Analysing Genomic Data with **dartRverse**: Accessible Tools for Conservation

UNIVERSITY OF CANBERRA

DArT diversity arrays technology

MONASH University

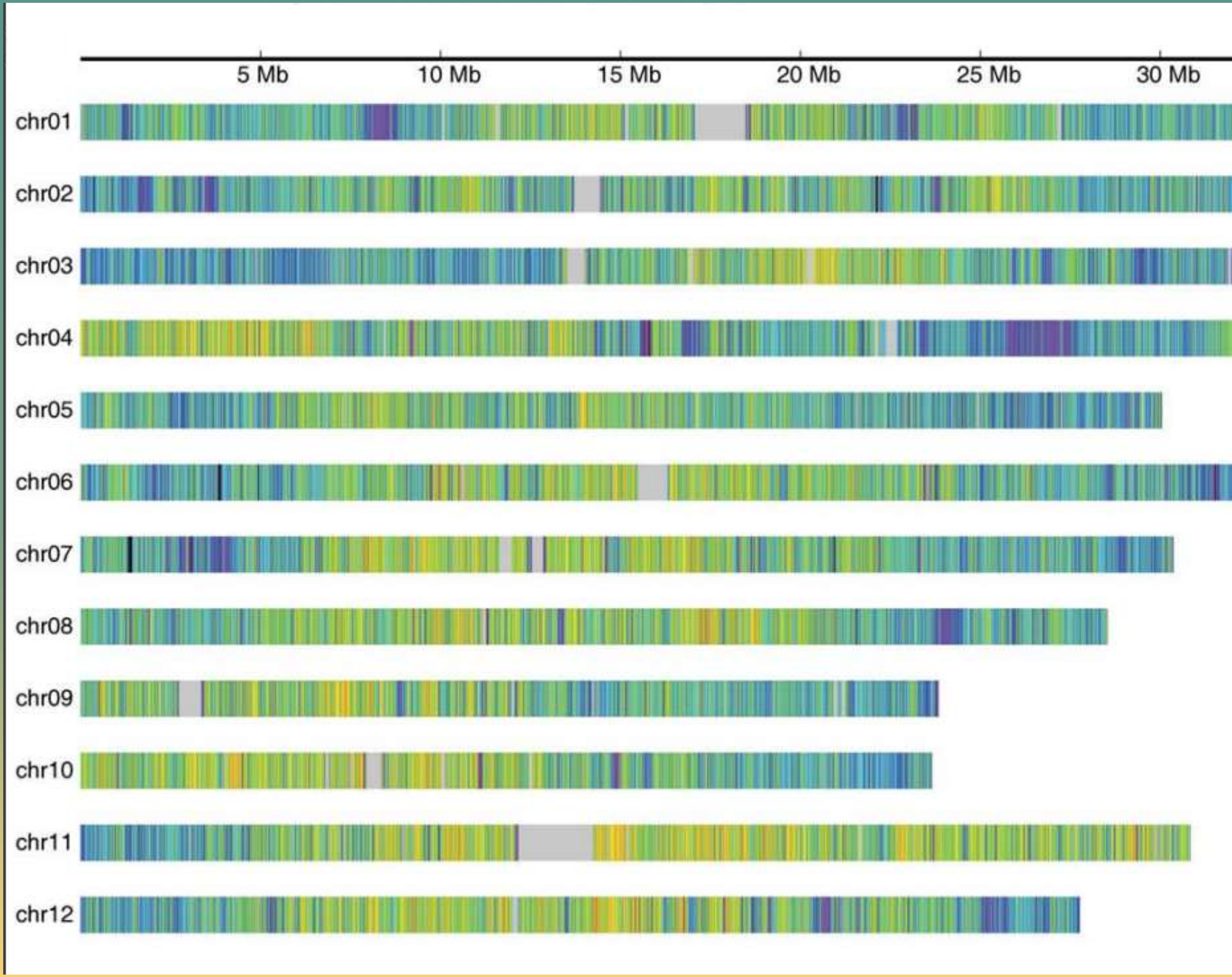
Centre for Biodiversity Analysis

Australian Government
Australian Research Council

Session 6: SNP Panel Selection

Elise Furlan
Andrzej Kilian
Bernd Gruber

Elise.Furlan@canberra.edu.au



Aim

Provide the knowledge and practical tools to reduce large SNP datasets into smaller, targeted SNP panels for conservation monitoring.

Background

- Targeted sets of SNP markers – 10s to100s
- Reproducible
- Cost-effective for high sample volume
- Suitable for low-quality or low-quantity DNA samples

Purpose

SNP panels can be used to address either specific questions or to span multiple conservation genetic applications

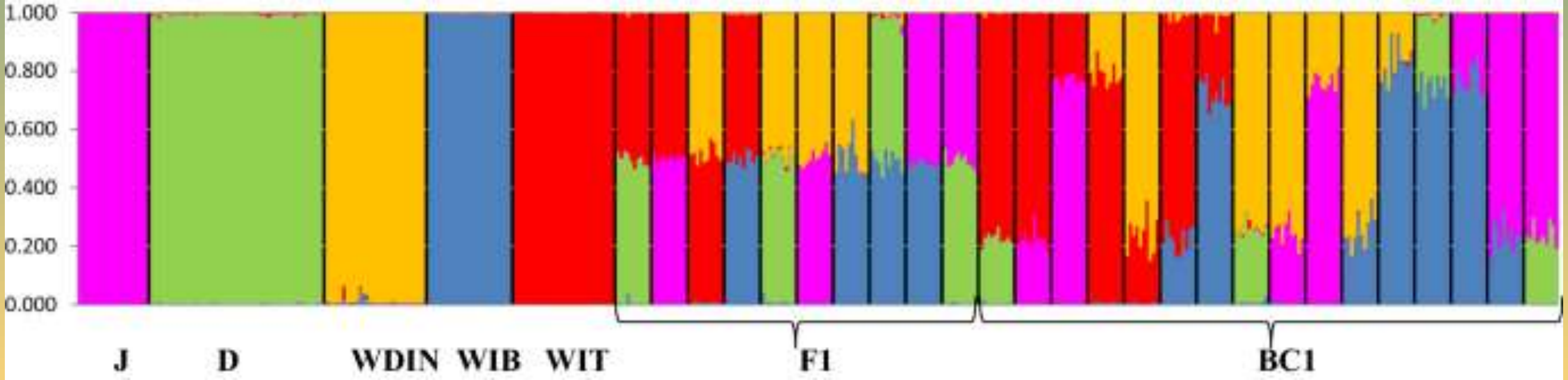
- **Population assignment**
- **Parentage or relatedness**
- **Individual ID**
- **Hybridisation**
- **Additional metrics**
  - Sex-linked SNPs
  - Candidate adaptive markers
  - Diagnostic SNPs for population ID
  - Phenotypic markers

Cheetah Conservation Fund, Magliolo et al. 2021

Example - Hybridisation

Stronen et al., 2022

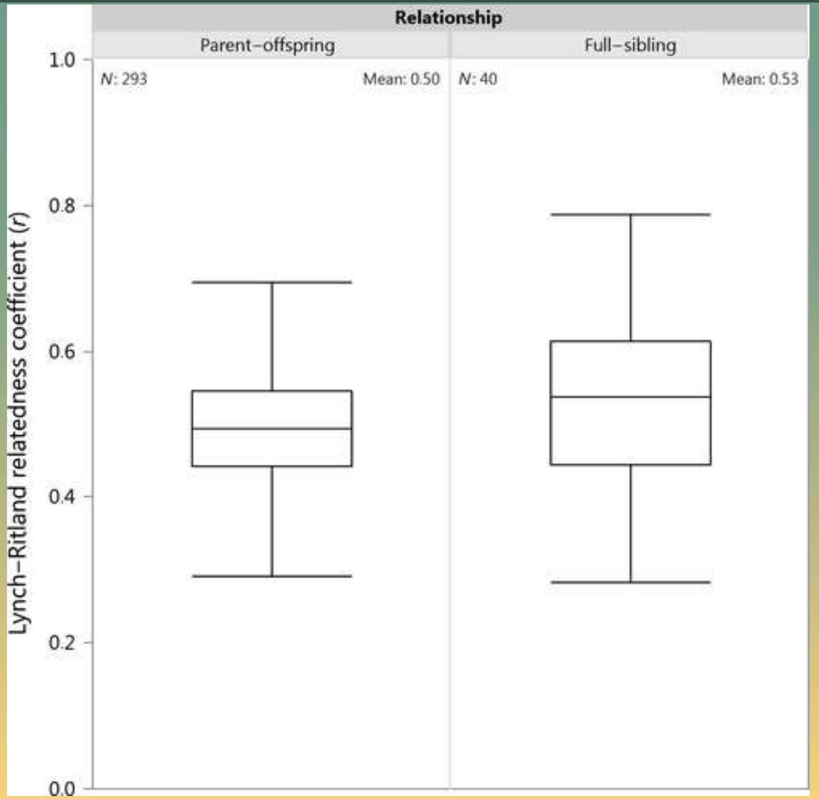
- 192 SNP genotypes
- Differentiate 5 canid species.
  - Jackals (J), dogs (D), Dinaric wolves (WDIN), Iberian wolves (WIB) and Italian wolves (WIT)
- Identify hybrids
  - First-generation (F1) hybrid and first-generation backcross (BC1) genotypes
- Included 3x phenotypic markers relating to coat colour, nail colour and dewclaw presence (absent in wild canids)



Example - Non-invasive samples

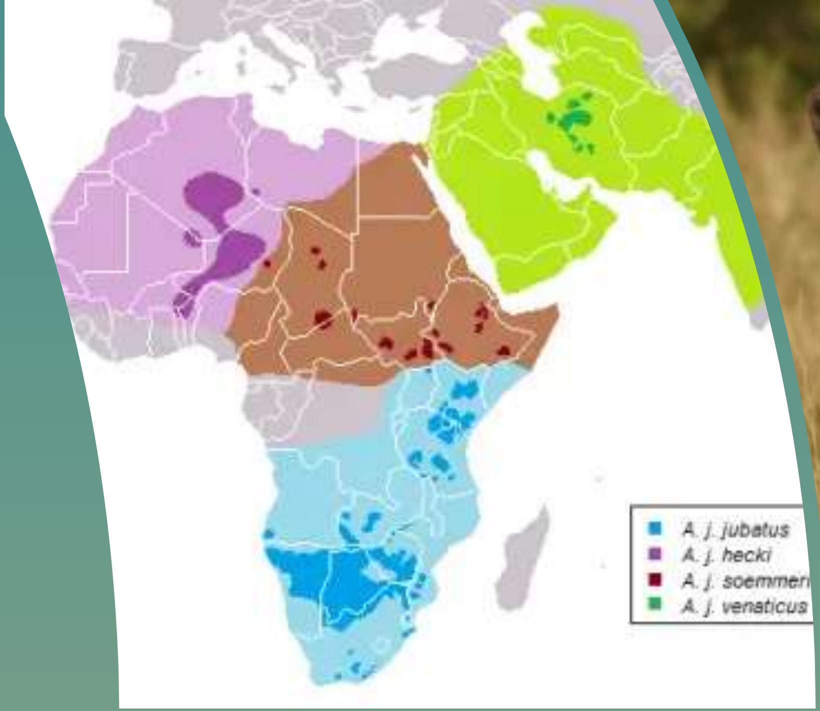
- 96-SNPs used on faecal samples in brown bears
- Estimated population size
  - Fell within the 95% CI of Capture-Mark-Recapture estimates
- Estimated relatedness
- Determined sex
  - using sex-linked markers

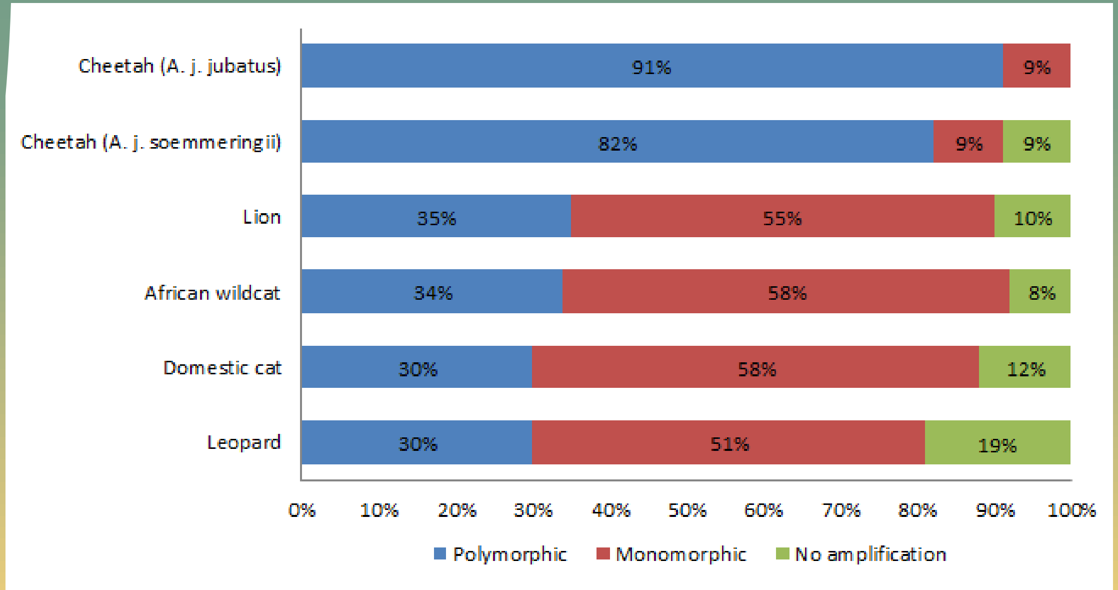
Spitzer et al., 2016, Source: Nyhetsbyrån



Marker Selection

- SNP panels must be carefully selected to maximise informativeness for the specific application
  - Less room for redundancy
- Requires genome-wide SNP data from individuals spanning the full distribution
  - Avoid ascertainment bias and loss of power in other species/populations
  - SNP panels can be expanded later if needed





Magliolo et al. 2021

Considerations

- Targeted SNP panels address specific applications
  - may not support broader analyses
- Reduced representation
  - some genetic signals may be lost (e.g., selection, subtle structure)
- A new measure of genetic diversity
  - Not comparable to genome-wide diversity

Requirements:

- Existing genome-wide SNP data
- Good geographic coverage

When to use SNP Panels

✅ Use SNP panels when:

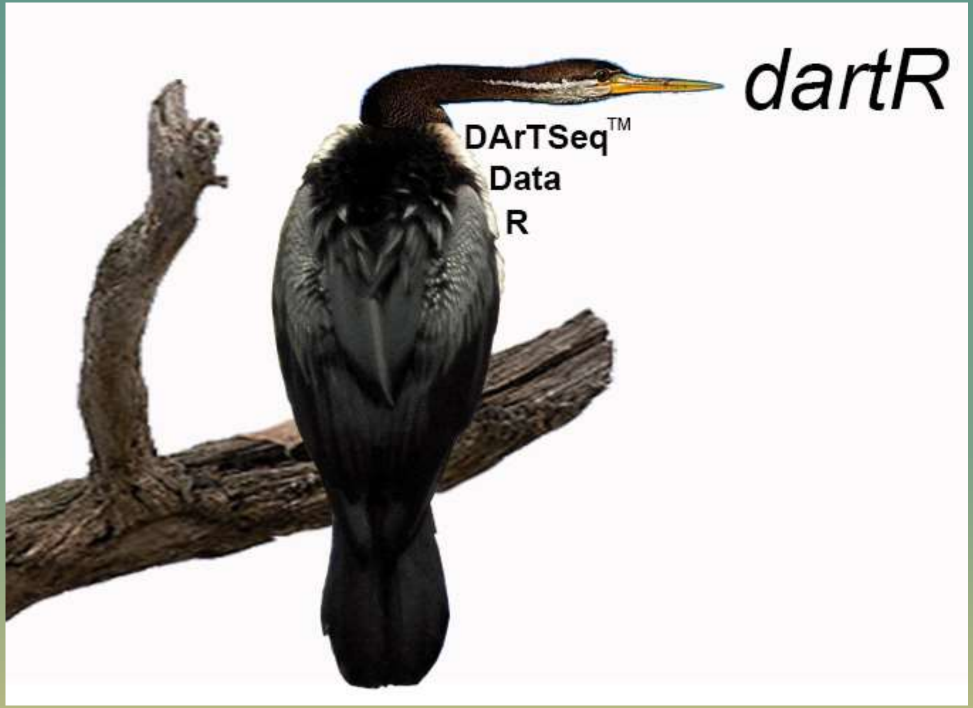
- monitoring large numbers of individuals,
- long-term surveillance,
- DNA samples are low quality or degraded (e.g., scats, feathers, eDNA)

❌ Avoid SNP panels when:

- sample sizes are small,
- genome-wide resolution is required (e.g., adaptation studies)
- genetically distant populations will be targeted

Using dartR for SNP panel selection

- Purpose: To select a subset of informative SNPs
- Versatility
  - Modify the *number* of SNPs
  - Find the best panel to address one *specific* conservation question or *multiple* conservation questions

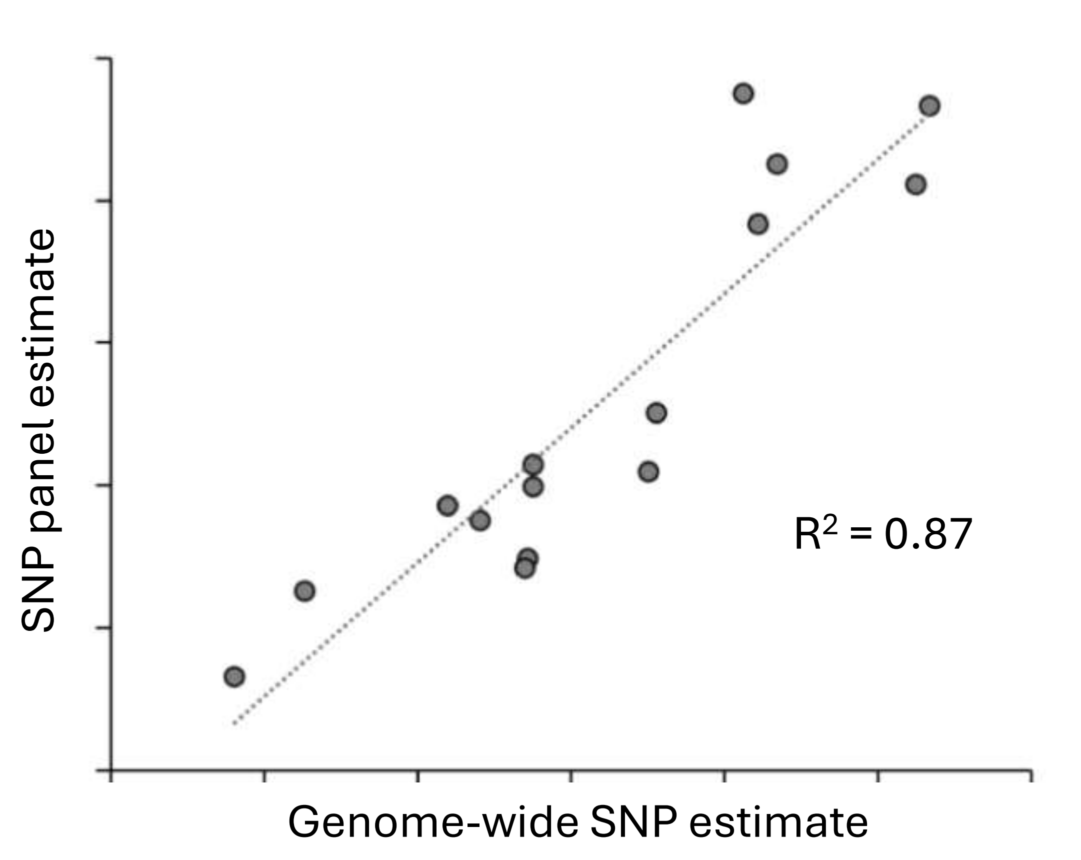


Key metrics for SNP selection

- Population structure, population assignment
  - **dapc**: Select loci contributing most to discrimination between populations using DAPC (Discriminant Analysis of Principal Components).
  - **pahigh**: Select loci with private alleles having high frequency i.e., diagnostic
  - **monopop**: Select monomorphic loci within populations i.e., fixed
- Individual-level resolution e.g., individual IDs, parentage, relatedness
  - **PIC**. Select loci with high Polymorphic Information Content i.e., high minor allele frequency.
  - **PICdart**. Similar to PIC but based on allele presence/absence rather than frequencies.
- Heterozygosity estimates e.g., diversity, inbreeding
  - **hafall**: Select loci with the highest minor allele frequencies across all populations. These are likely to be more polymorphic and informative across all populations.
  - **hafpop**: Select loci with the highest minor allele frequencies within each population. Increases within-population informativeness e.g., within-population diversity
- Genome-wide diversity
  - **random**: Randomly select loci. Provides an unbiased snapshot of diversity across the genome
  - **stratified**: Stratified sampling of loci based on allele frequencies. Similar to random but ensures broad coverage of genetic variation

Evaluate panel performance

- The final panel can be checked for concordance with:
  - F_{ST} - genetic differentiation
  - F_{IS} - inbreeding
  - N_{ALL} - number of alleles
  - H_E – expected heterozygosity
  - H_O – observed heterozygosity
  - N_E – effective population size



Next steps

- Select sequence provider and prepare data according to their requirements
- Primer design
  - Bioinformatics to avoid primer interactions
- Lab testing
  - Identify over-amplified or under-amplified loci. Filter as necessary and retest

Genomic Services @ DArT

Started with DNA array-based methods but moved to using Next Generation **Sequencing (NGS) supported by DArTdb/LIMS application**

DArTseq - **leading genotyping by sequencing technology**

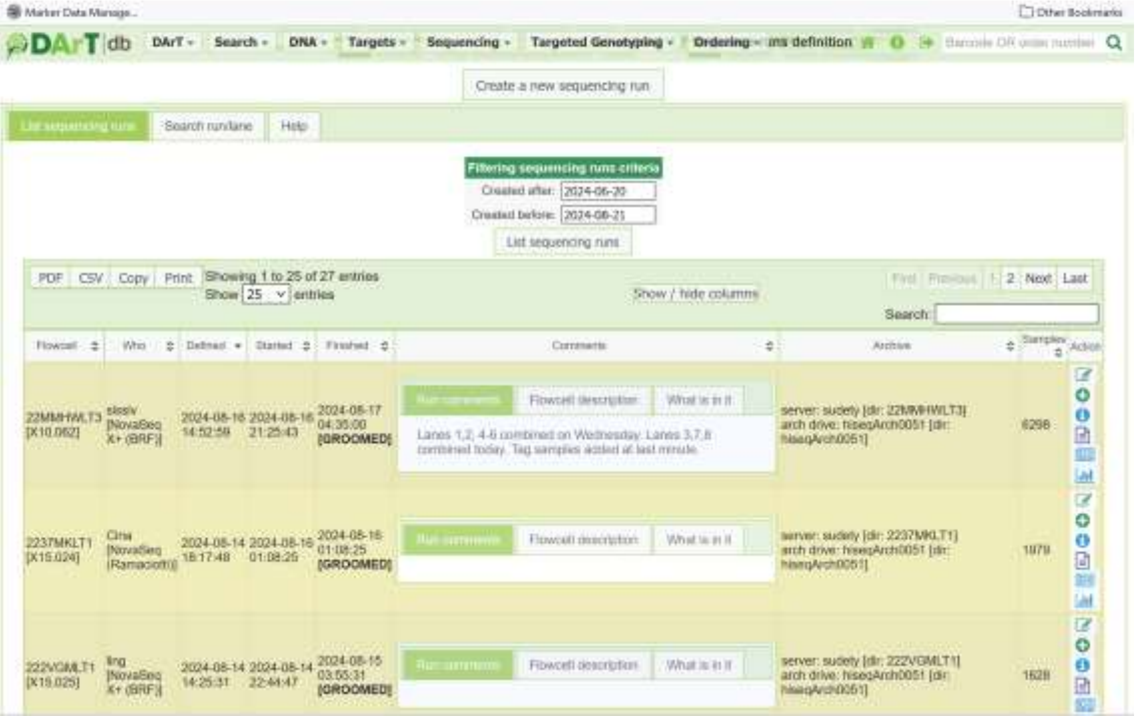
- Sequencing random genome fragments creating “genome representations”
- Highly scalable – adjust marker number
- High data quality
- Ability to resolve closely related material
- “De novo” and “SNP recall” analytical pipelines – reference free

Targeted Genotyping - amplicon sequencing

- **DArTag** – 300- 10,000 selected SNPs
    - Predominantly breeding tool, increasingly adopted in ecology
- **DArTcap**– 100- 10,000 selected SNPs used heavily in agriculture and in ecology
- **DArTmp** – up to 300 amplicons sequenced, used for genetic identification and paternity testing

DArTseqMet - DNA methylation analysis both at specific loci and genome-wide

DArTreseq and WG sequencing – gene cloning and pangenome construction



Filtering sequencing runs criteria

Created after: 2024-06-20

Created before: 2024-08-21

List sequencing runs

Showing 1 to 25 of 27 entries

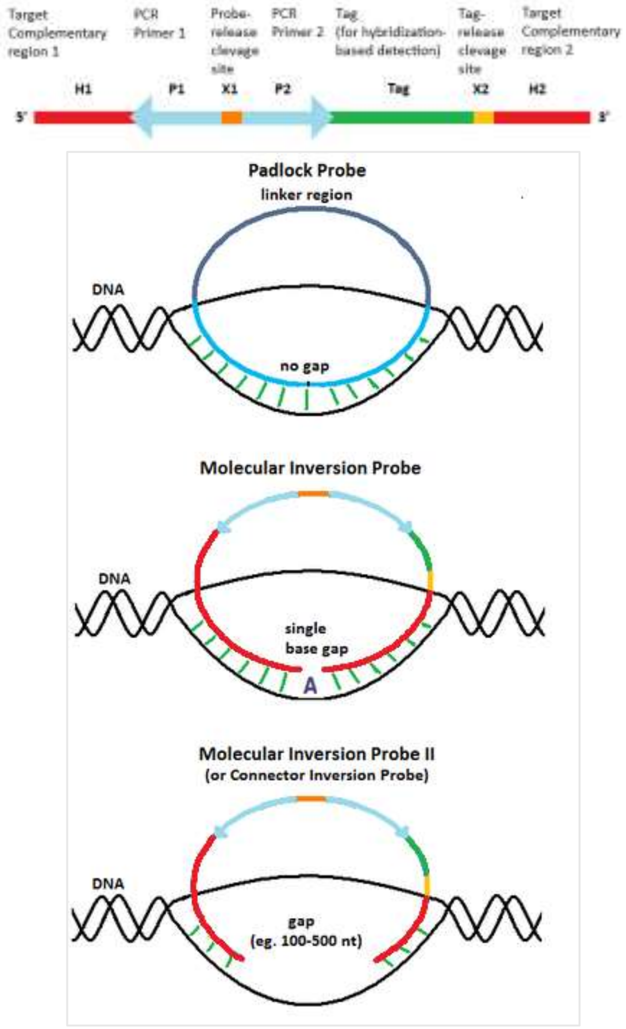
Show 25 entries

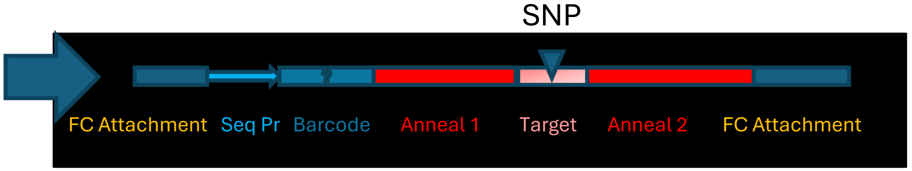
Search:

Flowcell	Who	Defined	Started	Finished	Comments	Archive	Samples	Action
22MMHWLT3 [X10,062]	slesiv [NovaSeq X+ (BRF)]	2024-08-16 14:52:59	2024-08-16 21:25:43	2024-08-17 04:35:00 **[GROOMED]**	Lanes 1,2, 4-6 combined on Wednesday. Lanes 3,7,8 combined today. Tag samples added at last minute.	server: sudety [dir: 22MMHWLT3] arch drive: hiseqArch0051 [dir: hiseqArch0051]	6298	
2237MKLT1 [X15,024]	Cina [NovaSeq (Ramaciotti)]	2024-08-14 18:17:48	2024-08-16 01:08:25	2024-08-16 01:08:25 **[GROOMED]**		server: sudety [dir: 2237MKLT1] arch drive: hiseqArch0051 [dir: hiseqArch0051]	1979	
222VGMLT1 [X15,025]	ling [NovaSeq X+ (BRF)]	2024-08-14 14:25:31	2024-08-14 22:44:47	2024-08-15 03:55:31 **[GROOMED]**		server: sudety [dir: 222VGMLT1] arch drive: hiseqArch0051 [dir: hiseqArch0051]	1628	

DArTag platform

- **Adopted padlock molecule concept to capture target region**
- **Three step process:**
  1. capture of region with SNP/INDEL via padlock oligos
  2. addition of sample barcode and flowcell attchement via PCR
  3. Sequencing of DArT libraries and marker data extraction
- **Dramatically simplified previous attempts at utilising padlocks in genotyping (MIPs)**
  - Eliminated some molecular “features”, several steps and some expensive enzymes
  - Flexible sequence capture window (mostly 70-110 bp range)
  - Moved assay to 384 plate format reducing assay volume and therefore the cost while increasing throughput
  - Tested scalability beyond 10,000 markers in a single assay

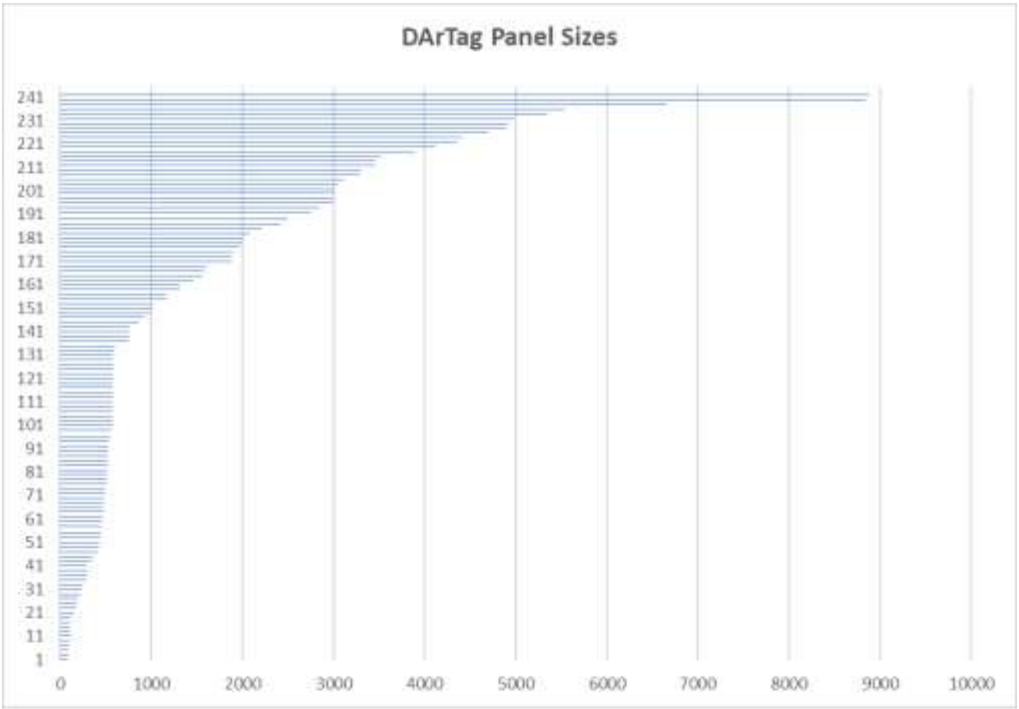

Target Complementary region 1
PCR Primer 1
Probe-release clevage site
PCR Primer 2
Tag (for hybridization-based detection)
Tag-release clevage site
Target Complementary region 2
H1
P1
X1
P2
Tag
X2
H2
5'
3'
Padlock Probe
linker region
DNA
no gap
Molecular Inversion Probe
DNA
single base gap
A
Molecular Inversion Probe II
(or Connector Inversion Probe)
DNA
gap
(eg. 100-500 nt)



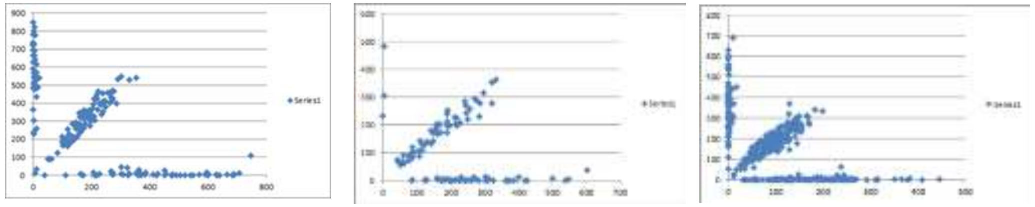
DArTag molecule final structure

Technical performance

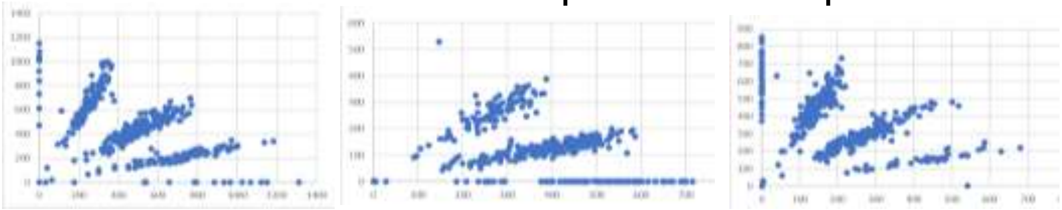
- **Conversion rate ~97% even** in species with limited sequencing resources
- **Technology works well from a few markers to over 10,000**
- **Outperformed other technologies on the market, including in species with polysomic inheritence (potato, blueberry, alfalfa…)**
- **Typical call rate: >98%**
- **Calling reproducibility: >99.95%**
- Average marker read depth ~100 X for diploids and 200- 500 X for polysomic species
- Cost dependent on the number of markers and required sequencing depth: -----> application!
- Main use at the moment in Genomic Selection of crops and animals
- In Ecology mostly large volume monitoring/Close Kin Mark Recapture applications



Distribution of counts in discomic species

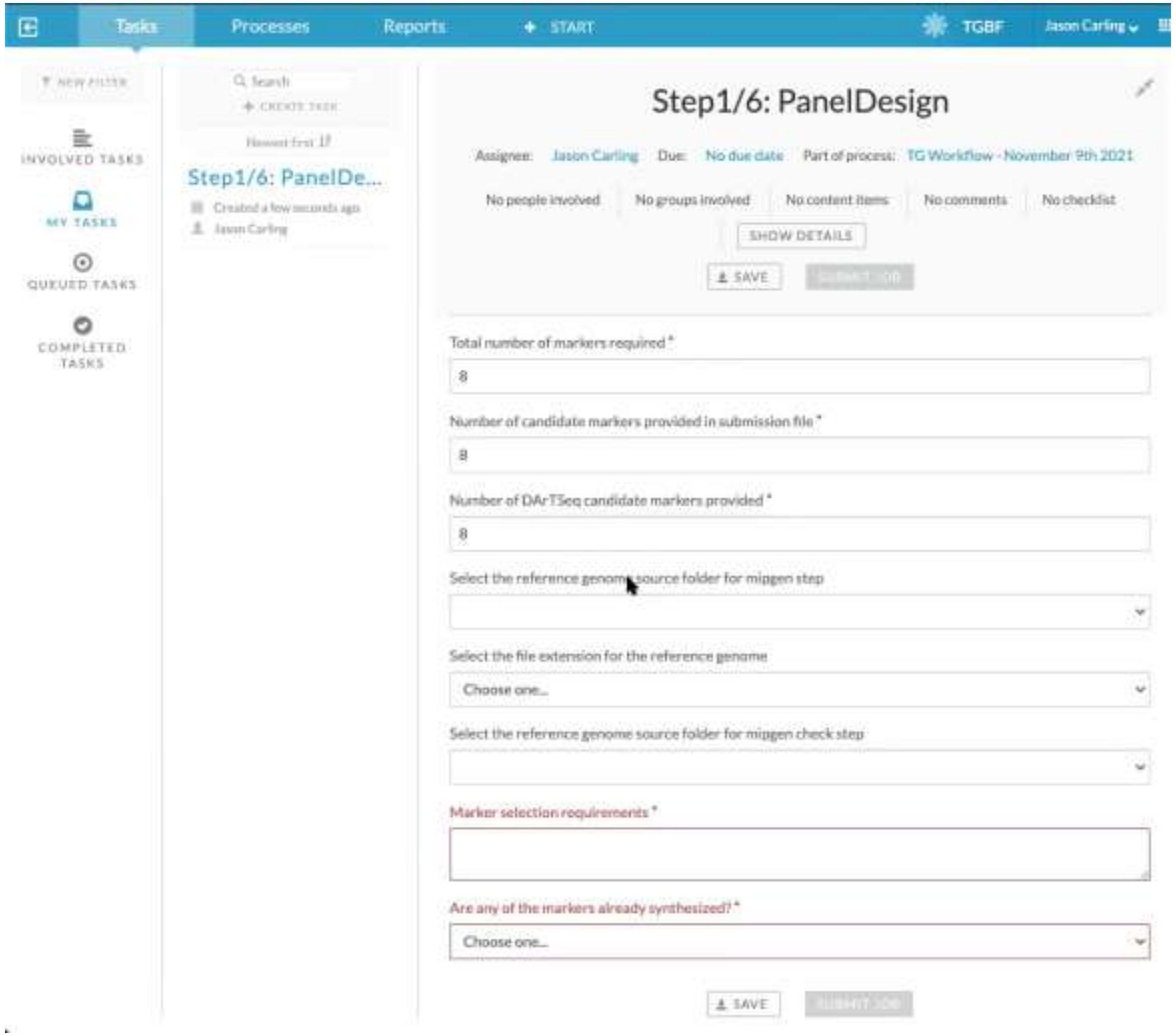


Distribution of counts in poliscomic species



Panel design and $ considerations

- Fully automated panel design process when the submission file is formatted and filled properly
- Detailed description of data and format requirements downloadable from https://www.diversityarrays.com/services/targeted-genotyping/
- Over 250 panels established since technology launched in 2015
- Median panel size: 577
- Average panel size: 1640
- Reference genome and marker data quality very important for design success
- DArTag outperforms other technologies in $ in medium-to-large scale applications
- Panel development cost depended on service volume (synthesis scale of oligos)
  - Cost per marker between $5-$15 for 20 K - 2 M assays
- Pricing strictly "per plate" as cost the same for full (94 samples) and partial plates
- Genotyping costs between $750/plate (small panels, very large service volume) to $3,000 (large panels, small volume)



Step1/6: PanelDe...

Created a few seconds ago

Jason Carling

Step1/6: PanelDesign

Assignee: Jason Carling Due: No due date Part of process: TG Workflow - November 9th 2021

No people involved

No groups involved

No content items

No comments

No checklist

Total number of markers required *

8

Number of candidate markers provided in submission file *

8

Number of DArTSeq candidate markers provided *

8

Select the reference genome source folder for mipgen step

Select the file extension for the reference genome

Choose one...

Select the reference genome source folder for mipgen check step

Marker selection requirements *

Are any of the markers already synthesized? *

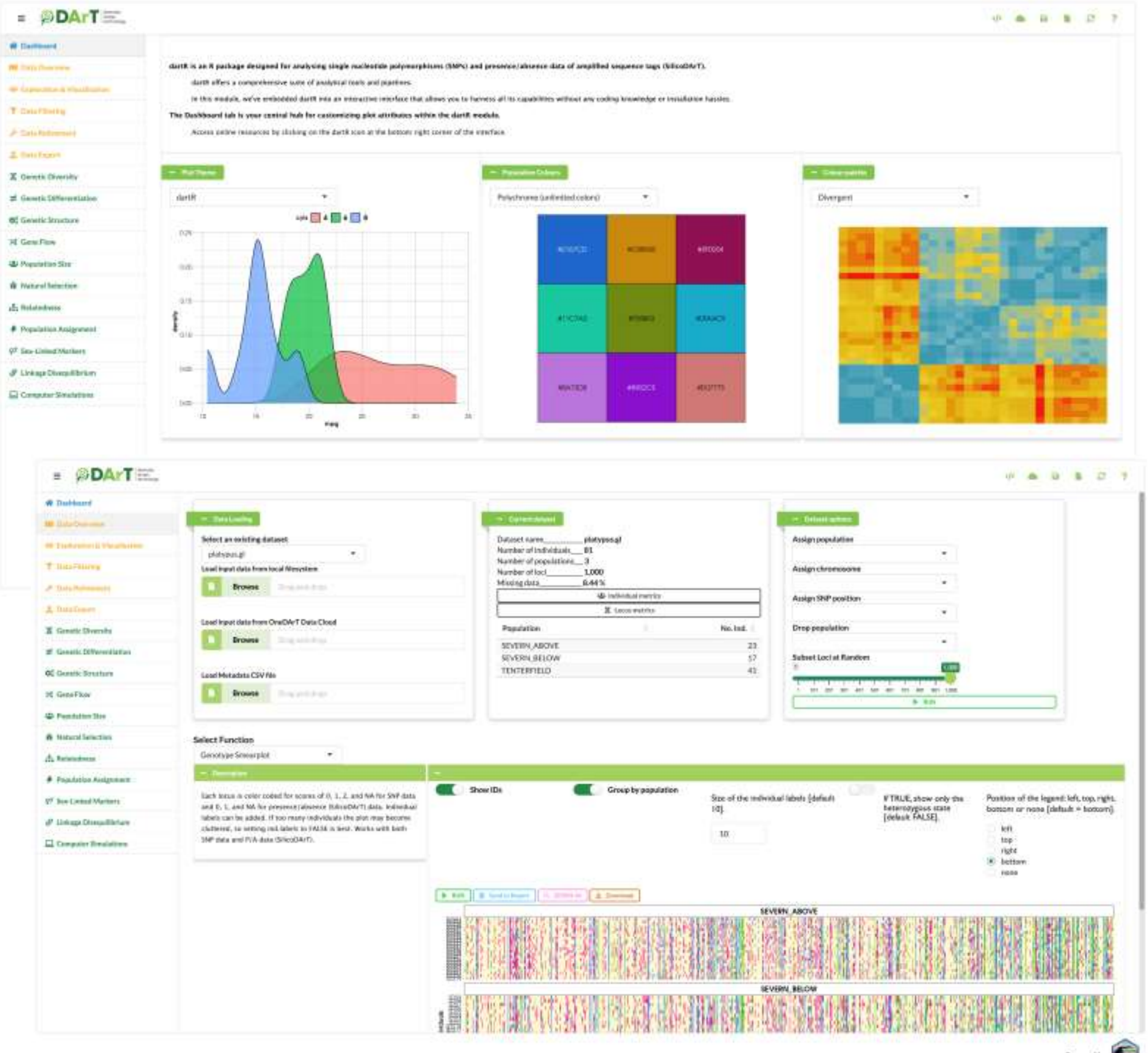
Choose one...

Automated pipeline for panel design consisting of 6 plugins, with dedicated application (TagGen) for oligonucleotides (padlocks) developed @DArT

OneDArT release in July!

dartR integration in Analytics+

- ✓ Partnership between DArT and Australian academics
- ✓ Extending user base of dartR
- ✓ Expanding from ecology-focused application to more general utility
- ✓ A broad range of analytical functions including complex modelling
- ✓ Accessible in OneDArT for people with no skill (or interest) in using R
- ✓ Providing expandable compute resources at low cost
- ✓ Bringing genomic and environmental data together with mobile app collected sample metadata



dartR is an R package designed for analysing single nucleotide polymorphisms (SNPs) and presence/absence data of amplified sequence tags (SilicoDArT).

dartR offers a comprehensive suite of analytical tools and pipelines.

In this module, we've embedded dartR into an interactive interface that allows you to harness all its capabilities without any coding knowledge or installation hassles.

The Dashboard tab is your central hub for customizing plot attributes within the dartR module.

Access online resources by clicking on the dartR icon at the bottom right corner of the interface.

Select an existing dataset: platypus.gl

Load input data from local filesystem

Load input data from OneDArT Data Cloud

Load Metadata CSV file

Dataset name__________ platypus.gl
Number of individuals___ 81
Number of populations___ 3
Number of loci__________ 1,000
Missing data__________ 6.44 %

Population	No. Ind.
SEVERN_ABOVE	23
SEVERN_BELOW	17
TENTERFIELD	41

Assign population

Assign chromosome

Assign SNP position

Drop population

Subset Loci at Random

Select Function: Genotype Smearplot

Each locus is color coded for scores of 0, 1, 2, and NA for SNP data and 0, 1, and NA for presence/absence (SilicoDArT) data. Individual labels can be added. If too many individuals the plot may become cluttered, so setting ind.labels to FALSE is best. Works with both SNP data and P/A data (SilicoDArT).

Show IDs

Group by population

Size of the individual labels [default 10]. 10

If TRUE, show only the heterozygous state [default FALSE].

Position of the legend: left, top, right, bottom or none [default = bottom].
○ left
○ top
○ right
● bottom
○ none

BUSH HERITAGE
AUSTRALIA

Example - Redfin blue eye

Australian Government
Australian Research Council

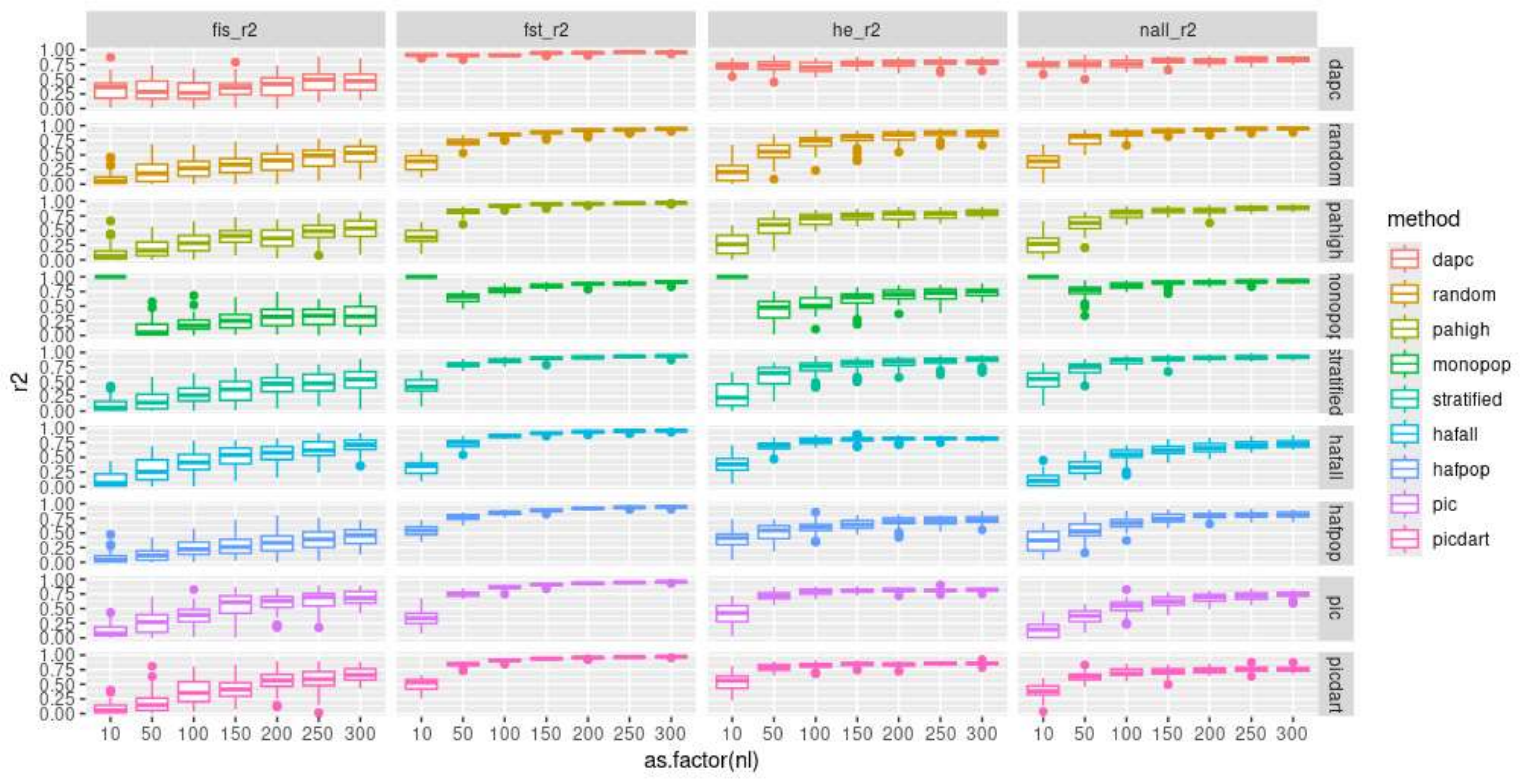
SNP panel development required for species monitoring from non-destructive, trace DNA samples

Example - Redfin blue eye

- Decide on aim and number of SNPs in the panel

- Filter for quality of SNPs
- Filter for sequence quality and suitability
- Select and check panel

Example - Redfin blue eye


fis_r2
fst_r2
he_r2
nall_r2
dapc
random
pahigh
monopop
stratified
hafall
hafpop
pic
picdart
r2
as.factor(nl)
method
dapc
random
pahigh
monopop
stratified
hafall
hafpop
pic
picdart