Appendix – supplementary figures and tables

Tools

KNIME4NGS tools

KNIME4NGS depends on a number of tools that need to be installed in the system. The workflows we devised and ran in a VirtualBox environment (see main manuscript) had the following versions of the tools required by KNIME4NGS installed.
Bowtie2: 2.1.0 (32bit)
BWA: 0.7.17 (32bit)
FastQC: 0.11.7 (32bit)
featureCounts: 1.6.0 (32bit)
R: 3.2.3 (32bit)
R Deseq: (32bit) - Installed via BioConductor
samtools: 1.7 (32bit)
subread: 1.6.0 (32bit)

Galaxy 18.01 tools

Galaxy workflows like KNIME workflows depend on tools which may need to be installed. The workflows we devised used the tools at the given revisions (Table S1).

Table S1: Installed Galaxy tools. Rows indicate tools used in Galaxy to construct workflows. For each tool, the repository from which the tool was obtained, the revision and version are noted in separate columns.

	Repository	Revision	Version
Bowtie2	devteam	c3dd1aeb7d07	2.3.4
BWA	devteam	8d2a528a9513	0.7.17
DESeq2	iuc	d0c39b5e78cf	2.11.40.1
FastQC	devteam	ff9530579d1f	0.11.6
FeatureCounts	iuc	386220cf6877	1.6.0.5
Kallisto pseudo	iuc	1c75aa5de15e	0.43.1
Kallisto quant	iuc	b818b23df1e0	0.43.1
Sickle	iuc	3905ccd5c631	1.33.1
Trimmomatic	pjbriggs	dfa082f84068	0.36.2

Differential expression

Count data

Count data produced by the differential expression from transcriptome mapping workflows of each WFMS, were transformed using log2 and subsequently plotted as box and whisker plots (Figure S1). CA and CB represent the two conditions (conditions A and B). The replicates for each condition are represented by R1 through R4. The Galaxy workflow used kallisto for mapping and quantification, which produces estimated counts. Some of these count values are between 0 and 1, resulting in negative log2 values as seen in Figure S3. Since counts varied among tools, the box and whisker plots were used to decide on cutoff values for filtering of transcripts with low evidence (see next section).

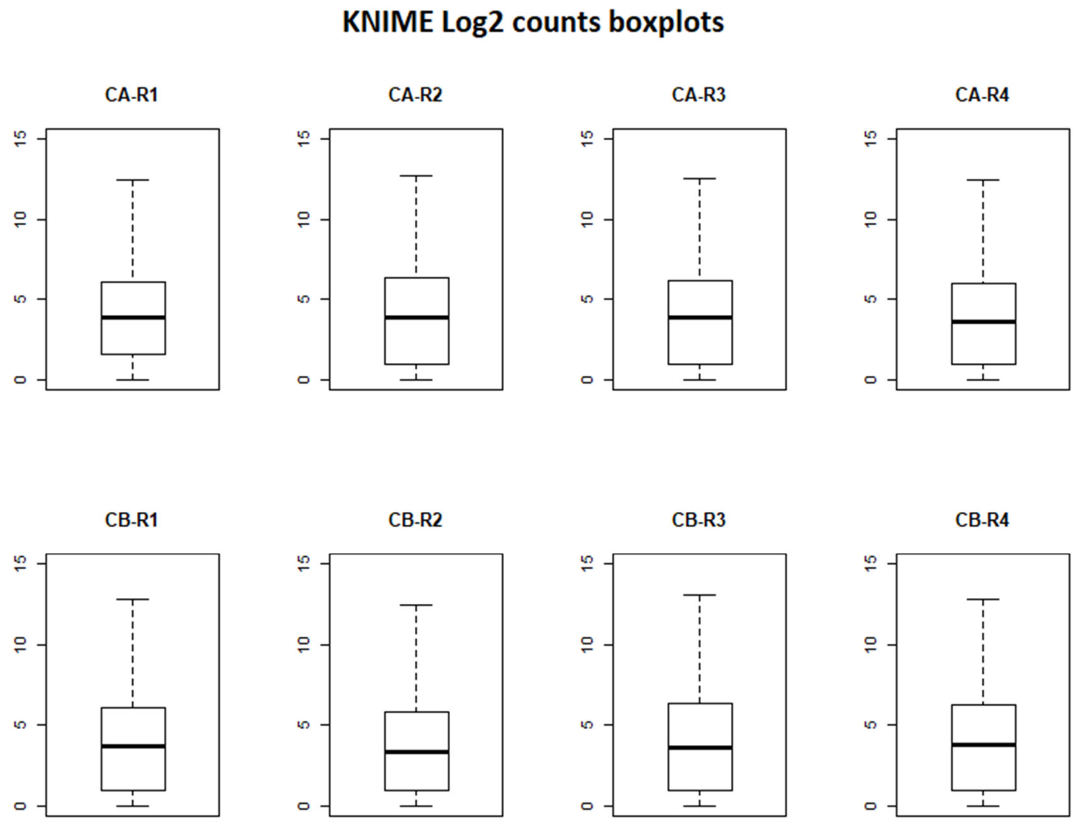


Figure S1: Boxplots of KNIME log2 transformed count data. The top row represents four replicate paired end read datasets (R1–R4) for the first condition (CA). The bottom row represents four replicate paired end read datasets (R1–R4) for the second condition (CB).

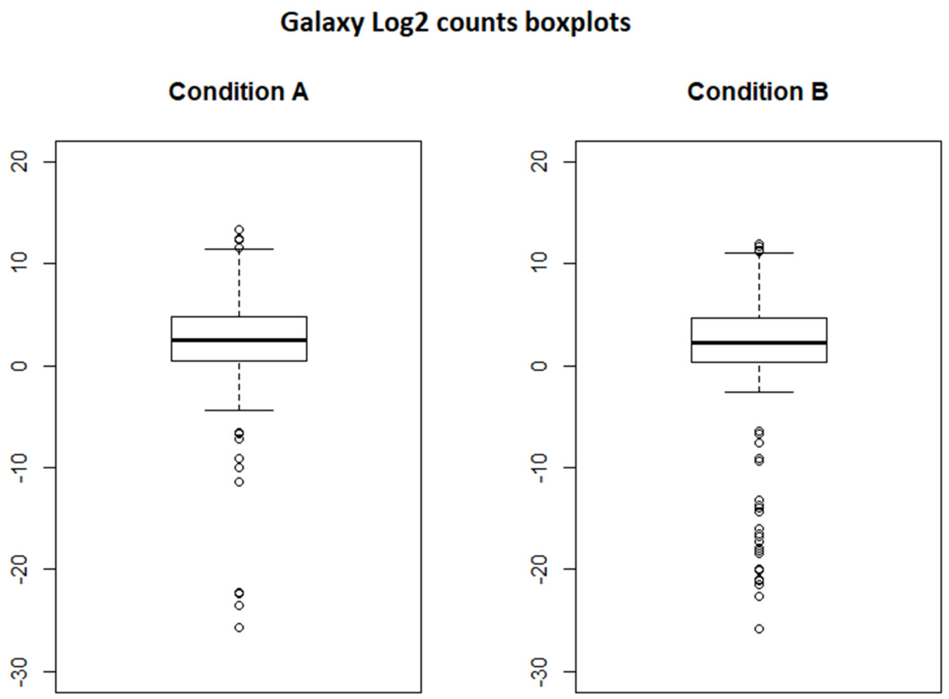


Figure S2: Boxplots of Galaxy log2 transformed count data. Each boxplot represents the count data.

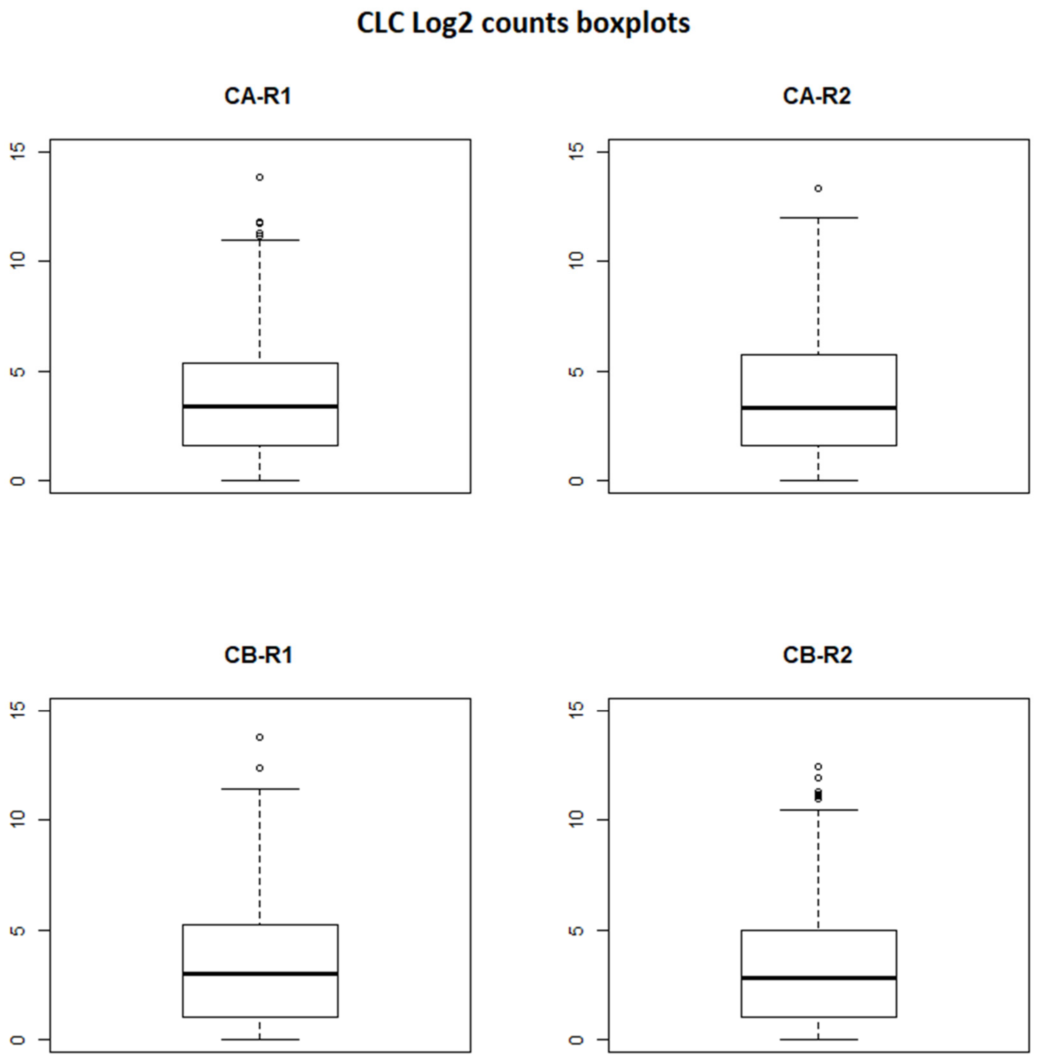


Figure S3: Boxplots of CLC log2 transformed count data. The top row represents counts of the two replicates for the first condition (condition A). The bottom row represents counts of the two replicates for the second condition (condition B). The two replicate datasets are represented by R1 and R2.

Cut off values

Count data was filtered with several cutoff values selected by different percentiles of the count data (see box and whisker plots above and Table S2). The overlap of identified transcripts among tools changes with the cutoff values (Figure S4). The values in the table below are the log2 cutoffs derived from the boxplots (Table S2). CA and CB represent the two conditions (conditions A and B). Replicates for each condition are represented by R1 through R4.

Table S2: KNIME boxplot cutoff values. Each row represents the specific cutoff values for datasets for a specific percentile. CA and CB represent the two conditions and R1 to R4 the replicate data sets. Specific cutoff values in this table are log2 transformed values. Most replicates have the same cutoff value for a given percentile.

Cutoff	CA-R1	CA-R2	CA-R3	CA-R4	CB-R1	CB-R2	CB-R3	CB-R4
>10th percentile	0	0	0	0	0	0	0	0
>25th percentile	1	1	1	1	1	1	1	1
>50th percentile	4	4	4	3	3	3	3	3
>75th percentile	6	6	6	6	6	6	6	6
>90th percentile	12	13	12	12	13	12	13	13

Table S3: Galaxy boxplot cutoff values. Count datasets from four replicates were already combined by the Galaxy workflow. Columns represent the specific cutoff values for each condition. Specific cutoff values are log2 transformed values. The negative values represent low count values between 0 and 1. Both conditions have the same cutoff values for each percentile.

Cutoff	Condition A	Condition B
>10th percentile	–5	–5
>25th percentile	0	0
>50th percentile	2	2
>75th percentile	4	4
>90th percentile	12	12

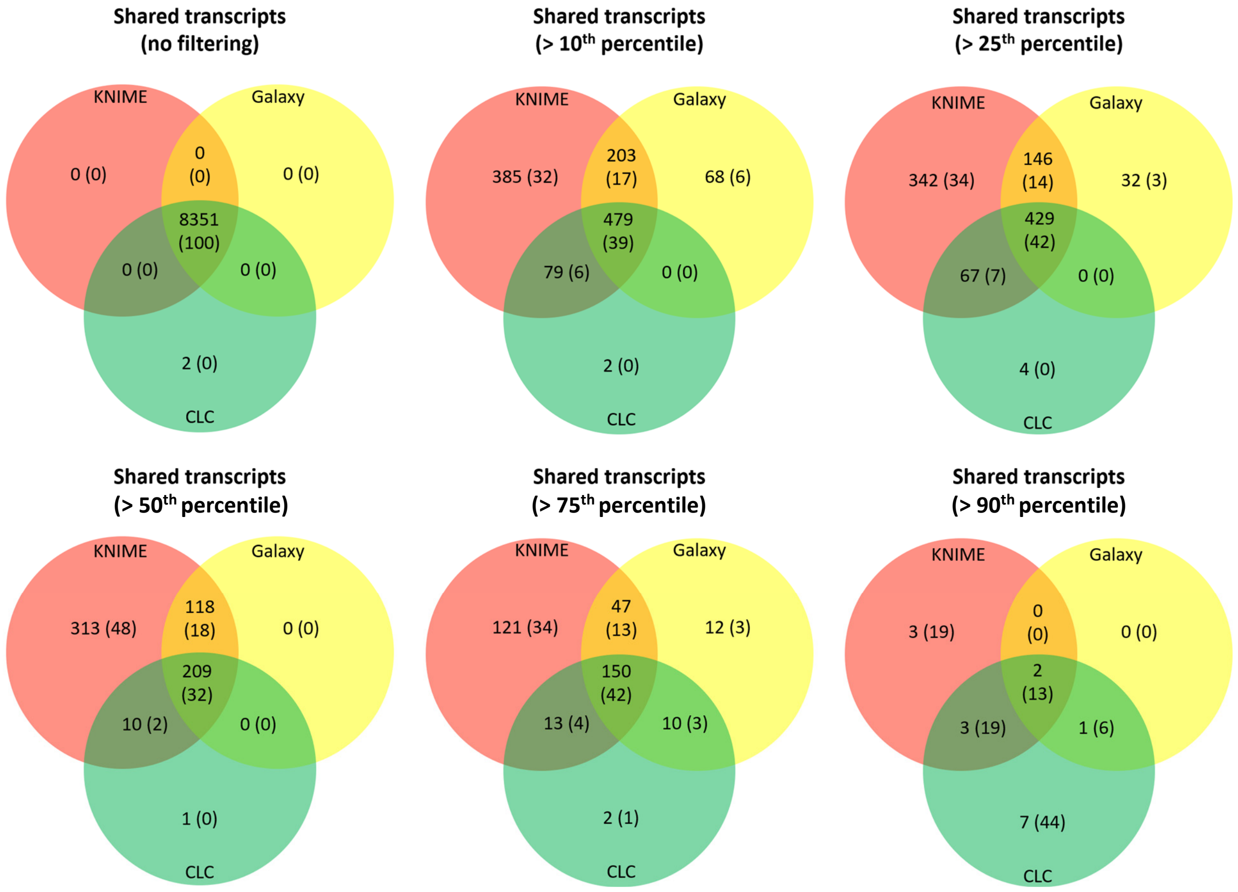
Table S4: CLC boxplot cutoff values. Each row represents the specific cutoff values for each percentile. CA and CB represent the two conditions, R1 and R2 represent the two replicate data sets. Specific cutoff values in this table are log2 transformed values. Specific cutoff values are the same for most percentiles except the 90th percentile.

Cutoff	CA-R1	CA-R2	CB-R1	CB-R2
>10th percentile	0	0	0	0
>25th percentile	1	1	1	1
>50th percentile	3	3	3	3
>75th percentile	5	5	5	5
>90th percentile	11	12	11	10

Shared transcripts

Venn diagrams

The Venn diagrams below were constructed using the list of transcripts remaining after filtering the count data. Count datasets for each condition and replicate, in the case of KNIME and CLC, were filtered using the appropriate cutoff values (Tables S3, S4, S5), resulting in a list of transcripts. The transcripts lists for each WFMS were combined by performing unions. These combined lists were subsequently used to construct the Venn diagrams depicted in Figure S4.



Shared transcripts (no filtering)

Shared transcripts (> 10th percentile)

Shared transcripts (> 25th percentile)

Shared transcripts (> 50th percentile)

Shared transcripts (> 75th percentile)

Shared transcripts (> 90th percentile)

Figure S4: Venn diagrams of shared transcripts. The Venn diagrams, arranged from left to right and top to bottom, display the intersecting and unique transcripts between each WFMS for the selected cutoff values. The percentage of the overall predictions are given in parentheses. Galaxy and CLC share most transcripts with either one or both other WFMS. KNIME shows the most transcripts not shared with either Galaxy or CLC except for at >10th percentile.

Table S5: Unions of transcripts between WFMS. Rows indicate several cutoffs used to filter transcripts. Cutoff values were based on boxplots created from log2 transformed count data. Columns represent the combined number of unique transcripts. Union with KNIME generally results in more transcripts, likely due to the fact that KNIME in general has more transcripts (see Figure S4).

	KNIME	Galaxy	CLC	KNIME ∪ Galaxy	KNIME ∪ CLC	Galaxy ∪ CLC	KNIME ∪ Galaxy ∪ CLC
No filtering	8,351	8,351	8,351	8,351	8,351	8,351	8,351
>10th percentile	1,146	750	560	1,214	1148	831	1,216
>25th percentile	984	607	500	1,016	988	678	1,020
>50th percentile	650	219	328	650	651	338	651
>75th percentile	331	219	175	353	343	234	355
>90th percentile	8	3	13	9	16	13	16