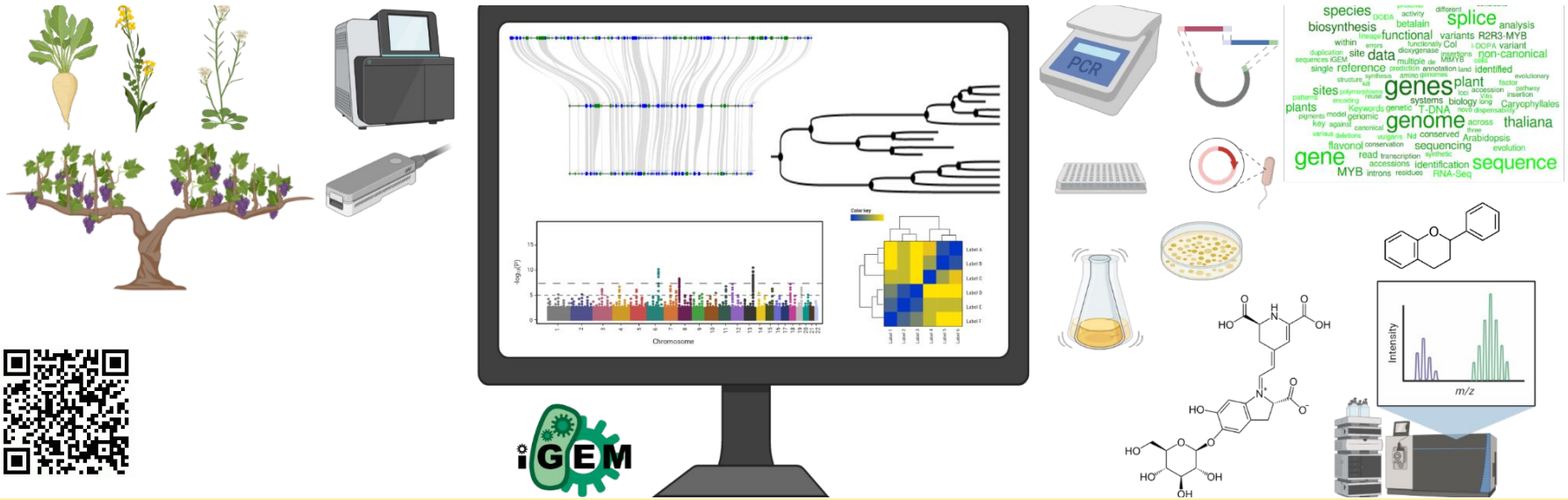
Technische Universität Braunschweig



GE32/MM12 - Data Life Cycle

Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

Availability of slides

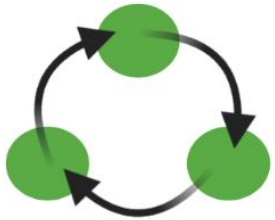
- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE32/MM12**
  - GitHub: https://github.com/bpucker/teaching

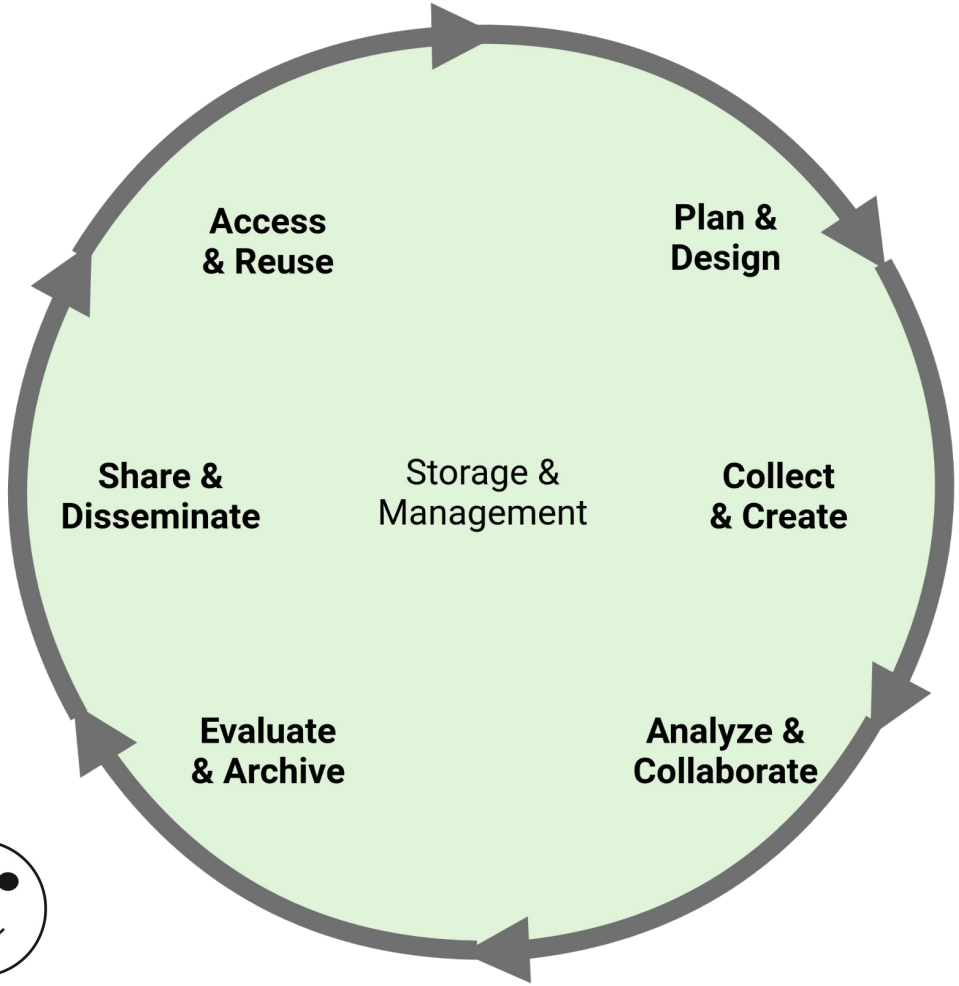
- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-braunschweig.de

What is a data life cycle?

Data life cycle





PLAN

https://commons.wikimedia.org/wiki/File:RDM_02_Collect-and-Capture_frame01.svg
https://www.picpedia.org/keyboard/a/analyze.html

Plan & design

- What are the research questions/objectives?
- Plan analysis/experiment that will generate data OR use existing data sets
- How many replicates? Which statistical tests (power)?
- Data management plan
- How to ensure data safety (backups)?

EXAMPLE: Plan & design

- Research question: What are the molecular mechanisms explaining white and red pigmentation of flowers?

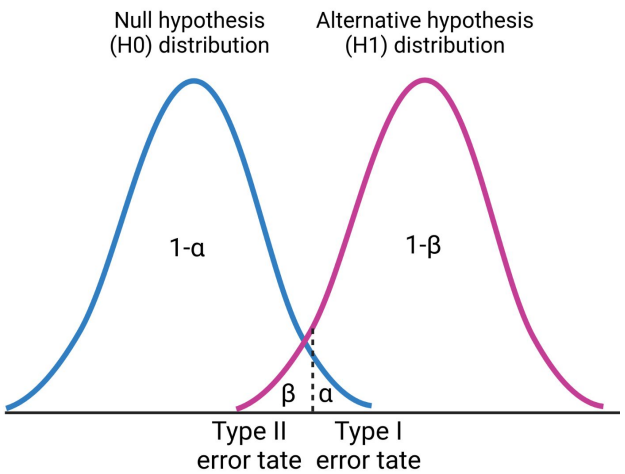
- Experiment:
  - RNA-Seq
  - >=3 replicates per morphotype
  - paired-end sequencing
  - 30 million tags

- Data storage:
  - on a local hard drive
  - backup in cloud
  - cold storage

Power of statistical tests

- Concept of statistical test: reject null hypothesis (e.g. no difference between samples)
- Power = probability to reject null hypothesis (H0) when it is false (1-ß)
- Size of the effect and size of sample determine power: larger samples lead to higher sensitivity

		Truth about the population	
		H0 true	H0 false
Decision based on sample	Reject H0	Type I error	Correct decision
	Accept H0	Correct decision	Type II error



Data management plan

- Tools are available to generate data management plans
- Research Data Management Organiser (RDMO)
- Funder-specific differences in expectations
- Content:
  - Information about data and data format
    - method/time of collection; version control; backup
  - Metadata content and format
  - Policies for access, sharing, and reuse
  - Long-term storage
  - Budget: considerable costs might arise

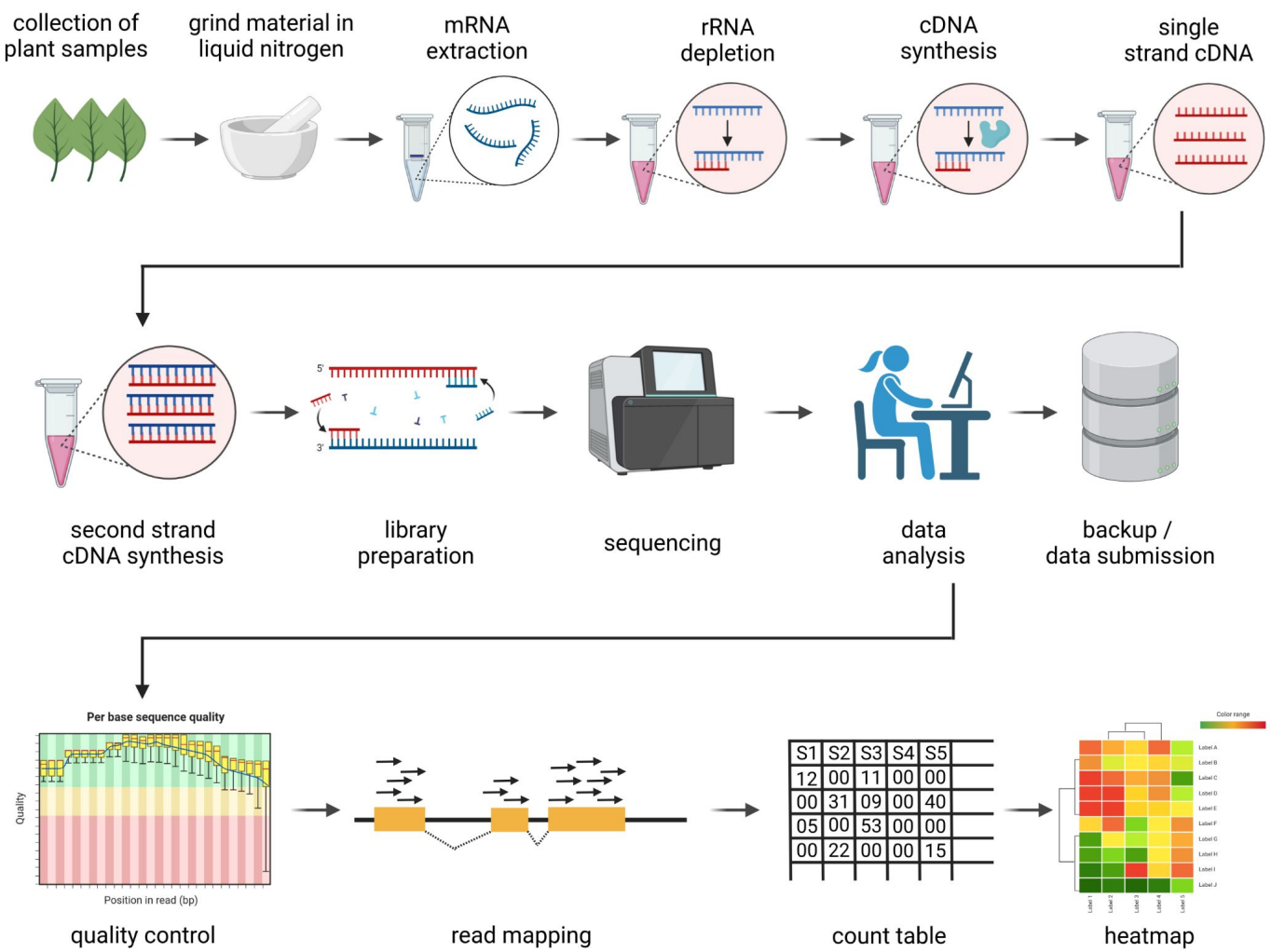
Collect & create

- Perform planned experiment with replicates
- Document all steps
- Collect the results (data)
- Synchronize results across all storage/backup locations

EXAMPLE: Collect & create

- Grow plants under precisely controlled conditions
- Harvest material in a reproducible manner; replicates are important!
- Extract RNA and subject to RNA-seq experiment
- Store resulting FASTQ files

RNA-Seq


collection of plant samples
grind material in liquid nitrogen
mRNA extraction
rRNA depletion
cDNA synthesis
single strand cDNA
5'
3'
second strand cDNA synthesis
library preparation
sequencing
data analysis
backup / data submission
Per base sequence quality
Quality
Position in read (bp)
S1 S2 S3 S4 S5
12 00 11 00 00
00 31 09 00 40
05 00 53 00 00
00 22 00 00 15
Color range
quality control
read mapping
count table
heatmap

Synchronize & backup

- Synchronization of data between to different locations:
  - $ rsync <SOURCE> <TARGET>

- Backup solutions:
  - Desktop computers & laptops
  - External hard drives
  - Network drives
  - Central storage options (e.g. cold storage; tape storage)
  - Cloud storage
  - Optical storage

- 3 copies of data are recommended; one off-site

Analyze & collaborate

- Quality control
- Plausibility checks
- Sample identity checks
- Perform actual analysis alone/in collaboration
- Interdependence of collaborators possible

EXAMPLE: Analyze & collaborate

- Quality control via principal component analysis (PCA)
- Differential gene expression analysis (DESeq2)
- Pathway enrichment analysis (KEGG, GO)

- How to exchange files with collaborators
  - cloud
  - hard drives

- How to document collaboration
  - Documentation of meetings
  - Documentation of contributions to project
  - Documentation of progress
  - Writing reports for funding agencies

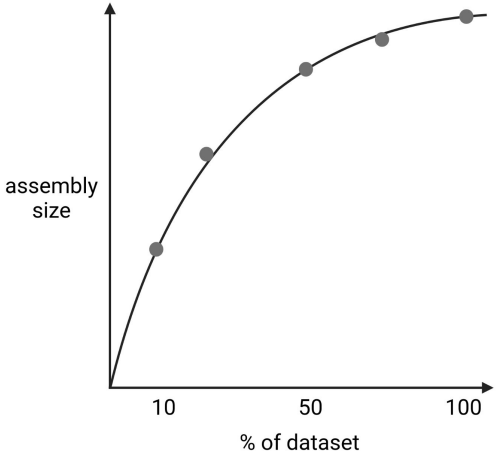
https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html

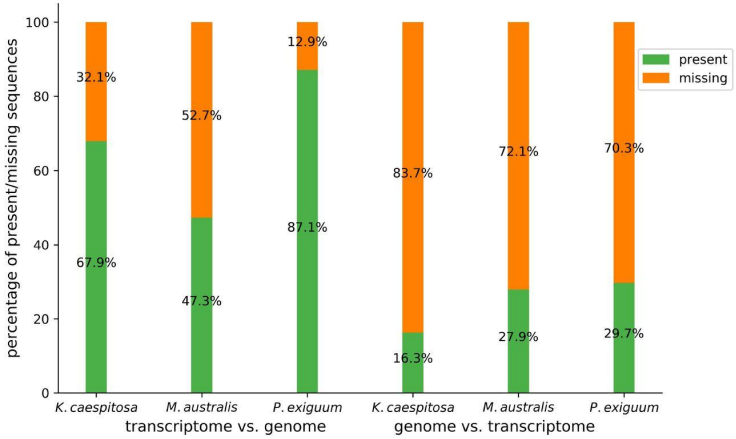
Evaluate & archive

- Are the data sets large enough?
- What could have been done better?
- Was the number of replicates sufficient?
- Are there clear differences between sample groups?
- Low variation between biological replicates?
- Archive all research data on tape storage for at least 10 years
- Off-site backups; Rsync to transfer only modified files
- Commercial clouds (Dryad, GigaDB); tape storage @ TUBS (contact GITZ)

Enough data

- Check if subsets lead to complete assemblies (saturation)
- Comparison of genome and transcriptome assembly
- Genome sequence assemblies cover lowly expressed genes
- Transcriptome assemblies can be advantageous for genes with large introns

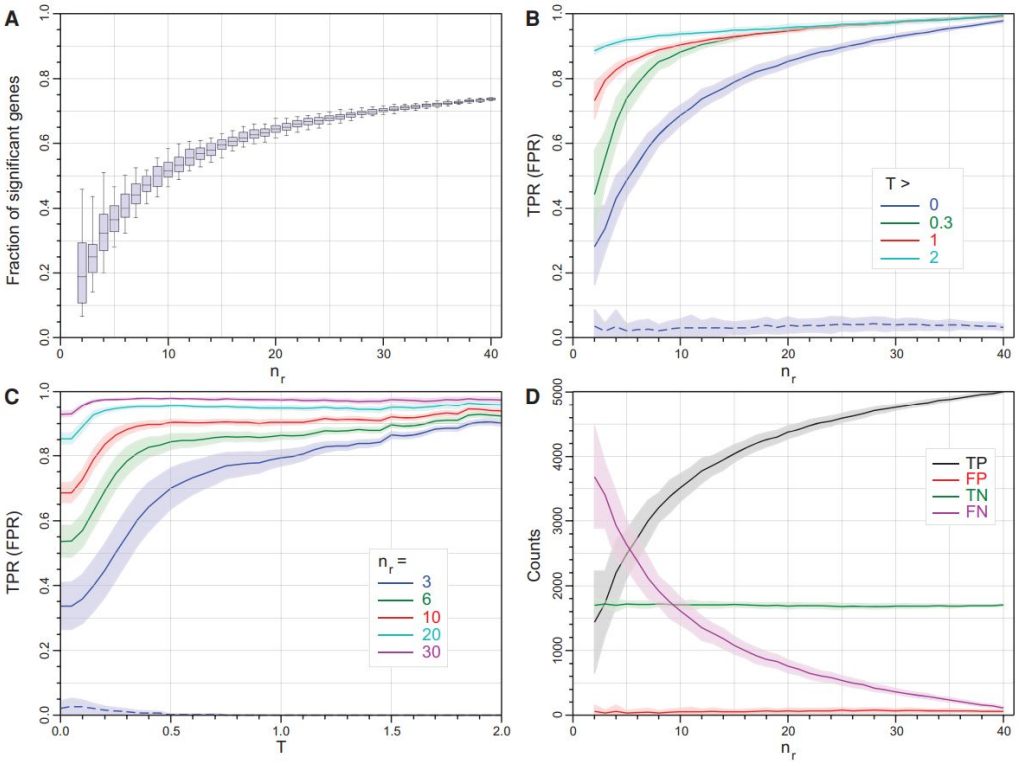




Haak et al., 2018: 10.3389/fmolb.2018.00062
Pucker et al., 2019: 10.1101/646133

Sufficient replicates

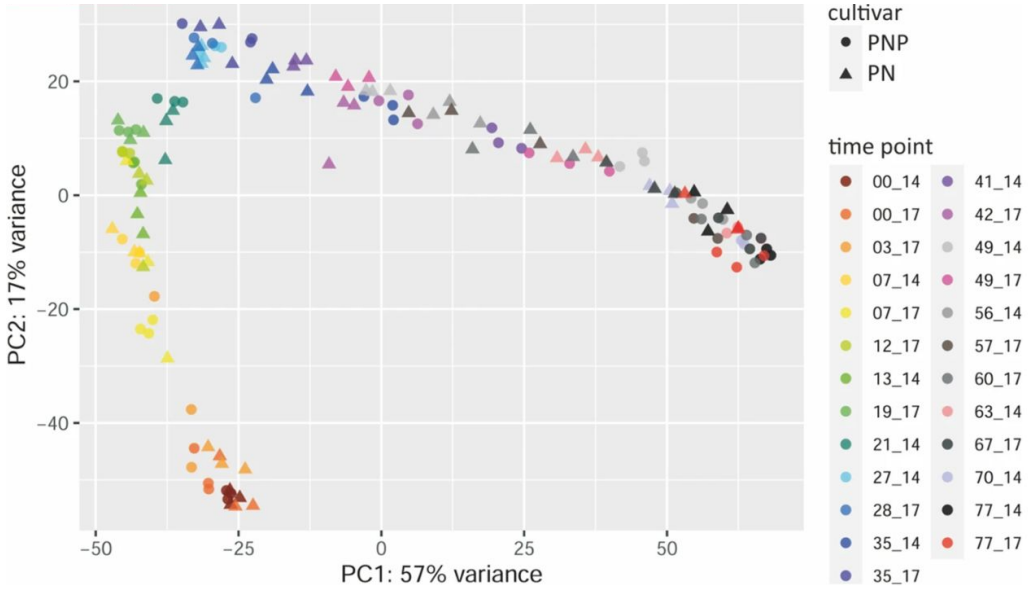
- More replicates allow the identification of a larger number of DEGs
- Reliability of gene classification increases with number of replicates
- Up to 40 replicates can boost signal strength



Schurch et al., 2016: 10.1261/rna.053959.115

Differences between groups

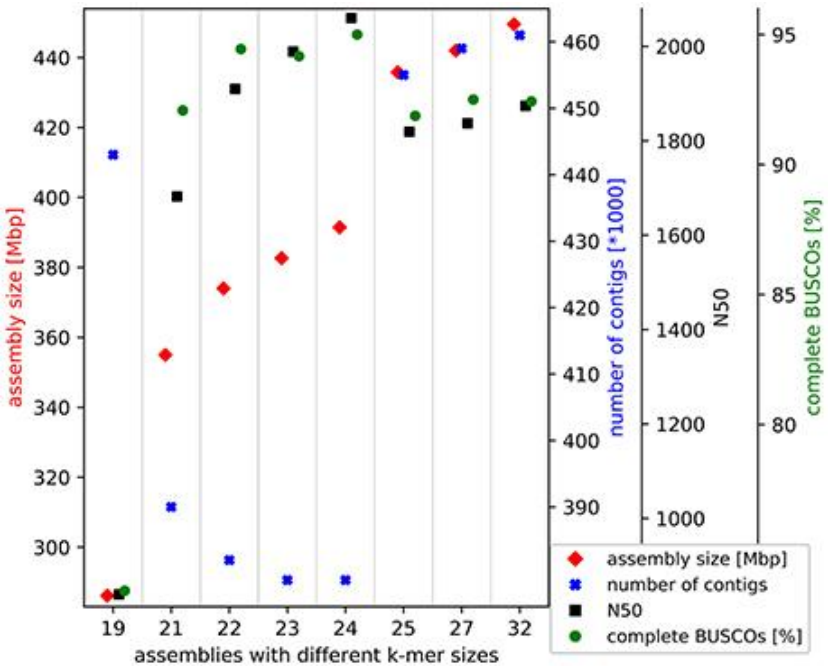
- Principal Component Analysis (PCA) separates RNA-seq samples based on gene expression patterns
- Similar samples are grouped together
- Principal Components (axes) are artificial axes



Theine et al., 2021: 10.1186/s12870-021-03110-6

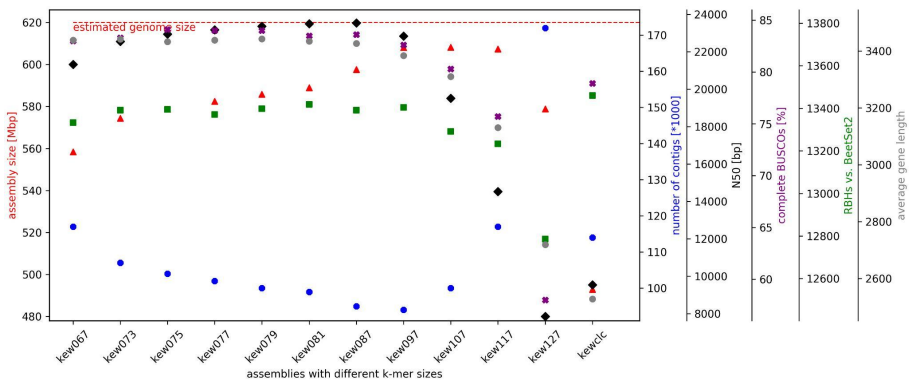
EXAMPLE: find best k-mer

- Central parameter in transcriptome assembly is k-mer size
- Empirical identification of best k-mer for data set leading to best transcriptome assembly
- Some parameters need to be optimized for each project



Haak et al., 2018: 10.3389/fmolb.2018.00

EXAMPLE: find best assembly settings



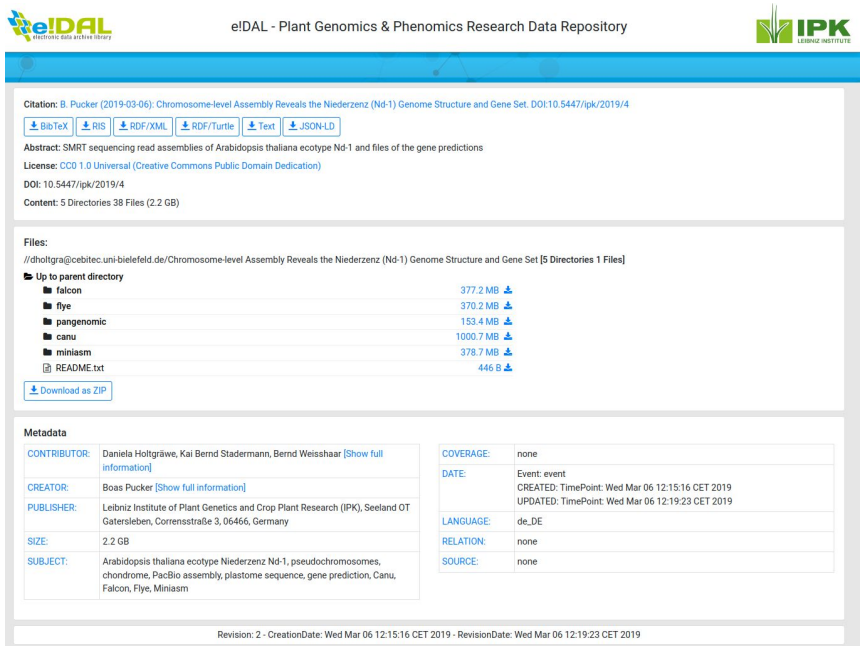
	kew067	kew073
Species	*Kewa caespitosa*	*Kewa caespitosa*
Assembler	SOAPdenovo2	SOAPdenovo2
K-mer size	67	73
Insert size	750	750
Number of contigs	117829	107585
Maximal contig length	308789	328441
N50	21320	22551
N90	2824	2977
assembly size	558347289	574375745
GC content	0.381104357793	0.381438787949
predicted genes	83397	88487
average gene length	3438.46340995	3441.81079707
RBHs vs. BeetSet2	13333	13391
BUSCO result	C:83.0%[S:81.7%,D:1.3%],F:6.7%,M:10.3%,n:1440	C:83.3%[S:82.0%,D:1.3%],F:6.5%,M:10.2%,n:1440

Assembly-ID	gene_numbe	gene_length	exons_per_gen	average_mRNA_lengt	average_peptide_length
Kewa_caespitosa_CDS	80296	3492	5.543396226415	1544	378
Kewa_caespitosa_FINAL	50661	5494	6.613225424398	2143	447
Kewa_caespitosa_RNA_seq	63573	3203	5.130969609262	1330	363
Kewa_caespitosa_ab_initio	80728	3424	5.442231223754	1548	378
Kewa_caespitosa_masked_ab_initio	70478	3406	5.456631241398	1522	374
Macarthuria_australis_CDS	110884	1615	2.682485141997	888	215
Macarthuria_australis_FINAL	80236	1936	2.938829687792	1018	241
Macarthuria_australis_ab_initio	98131	1907	2.467706104148	911	228
Macarthuria_australis_masked_ab_initio	97530	1589	2.688461893386	871	213
Pharnaceum_exiguum_CDS	39972	3770	5.438743088739	1643	383
Pharnaceum_exiguum_FINAL	26155	5090	6.077690690117	2154	435
Pharnaceum_exiguum_ab_initio	40612	3614	5.20154637906	1649	382
Pharnaceum_exiguum_ab_inito_spec_param	29558	5550	6.087018303617	2335	483
Pharnaceum_exiguum_masked_ab_initio	34163	3585	5.128007727885	1585	368

Pucker et al., 2020: 10.1101/646133

EXAMPLE: archive

- Submission of Nd-1 data sets to e!DAL
- Citable in corresponding publication via DOI
- Data are also archived on tape storage



e!DAL - Plant Genomics & Phenomics Research Data Repository

Citation: B. Pucker (2019-03-06): Chromosome-level Assembly Reveals the Niederzenz (Nd-1) Genome Structure and Gene Set. DOI:10.5447/ipk/2019/4

BibTeX | RIS | RDF/XML | RDF/Turtle | Text | JSON-LD

Abstract: SMRT sequencing read assemblies of Arabidopsis thaliana ecotype Nd-1 and files of the gene predictions

License: CC0 1.0 Universal (Creative Commons Public Domain Dedication)

DOI: 10.5447/ipk/2019/4

Content: 5 Directories 38 Files (2.2 GB)

Files:

//dholtgra@cebitec.uni-bielefeld.de/Chromosome-level Assembly Reveals the Niederzenz (Nd-1) Genome Structure and Gene Set **[5 Directories 1 Files]**

Up to parent directory	
falcon	377.2 MB
flye	370.2 MB
pangenomic	153.4 MB
canu	1000.7 MB
miniasm	378.7 MB
README.txt	446 B

Download as ZIP

Metadata

CONTRIBUTOR:	Daniela Holtgräwe, Kai Bernd Stadermann, Bernd Weisshaar [Show full information]
CREATOR:	Boas Pucker [Show full information]
PUBLISHER:	Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland OT Gatersleben, Corrensstraße 3, 06466, Germany
SIZE:	2.2 GB
SUBJECT:	Arabidopsis thaliana ecotype Niederzenz Nd-1, pseudochromosomes, chondrome, PacBio assembly, plastome sequence, gene prediction, Canu, Falcon, Flye, Miniasm

COVERAGE:	none
DATE:	Event: event CREATED: TimePoint: Wed Mar 06 12:15:16 CET 2019 UPDATED: TimePoint: Wed Mar 06 12:19:23 CET 2019
LANGUAGE:	de_DE
RELATION:	none
SOURCE:	none

Revision: 2 - CreationDate: Wed Mar 06 12:15:16 CET 2019 - RevisionDate: Wed Mar 06 12:19:23 CET 2019

Share & disseminate

- All data sets underlying a publication should be released
- Clearly refer to published data sets in data availability statement
- Submission of large data sets to suitable databases
- Select proper license for data sets
- Include all ‘customized scripts’ in publication

EXAMPLE: Share & disseminate

- Options to submit data sequencing data: ENA, SRA (GEO)
- Submission of individual sequences: NCBI
- Options for geoposition data sets: GBIF
- Options to share data sets (github, dryad, GigaDB, PUB, TUBS?)
- Scripts can be shared via github, bitbucket, gitlab, codeberg

EXAMPLE: data availability statement

- Sequence read datasets generated and analyzed during this study were made available at ENA under the accession PRJEB35658. Individual run IDs are included in Additional file 1. The Col-0 genome sequence assembly of the GABI-Kat Col-0 genetic background (Col-0_GKat-wt) is available at ENA under the accession GCA_905067165.

- **Availability of supporting code and requirements**
  - Project name: KIPEs3
  - Project home page: https://github.com/bpucker/KIPEs
  - Operating system(s): Linux (website is platform independent)
  - Programming language: Python3
  - Other requirements: BLAST, MAFFT, FastTree2, dendropy, scipy
  - License: GNU General Public License v3.0
  - RRID: SCR_022370

- All data sets analyzed in this study are publicly available. Data sets generated as part of this study are shared via GitHub (https://github.com/bpucker/KIPEs). A docker image is available via DockerHub (https://hub.docker.com/r/bpucker/kipes).

Pucker et al., 2021: 10.1186/s12864-021-07877-8
Rempel & Pucker, 2022: 10.1101/2022.06.30.498365

EXAMPLE: data publication

PUB - Publications at Bielefeld University

Arabidopsis thaliana methylation pattern analysis based on ONT sequence reads

Schilbert H, Kleinbölting N, Weisshaar B, Pucker B (2021)
Bielefeld University.

Research Data

Download
- GK_040A12.vcf.gz 176.58 MB
- GK_050B11.vcf.gz 183.65 MB
- GK_082G09.vcf.gz 185.46 MB
- + All

DOI https://doi.org/10.4119/unibi/2956654

Details | Files | Links

Creator: Schilbert, Hanna[UniBi]; Kleinbölting, Nils[UniBi]; Weisshaar, Bernd[UniBi]; Pucker, Boas[UniBi]

Department: Fakultät für Biologie > Genetik und Genomik der Pflanzen
Centrum für Biotechnologie > Research Group B. Weisshaar

Abstract / Notes:

We sequenced the genomes of 14 Arabidopsis thaliana GABI-Kat T-DNA insertion lines (Col-0 background), which eluded flanking sequence tag-based attempts to fully characterize their insertion alleles, with Oxford Nanopore Technologies (ONT) long reads. Detailed information about each line, e.g. their T-DNA insertion site(s), have been described by Pucker et al. (BMC Genomics (2021) 22:599; https://doi.org/10.1186/s12864-021-07877-8). The DNA that has been sequenced is derived from a mixtures of tissues as expected for DNA extracted from young plantlets. For 11 of the datasets from these lines, we called 5mC methylation patterns with the tool Megalodon v.2.2.9 provided by ONT (https://github.com/nanoporetech/megalodon/). Results of this analysis are available per line as individual VCF files. We are sharing unfiltered output files to grant full control over the down-stream analysis steps and to accelerate the research in epigenomics. We identified between 706,254 and 1,020,654 positions per line where at least 80% of 5 or more reads support a methylation site. The following datasets are available and the file names reveal the respective GK line:

1. GK_040A12.vcf
2. GK_050B11.vcf
3. GK_082G09.vcf
4. GK_290G05.vcf
5. GK_399C06.vcf
6. GK_410B07.vcf
7. GK_430F05.vcf
8. GK_433E06.vcf
9. GK_654A12.vcf
10. GK_767D12.vcf
11. GK_909H04.vcf

Keywords: long read sequencing; ONT; methylation; plant epigenomics; GABI-Kat; Megalodon

Year of Publication: 2021

Copyright and Licenses: Creative Commons Attribution 4.0 International Public License (CC-BY 4.0)

Page URI: https://pub.uni-bielefeld.de/record/2956654

Cite this

AMA | APA (6th ed.) | Frontiers | Harvard | IEEE | LNCS | MLA

Schilbert H, Kleinbölting N, Weisshaar B, Pucker B. *Arabidopsis thaliana methylation pattern analysis based on ONT sequence reads*. Bielefeld University; 2021.

PUB - Publications at Bielefeld University

Arabidopsis thaliana methylation pattern analysis based on ONT sequence reads

Schilbert H, Kleinbölting N, Weisshaar B, Pucker B (2021)
Bielefeld University.

Research Data

Download
- GK_040A12.vcf.gz 176.58 MB
- GK_050B11.vcf.gz 183.65 MB
- GK_082G09.vcf.gz 185.46 MB
- + All

DOI https://doi.org/10.4119/unibi/2956654

Details | Files | Links

All files available under the following license(s):

Creative Commons Attribution 4.0 International Public License (CC-BY 4.0):
https://creativecommons.org/licenses/by/4.0/
https://creativecommons.org/licenses/by/4.0/legalcode

Main File(s)

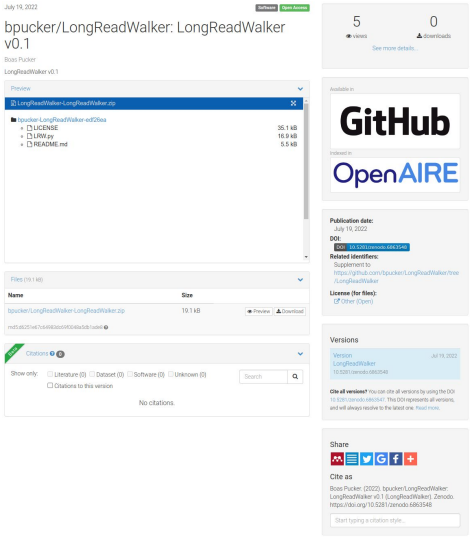
File Name	Access Level	Last Uploaded	MD5 Checksum
GK_040A12.vcf.gz 176.58 MB	Open Access	2021-08-09T16:28:06Z	5639514d2396b649688a1c08c9bb7ca7
GK_050B11.vcf.gz 183.65 MB	Open Access	2021-08-09T16:28:07Z	8ad8a957b3a7c48fa3f8bcc2eaff4178
GK_082G09.vcf.gz 185.46 MB	Open Access	2021-08-09T16:28:09Z	ae122833858f8e826f51d465c377443c
GK_290G05.vcf.gz 199.73 MB	Open Access	2021-08-09T16:28:09Z	df135692cab92ae1d4e1a0ff852c0691
GK_399C06.vcf.gz 181.61 MB	Open Access	2021-08-09T16:28:12Z	59202cb9168b9c5a9b5d72282ee17a66
GK_410B07.vcf.gz 200.89 MB	Open Access	2021-08-09T16:28:13Z	608d96b3879e79170b51cab0d78c01d9
GK_430F05.vcf.gz 197.97 MB	Open Access	2021-08-09T16:28:15Z	eb9e3d4f957f33fb623d9e6d44cb00f3
GK_433E06.vcf.gz 192.92 MB	Open Access	2021-08-09T16:28:16Z	4ecfa41f5aa385e5c8b851f700e1dc1f
GK_654A12.vcf.gz 187.28 MB	Open Access	2021-08-09T16:28:19Z	29822d0433e5178e0110573213d76886

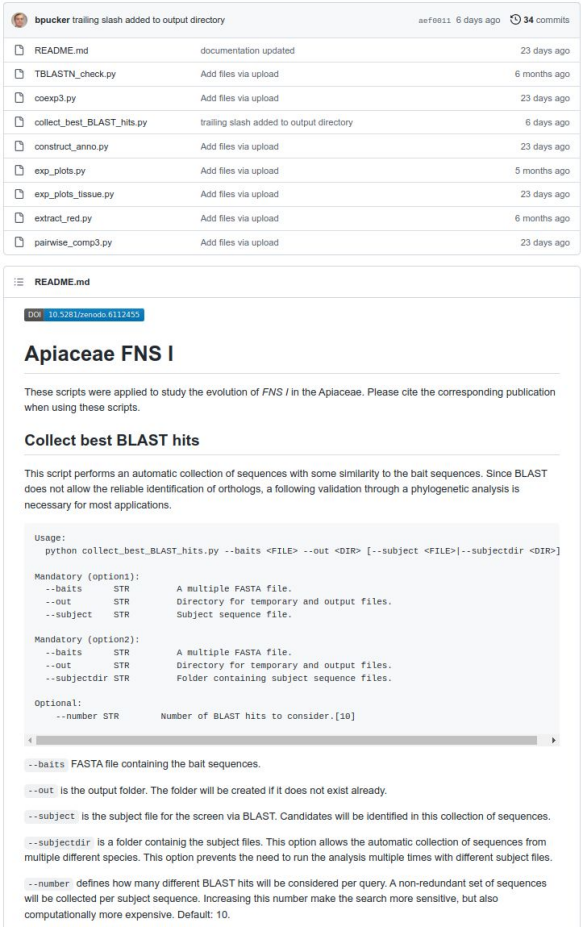
Schilbert et al., 2021: 10.4119/unibi/2956654

EXAMPLE: publication of scripts

- Scripts should be shared through repositories:
  - Github
  - Bitbucket
  - Codeberg
  - Gitlab

- Archive repositories in Zenodo (DOI assignment)





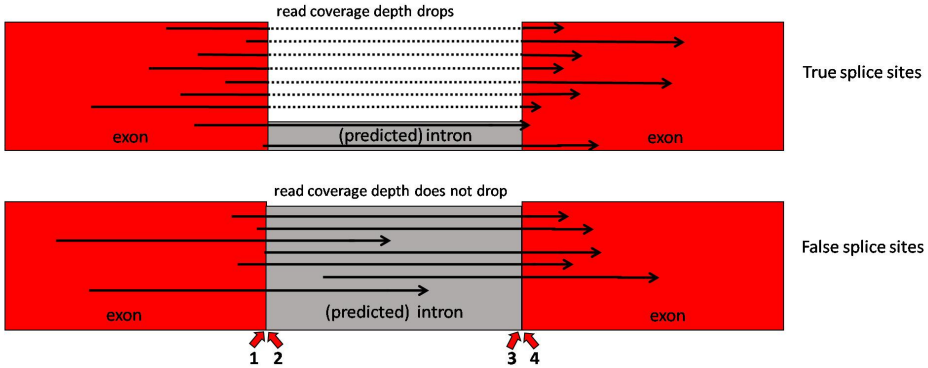
https://github.com/bpucker/ApiaceaeFNS1

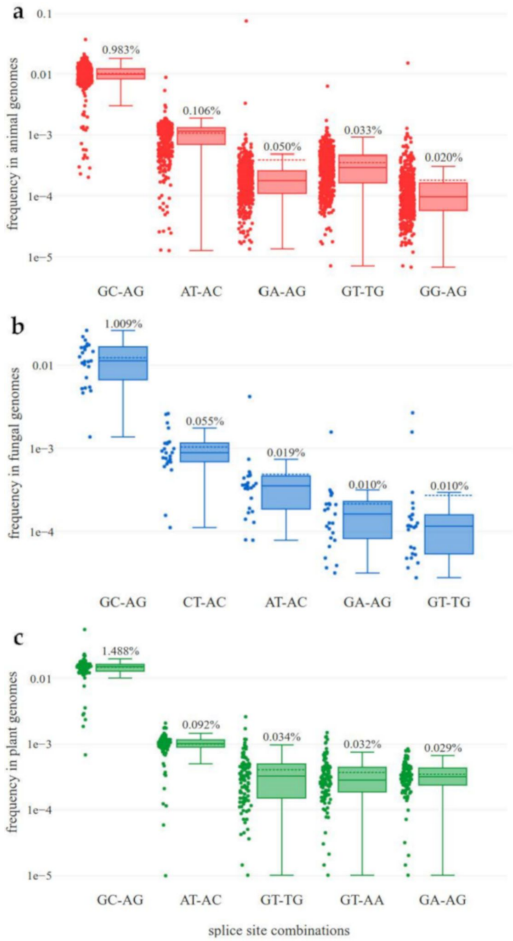
Access & reuse

- Data sets should be freely available to enable validation of findings
  - NOT: ‘Available upon reasonable request from the corresponding author’
- Public data sets are an excellent resource for re-use
- No costs for data generation
- Massive datasets can enable identification of small effect sizes
- Generation of novel hypothesis

EXAMPLE: Access & reuse

- Investigation of non-canonical splice sites in plants, animals, and fungi
- Screening genome sequences+annotations for splice sites
- Harnessing RNA-seq to quantify usage of (non-canonical) splice sites

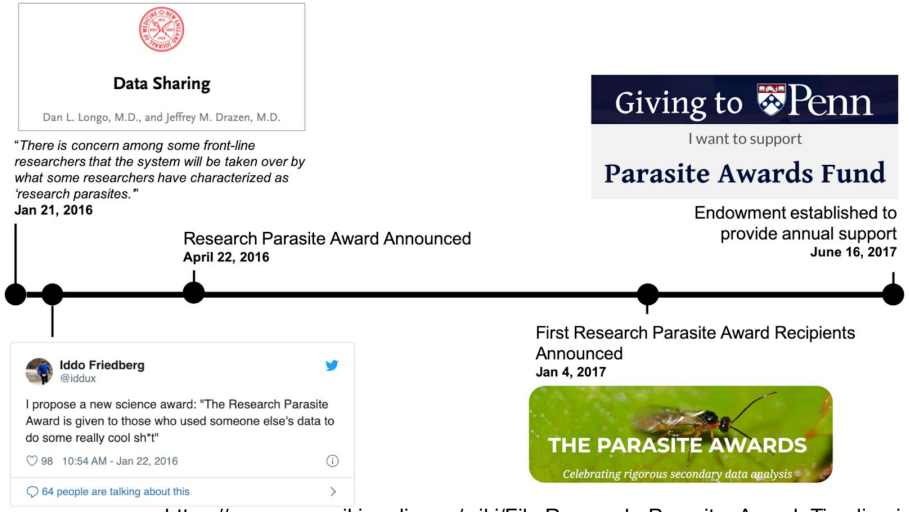




Pucker & Brockington, 2018: 10.1186/s12864-018-5360-z
Frey & Pucker, 2020: 10.3390/cells9020458

The Parasite Awards

- Celebrates comprehensive secondary data analysis
- Novel insights inferred from existing (underutilized) data sets
- Supported by GigaScience/GigaByte



https://commons.wikimedia.org/wiki/File:Research_Parasite_Award_Timeline.jpg

https://researchparasite.com/

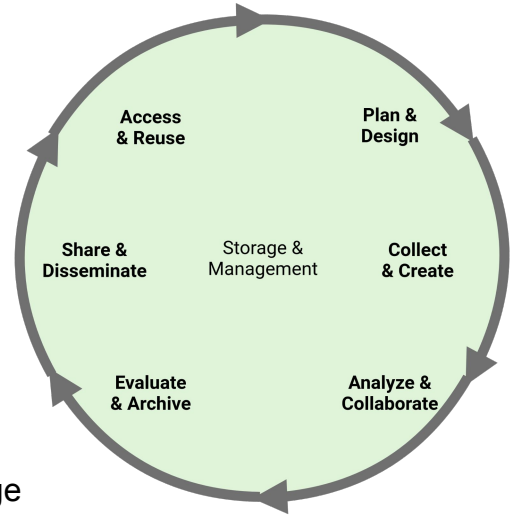
Full cycle: Nd-1

Data sets reused for variant caller benchmarking study

Publication about findings

data sets published at SRA

NGS not sufficient for complete genome sequence; on tape storage



Sequence Nd-1 genome, because it is a parent of a mapping population

Sequencing on Roche 454, GAIIx, & HiSeq1500

fastQC analysis, trimming with Trimmomatic, and assembly+annotation

Full cycle: *Croton tiglium*

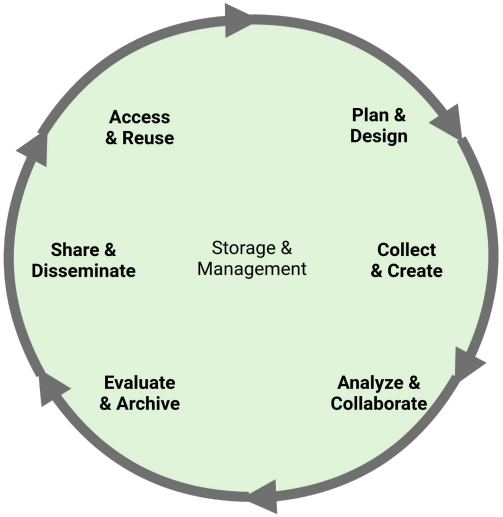
Data reuse for validation of KIPEs

Publication about findings

data sets published at SRA

Evaluation to identify optimal parameters

Data stored via tape storage



Identify specific genes in *Croton tiglium* via transcriptome assembly

RNA-seq with samples of different tissues & normalized library

Transcriptome assembly and characterization of candidate genes

Time for questions!

Questions

1. What are the important stages of a data life cycle?
2. What are the steps of a typical RNA-seq analysis?
3. What considerations about backups are important?
4. Where would you store data related to your publication?
5. How would you assess sequence data sets?
6. How can you share your scripts?