TAXONOMIC AND FUNCTIONAL ANALYSIS OF DE NOVO TRANSCRIPTOMES WITH TRAPID 2.0

François Bucchini [1,2], Andrea Del Cortona [1,2], Michiel Van Bel [1,2] and Klaas Vandepoele [1,2]

francois.bucchini@psb.vib-ugent.be

1. Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium
2. VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium

BACKGROUND

- Recent technological advances in sequencing have made it possible to take a **snapshot of gene expression** in a specific tissue, condition, unicellular organism, or community.
- The explosion of transcriptome RNA-Seq data comes at the cost of new challenges, as **reference genomes are rarely available**.
- In the absence of genome sequences, ***de novo* assembled transcriptomes** represent a basis for investigating the **gene repertoire** of previously uncharacterized organisms.
- *De novo* transcriptomes are however **challenging to analyze** and interpret. They often contain fragmented, spurious or contaminant sequences.
- To mitigate some of these challenges, we developed TRAPID 2.0, a **web application** for the fast and efficient **processing of assembled transcriptome data**.

TRAPID 2.0 WORKFLOW OVERVIEW

TRAPID 2.0's workflow (Fig. 1) consists of **two distinct phases**: an **initial processing phase**, and an **exploratory phase** that enable users to perform functional & comparative analyses interactively from the web application.

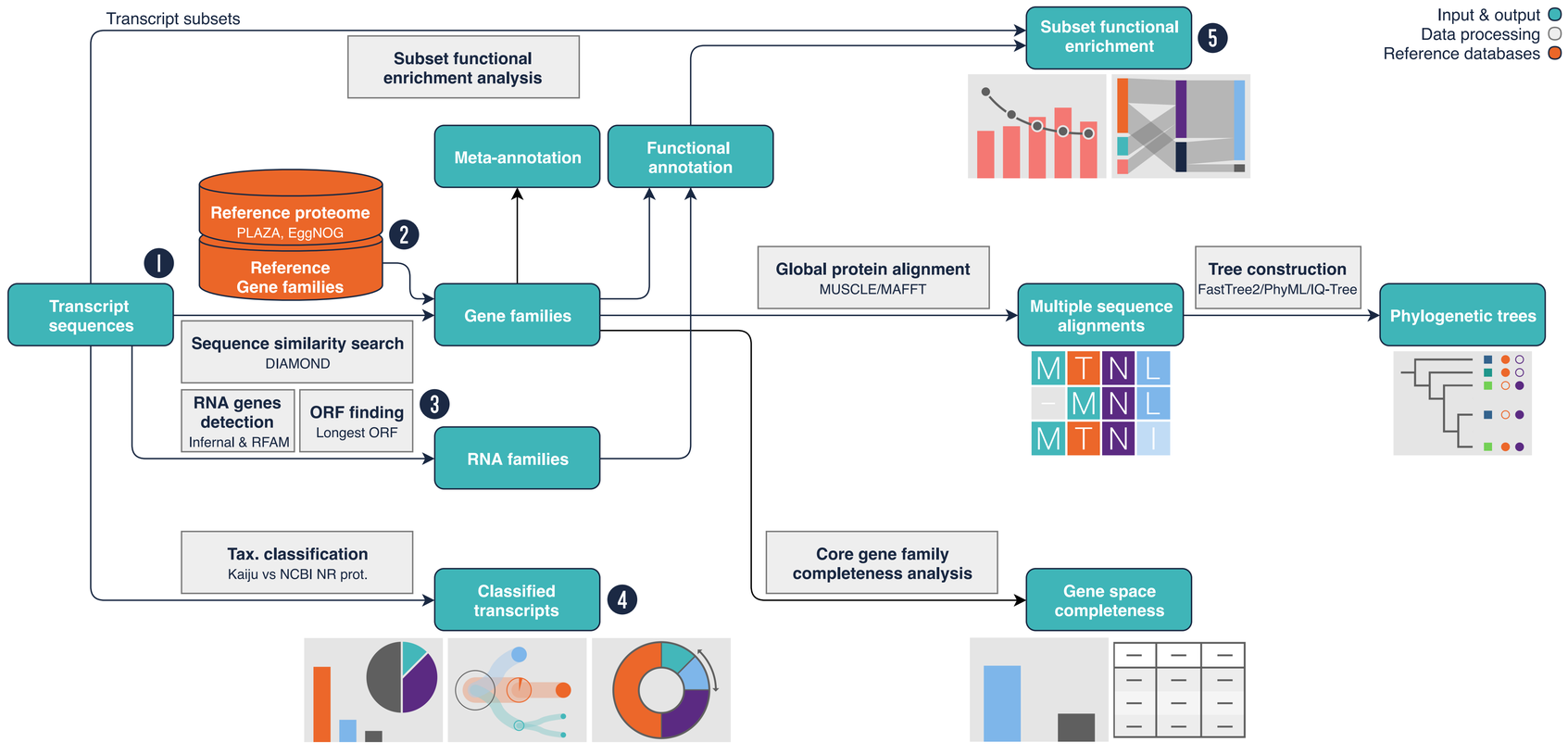

Fig. 1: Overview of the TRAPID 2.0 workflow & functionalities. Thumbnails depict caricatures of the available visualizations, & elements marked with circled numbers are further detailed in dedicated sections.

1 INPUT DATA AND METATRANSCRIPTOME PROCESSING

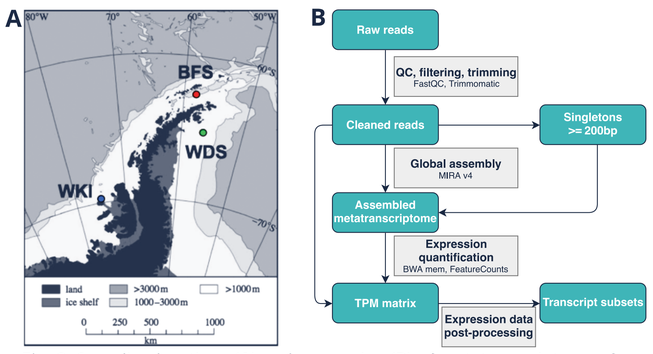

Fig. 2: Sampling locations (**A**) and processing (**B**) of metatranscriptomes from diatom-dominated communities. BFS: Bransfield Strait; WDS: Weddell Sea; WKI: Wilkins Ice Shelf. Panel A adapted from Pearson et al., 2015.

- TRAPID 2.0 takes any set of **assembled transcripts** as input.
- To demonstrate its efficiency in extracting biological knowledge from **metatranscriptomics data**, we used TRAPID 2.0 to study functional variations in **diatom-dominated phytoplankton communities** from the Antarctic peninsula (Fig. 2, data from Pearson et al. 2015).

2 TRAPID 2.0 REFERENCE DATABASES

- Collections of **functionally annotated sequences** from multiple species, clustered in precomputed **gene families** (GFs).
- Reference databases (Table 1): **broad phylogenetic range**, high-quality backbone for the comparative genomics features of TRAPID 2.0.

Table 1: Overview of TRAPID 2.0 reference databases. The gene family count only includes homology-based for PLAZA databases, and only orthologous groups at the root level for EggNOG 4.5.

	PLAZA 4.0 dicots	PLAZA 4.0 monocots	Pico-PLAZA 2.0	EggNOG 4.5
# Species	55	29	19	2,031
# Genes	3,065,012	1,056,271	302,559	14,116,949
# GFs	208,456	154,839	68,827	190,803
Taxonomic focus	Dicot plants	Monocot plants	Photosynthetic microeukaryotes	Archeae, Bacteria, Eukaryota
Funct. annotation	GO, InterPro	GO, InterPro	GO, InterPro	GO, KO
GF construction	Tribe-MCL	Tribe-MCL	Tribe-MCL	EggNOG

3 ORF FINDING: NON-CANONICAL GENETIC CODE SUPPORT

- **Homology-supported** ORF sequence detection using non-canonical genetic code.
- **Impact of appropriate genetic code** use confirmed by processing 16 ciliate MMETSP transcriptomes with TRAPID 2.0 (Fig. 3).

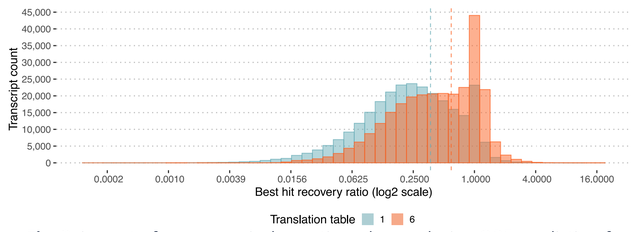

Fig. 3: impact of non-canonical genetic code use during ORF prediction for 272,214 sequences from 16 ciliate transcriptomes.

4 TAXONOMIC CLASSIFICATION OF TRANSCRIPTS

- Purpose: **flagging of potential contaminants**, examination of the **taxonomic composition** of complex samples. Supported by interactive vizualisations (Fig. 4).
- Performed using **Kaiju**, a tool particularly adapted to classify sequences of organisms from phylogenetic clades that are under-represented in databases.

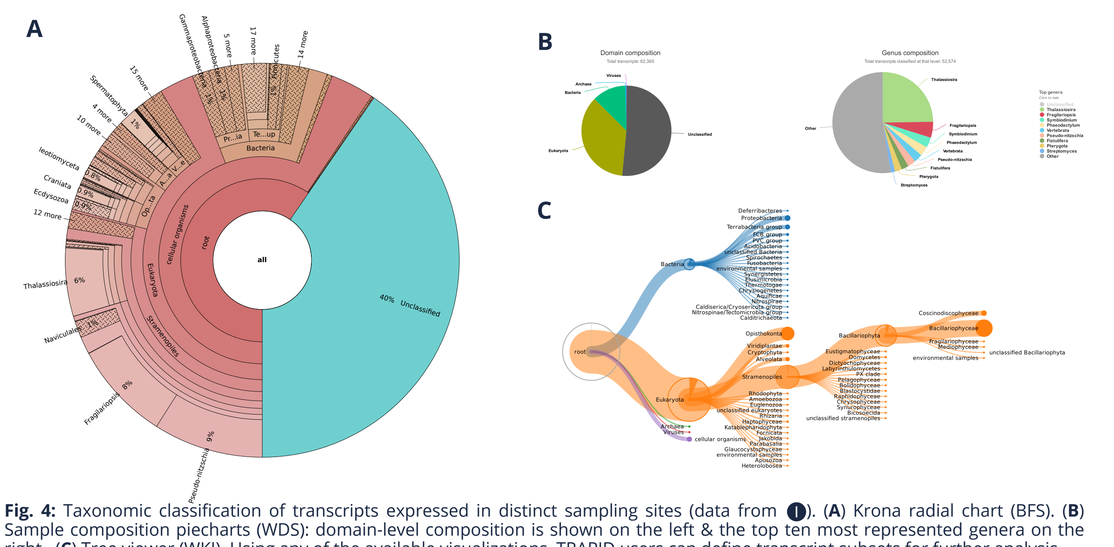

Fig. 4: Taxonomic classification of transcripts expressed in distinct sampling sites (data from ❶). (**A**) Krona radial chart (BFS). (**B**) Sample composition piecharts (WDS): domain-level composition is shown on the left & the top ten most represented genera on the right. (**C**) Tree viewer (WKI). Using any of the available visualizations, TRAPID users can define transcript subsets for further analysis.

5 TRANSCRIPT SUBSETS ANALYSIS AND COMPARISON

- The **analysis of transcript subsets** can provide additional biological insights.
- Available analyses: exploration of the relationships between subsets, functional annotations, and GFs (Fig. 2A); functional enrichment (Fig. 2B); and subset functional annotation comparison.

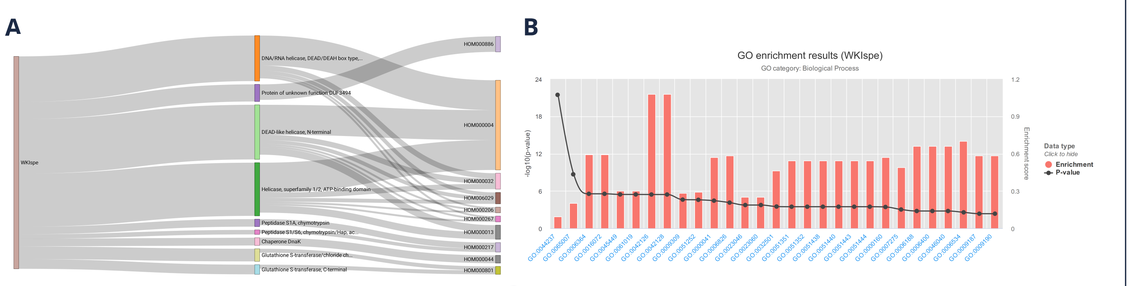

Fig. 5: Analysis of 49,998 WKI-specific transcripts (data from ❶). (**A**) Sankey diagram depicting the relationships between WKI-specific transcripts (left blocks), significantly enriched IPR domains (middle blocks) and Pico-PLAZA GFs (right blocks). Line width is proportional to transcript annotation (left lines) and GF membership (right lines). (**B**) WKI-specific transcripts GO enrichment results. GO terms are represented on the x-axis, enrichment p-value on the left y-axis (black dots), and enrichment score on right y-axis (red bars). Maximum enrichment p-value threshold is 1E-3 and only biological process GO terms are displayed.

TRAPID 2.0 IS AVAILABLE AT BIOINFORMATICS.PSB.UGENT.BE/TRAPID_02

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY
GHENT UNIVERSITY
COMPARATIVE NETWORK BIOLOGY
SINGEK training network
MARIE SKŁODOWSKA-CURIE ACTIONS

SCIENCE MEETS LIFE