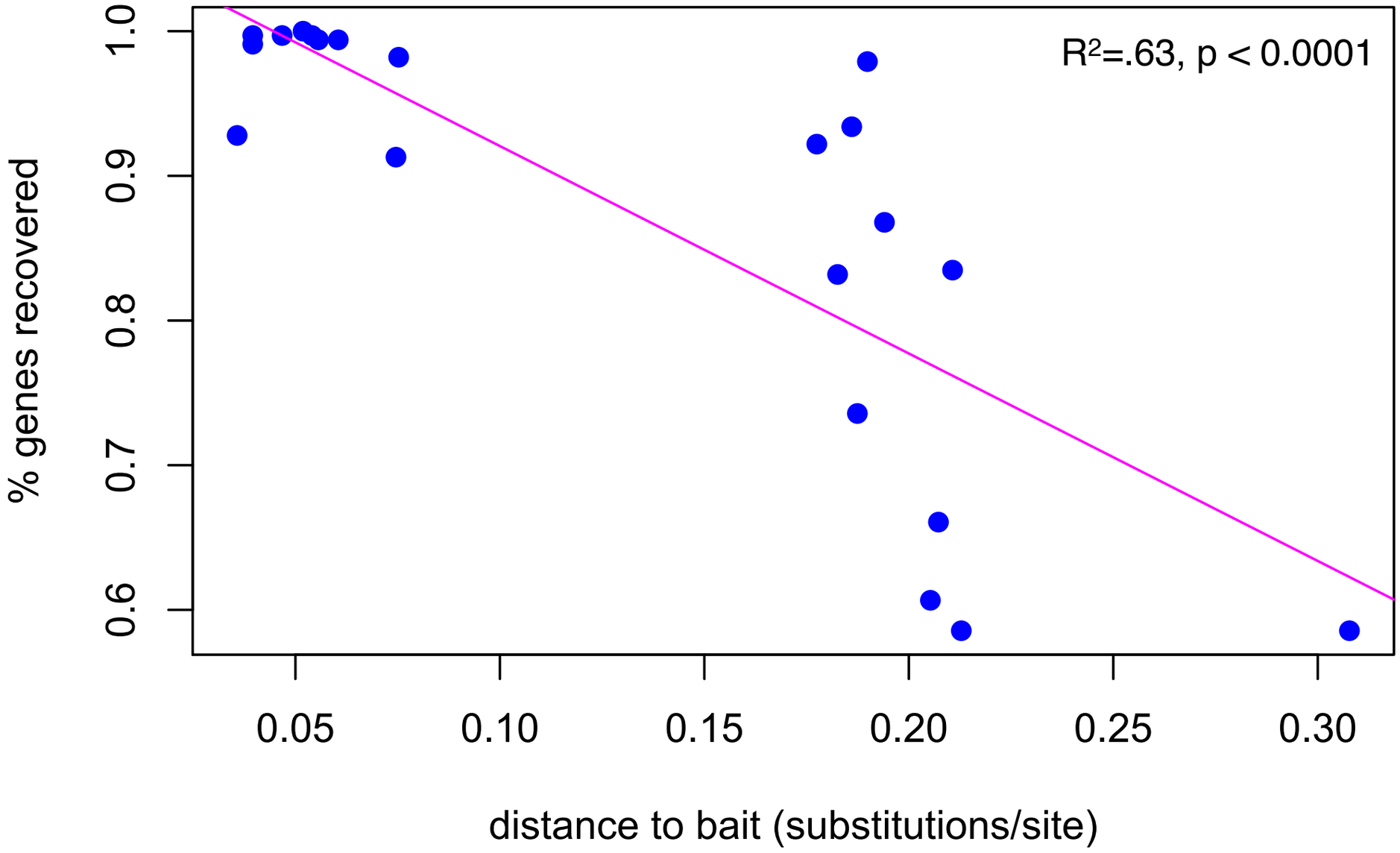
Example from Moraceae



Gardner et al. 2016
Gardner et al., *in prep.*

REQUIRED STARTER DATA

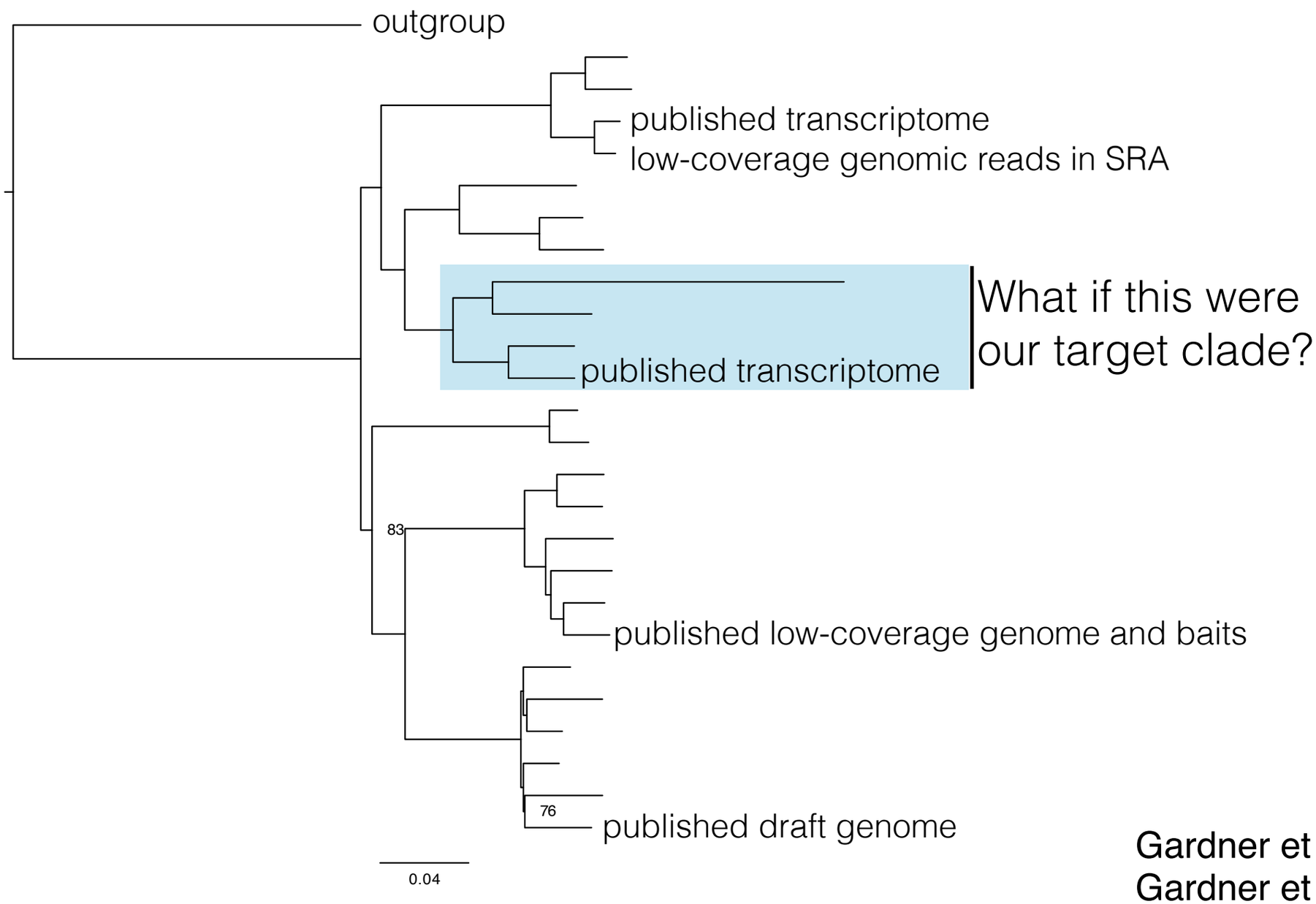
Example from Moraceae: effect of phylogenetic distance to baits on gene recovery



Gardner et al., *in prep.*

REQUIRED STARTER DATA

Example from Moraceae



Gardner et al. 2016
Gardner et al., *in prep.*

BASIC WORKFLOW

1. Sampling
2. DNA extraction and QC
3. Library preparation and QC
4. Hybridization and final QC
5. Sequencing

REQUIRED EQUIPMENT

DNA extraction:
- Tissue disruptor
- Centrifuge

DNA quality control:
- Qubit
- Vacuum centrifuge (SpeedVac)
- Gel rig
- (BioAnalyzer or TapeStation- for precious/degraded samples)
- (NanoDrop - for detecting impurities)
- (Tube magnet - for purifying samples)

Library preparation and hybridization:
- Sonicator (Covaris)
- Thermocycler
- Plate magnet (tube magnet is also useful for low-throughput hybridization)
- Vacuum centrifuge (SpeedVac)

Library quality-control
- BioAnalyzer or TapeStation

Sequencing
- MiSeq, NextSeq, or HiSeq

Other useful equipment:
- BluePippin or PippinPrep - for very accurate size selection
- QPCR machine - for very accurate library quantification

REQUIRED CONSUMABLES

DNA extraction:

- DNA kits (ca. £3 per sample)
  or
- CTAB reagents (< £0.25 per sample)
- May require purification (spin column or SPRI bead) (ca. £1 per sample)

DNA quality control:

- Qubit reagents - (£70 for 100 assays, £60 for 500 tubes)
- Agarose and loading dye (£0.20 per sample)
- (BioAnalyzer or TapeStation reagents) (ca. £50 for an 11-sample chip)
- (SPRI beads or spin-column cleanup kit) (ca. £1 per sample)

REQUIRED CONSUMABLES

Library preparation:

- Covaris tubes for sonicating DNA x 25 (£98)
- Library preparation kit
  - Comes with everything: Illumina TruSeq Nano (96) (£2400)
  - Need to buy beads and adapters:
    - NEBNext Ultra (£1,661)
    - KAPA Hyper Prep (£1,740)
- Ethanol and PCR grade water
- SPRI beads (for non-Illumina kits), 60 mL (£768)
- Adapters (for non-Illumina kits) (NEBNext® set 1 of 96: £432.00)
- 1.5 mL tubes and/or 15 mL tubes
- PCR plates: up to 7 per plate of libraries (Illumina kit)
- Plate seals x 100 (£135.00)
- Pipette tip boxes per plate of libraries:
  - 10 uL x 10
  - 100-200 uL x 20
  - 1000 uL < 1

REQUIRED CONSUMABLES

Library quality-control

- Qubit reagents (£70 for 100 assays)
- BioAnalyzer or TapeStation or qPCR reagents.

Hybridization:

- MYbaits kit (£2,000 for the smallest custom kit: baits for ca. 500 samples plus reagents for 12 samples — or £800 for the PAFTOL Angiosperm-wide kit)
- Dynabeads® MyOne™ Streptavidin C1, 2 mL (£310.00)
- KAPA HiFi PCR Kit x 100 reactions (£486)
  - Or other PCR master mix
- MYbaits reagent kit (£300, 48 reactions)
- PCR tubes (2 per reaction)
- LoBind tubes x 100 (£20) - need 2 per reaction
- PCR-grade water

Sequencing

- HiSeq 2000 2x100: ca. £2,500
- MiSeq v3 2x300: ca. £1,500

SAMPLE BUDGET - PILOT STUDY ON THE CHEAP

48 samples, silica-dried

• Low-coverage genome (1 HiSeq lane)	£2,500
• Qiagen DNeasy Plant Mini kit	£130
• Qubit assays x 100 + bag of tubes	£175
• Covaris tubes x 25 (will clean and reuse once)	£98
• Illumina TruSeq Nano HT	£2,400
• MYbaits kit -	£2,400
• Streptavidin beads	£310
• BioAnalyzer chip x 1	£50
• General lab consumables (tubes, plate, etc.)	£200
• Use existing PCR master mix	£0
• MiSeq v3 600 cycle kit	£1,500
	£9,263

SAMPLE BUDGET - PILOT STUDY ON THE CHEAP

48 samples, silica-dried

~~Low-coverage genome (1 HiSeq lane)~~	~~£2,500~~
Qiagen DNeasy Plant Mini kit	£130
Qubit assays x 100 + bag of tubes	£175
Covaris tubes x 25 (will clean and reuse once)	£98
Illumina TruSeq Nano HT	£2,400
MYbaits kit -	£2,400
Streptavidin beads	£310
BioAnalyzer chip x 1	£50
General lab consumables (tubes, plate, etc.)	£200
Use existing PCR master mix	£0
MiSeq v3 600 cycle kit	£1,500
	£6,763

Maybe you can use existing genomic resources …

SAMPLE BUDGET - PILOT STUDY ON THE CHEAP

48 samples, silica-dried

• ~~Low-coverage genome (1 HiSeq lane)~~	~~£2,500~~	
• Qiagen DNeasy Plant Mini kit	£130	
• Qubit assays x 100 + bag of tubes	£175	
• Covaris tubes x 25 (will clean and reuse once)	£98	
• Illumina TruSeq Nano HT	~~£2,400~~	1200
• MYbaits kit	£2,400	
• Streptavidin beads	£310	
• BioAnalyzer chip x 1	£50	
• General lab consumables (tubes, plate, etc.)	£200	
• Use existing PCR master mix	£0	
• MiSeq v3 600 cycle kit	~~£1,500~~	750
	£4,813	

…and split the library and run kits with a colleague…

SAMPLE BUDGET - PILOT STUDY ON THE CHEAP

48 samples, silica-dried

• ~~Low-coverage genome (1 HiSeq lane)~~	~~£2,500~~
• Qiagen DNeasy Plant Mini kit	£130
• Qubit assays x 100 + bag of tubes	£175
• Covaris tubes x 25 (will clean and reuse once)	£98
• Illumina TruSeq Nano HT	~~£2,400~~ 1200
• MYbaits kit -	~~£2,400~~ 800
• Streptavidin beads	£310
• BioAnalyzer chip x 1	£50
• General lab consumables (tubes, plate, etc.)	£200
• Use existing PCR master mix	£0
• MiSeq v3 600 cycle kit	~~£1,500~~ 750
	£3,713

…and use the PAFTOL kit!

ACTIVITY: DESIGN A PROJECT

1. Fresh samples, no genomic resources, phylogenetics
You would like to reconstruct the phylogeny of a medium-sized genus of rosids. Due the industrious collecting of your colleagues, you have at least two accessions each of all 47 species growing in your research greenhouse. However, due to the ugliness of the plants and their total lack of commercial value, no genomic resources exist within the relevant family (although published draft genomes exist for two taxa in the same order).

2. Fresh samples, transcriptome resources, functional genes
You are interested in co-evolution of inflorescence size in your pet genus with jaw width in a clade of florivorous Rodents of Unusual Size; as part of this investigation, you would like to test for selection in candidate genes involved in floral development. You have therefore joined a large collaborative project investigating directional selection in 108 MADS-box genes in a clade of angiosperms containing two orders, six families, 76 genera (including yours), and 2,382 species. After missing a critical task-assignment meeting, you find yourself tasked with designing a HybSeq study to capture the genes. You have access to one transcriptome assembly per family for five families, and three transcriptomes for the sixth family, to aid you in developing your markers.

3. Mix of herbarium and fresh samples, low-coverage genome, functional genes
In a welcome stoke of good fortune, you have been awarded a seed grant to investigate pseudogenization and copy-number variation in 112 putative defense genes in a 4-species complex of a commercially-important horticultural plant with beautiful but disease-prone flowers. You suspect that a domestication bottleneck has led to disease-prone plants in cultivation. Available material includes living collections of 27 modern named varieties, and historical specimens in herbaria from 18 wild populations, some of which you suspect may have been sources of material brought into cultivation in the 19th century. You have a low-coverage genome assembly for one of the modern varieties as well as a published draft genome from the sister family.

4. Mix of herbarium and fresh samples, no genomic resources, phylogenetics
Due to an extremely ill-advised bet, your general well-being heavily depends on reconstructing an accurate phylogeny of two sister ascomycete genera containing 142 species. Although you have access to silica-dried material for 27 species, the remaining species are available only in herbaria due to a combination of extinction and a prolonged civil war making fieldwork impractical. For those 115 species, the most recently collected specimen ranges from 110 to 7 years old. Four species exist only as type specimens in a small regional herbarium, so, in the words of the curator, “this had better work.” No genomic resources exist within the family, but one low-coverage published draft genome exists in the sister family.

ACTIVITY: DESIGN A PROJECT

Outline a strategy for developing HybSeq baits appropriate for your study.
Be sure to include

- A proposed method for marker development, including any required sequencing
- Sources of phylogenetically-appropriate bait sequences
- The estimated total cost of bait development.

Roughly estimate the cost of the lab work required to carry out the project.
At this stage in the workshop it will only be a rough estimate. Estimate the number of samples and cost of consumable required (including kits) for the following:

- DNA extraction
- Library preparation
- Hybridization
- Sequencing

Add up all your cost estimates.
Then suppose you only have one-quarter of these funds available. How would you structure a pilot study that would (hopefully) produce enough results to justify a request for full funding?