Supplementary Figures and Table for “A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data”

Authors :

Paul Simion[1,3], Khalid Belkhir[1], Clémentine François[1], Julien Veyssier[1], Jochen C. Rink[2], Michaël Manuel[3], Hervé Philippe[4,5], Maximilian J. Telford[6]

Affiliations :

1 Institut des Sciences de l'Evolution (ISEM), UMR 5554, CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

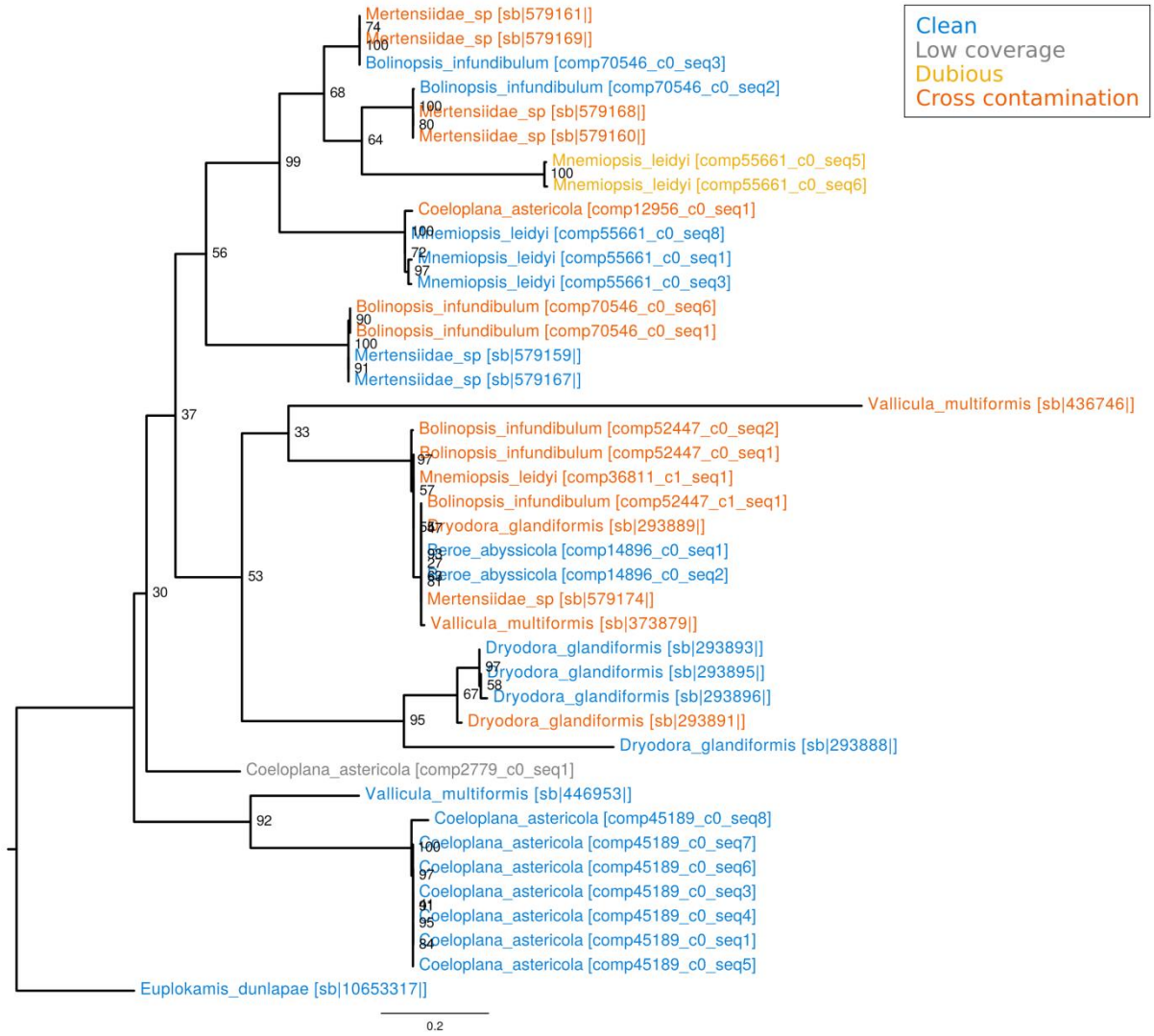
2 Max Plank Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany

3 Sorbonne Universités, UPMC Univ Paris 06, CNRS, Evolution Paris-Seine UMR7138, Institut de Biologie Paris-Seine, Case 05, 7 quai St Bernard, 75005 Paris, France

4 Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321, Moulis, 09200, France

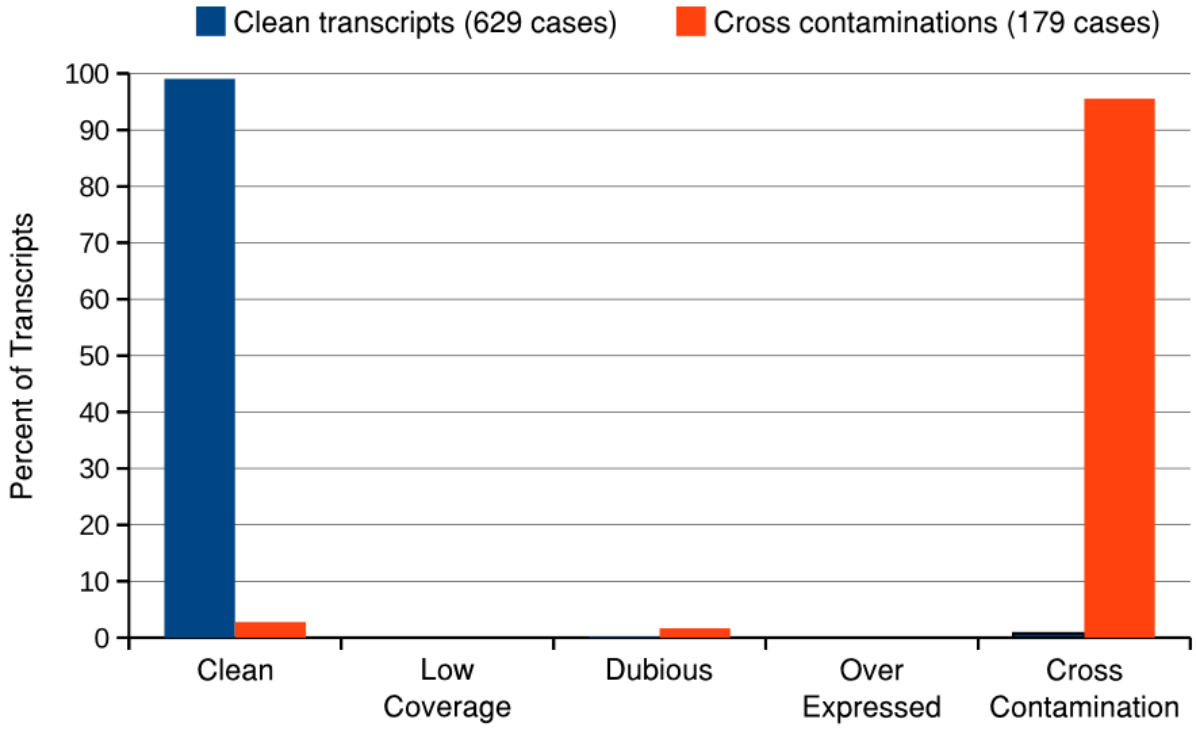
5 Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, H3C 3J7 Québec, Canada

6 University College London, Centre for Life’s Origins and Evolution, Department of Genetics, Evolution and Environment, Darwin Building, Gower Street, London WC1E 6BT, UK



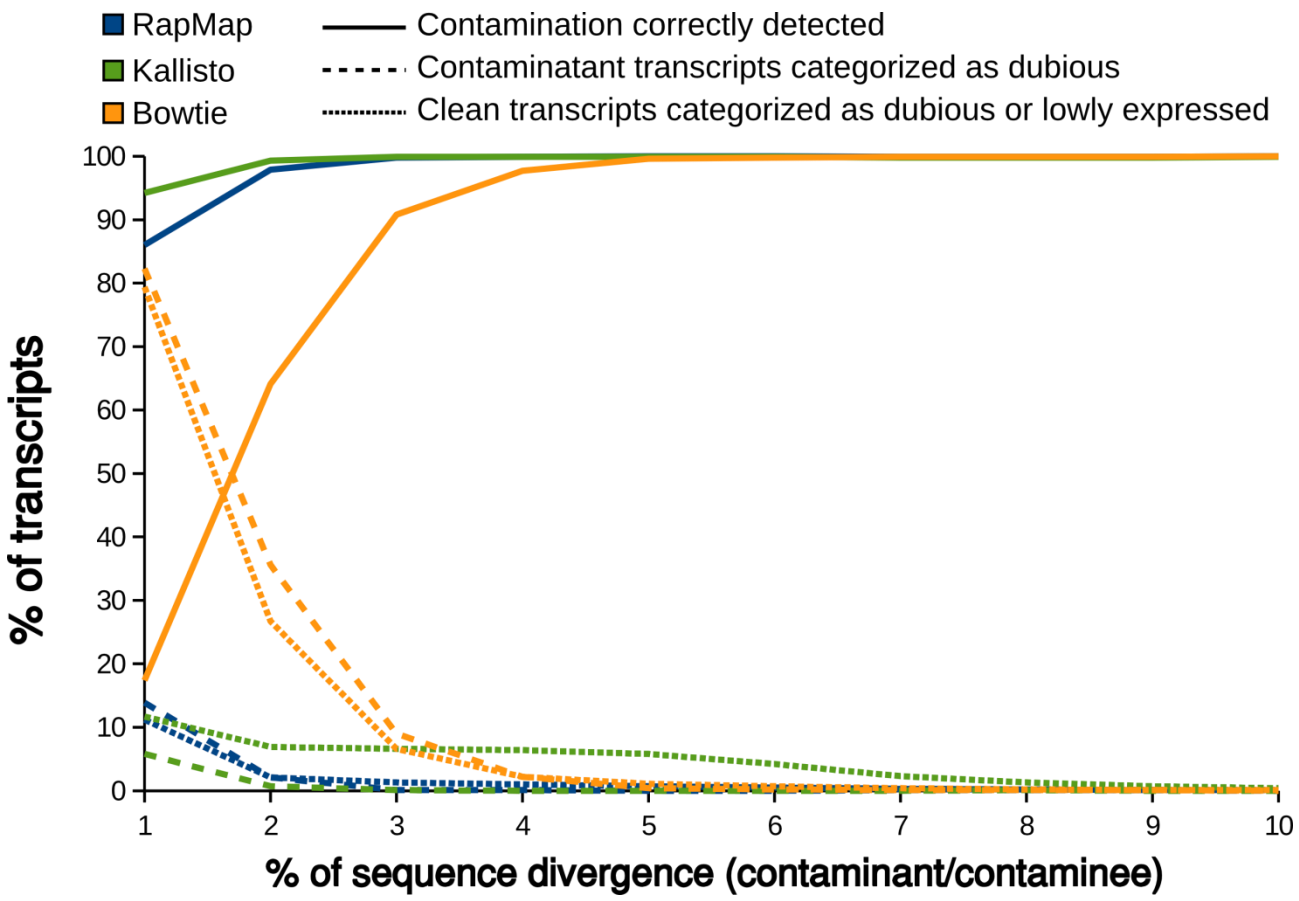
Additional file 1: Figure S1:

Single-gene phylogeny with multiple cross contaminations. Single-gene phylogeny reconstructed from a gene of dataset A belonging to the 14-3-3 gene family (see Methods section for details), showing at least 11 instances of cross contamination. We used CroCo to categorise transcripts and coloured them accordingly: blue for clean transcripts, grey for low coverage transcripts, orange for dubious transcripts and red for cross contaminations.



Additional file 1: Figure S2:

Comparison between transcript categorization by CroCo and a reference set of manually detected cross contaminations. CroCo categorization into five categories of transcripts previously classified as clean (629 cases, in blue) or as cross contaminations (179 cases, in red) using default parameters.

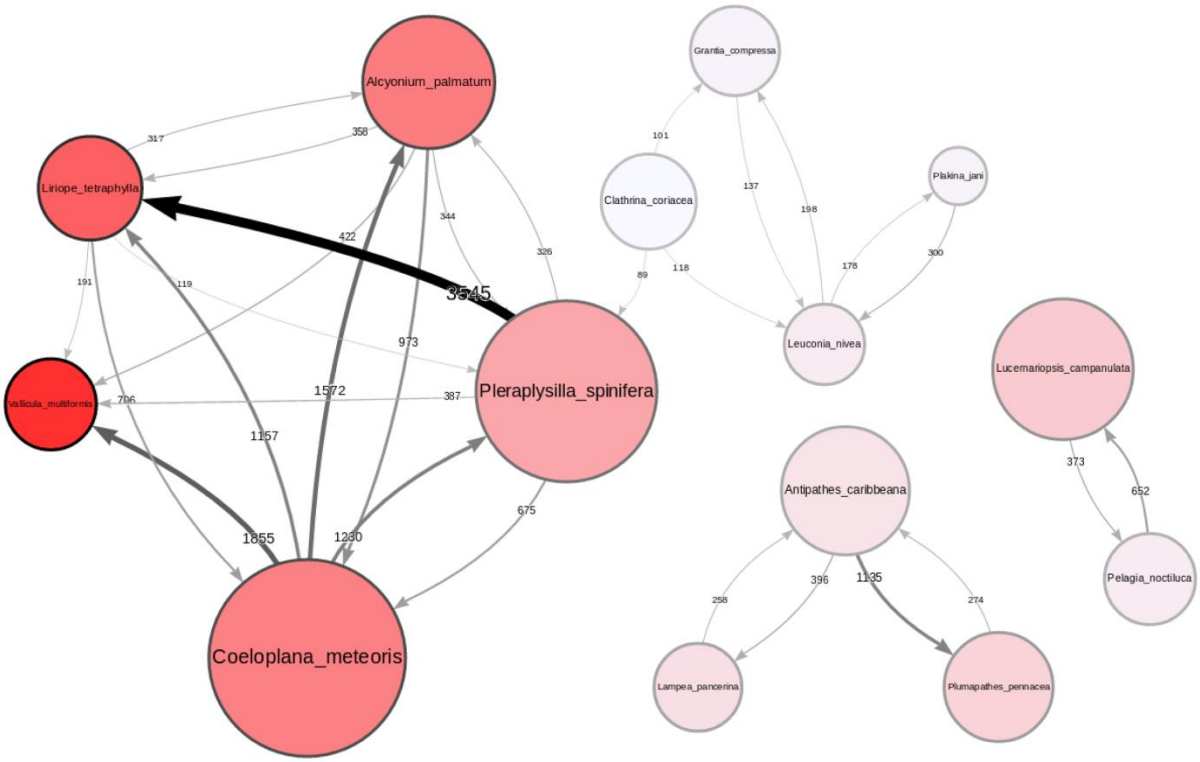


Additional file 1: Figure S3:

Benchmarking CroCo using simulations. Impact of genetic distance between the contaminant and the contaminee for three different mapping tools on the proportion of cross contamination correctly detected, cross contamination detected as dubious and clean transcripts categorized as anything other than clean. RapMap and Kallisto outperform Bowtie for this task.

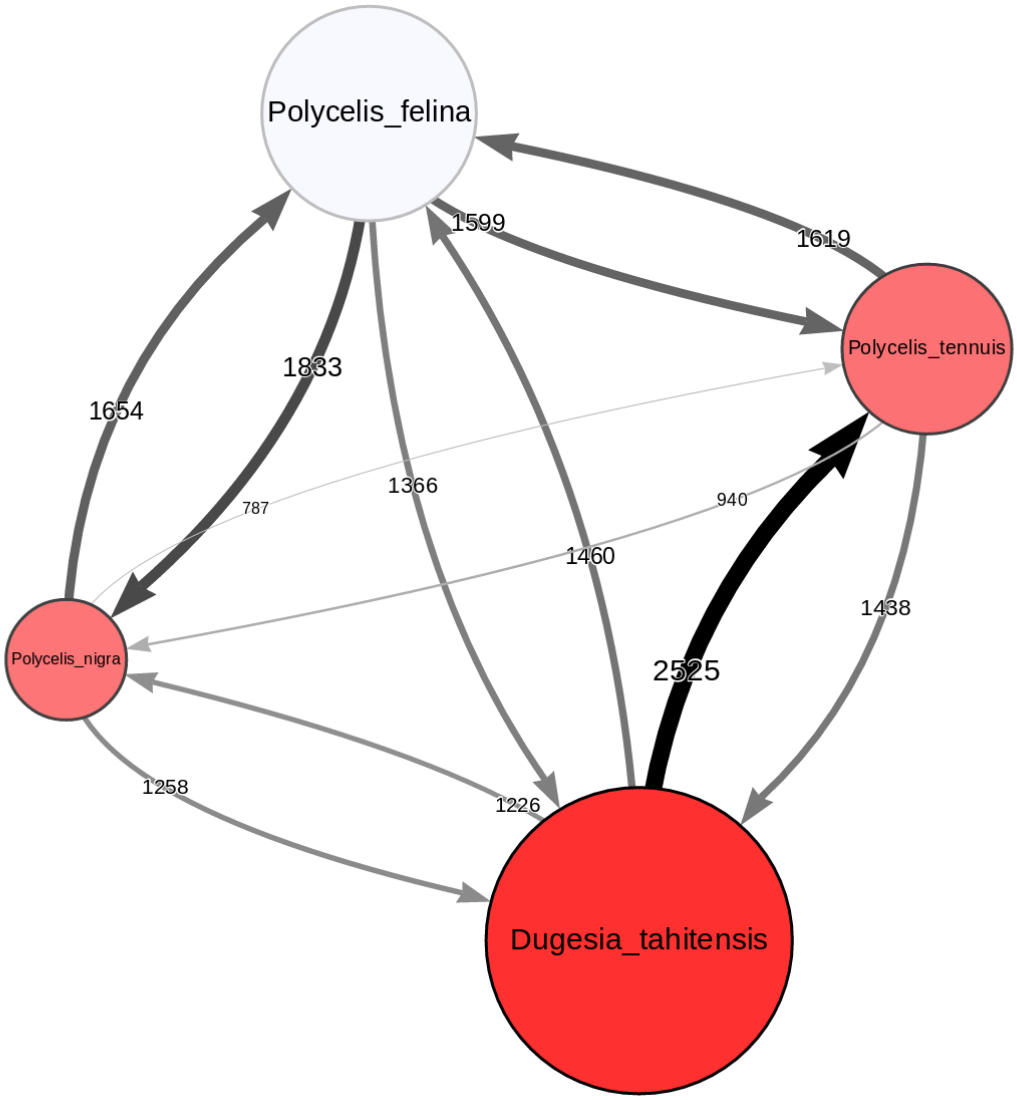
Note for network graph interpretation :

Colors, nodes diameter and arrow sizes in networks are relative to the sampling used and therefore cannot be compared across different sequencing experiments. Example : although *Dryodora glandiformis* looks cleaner in fig. 2a than *Polycelis nigra* in Additional file 1: Figure S5 based on respective colors, both species have ~6% of their transcriptome that is contaminated.



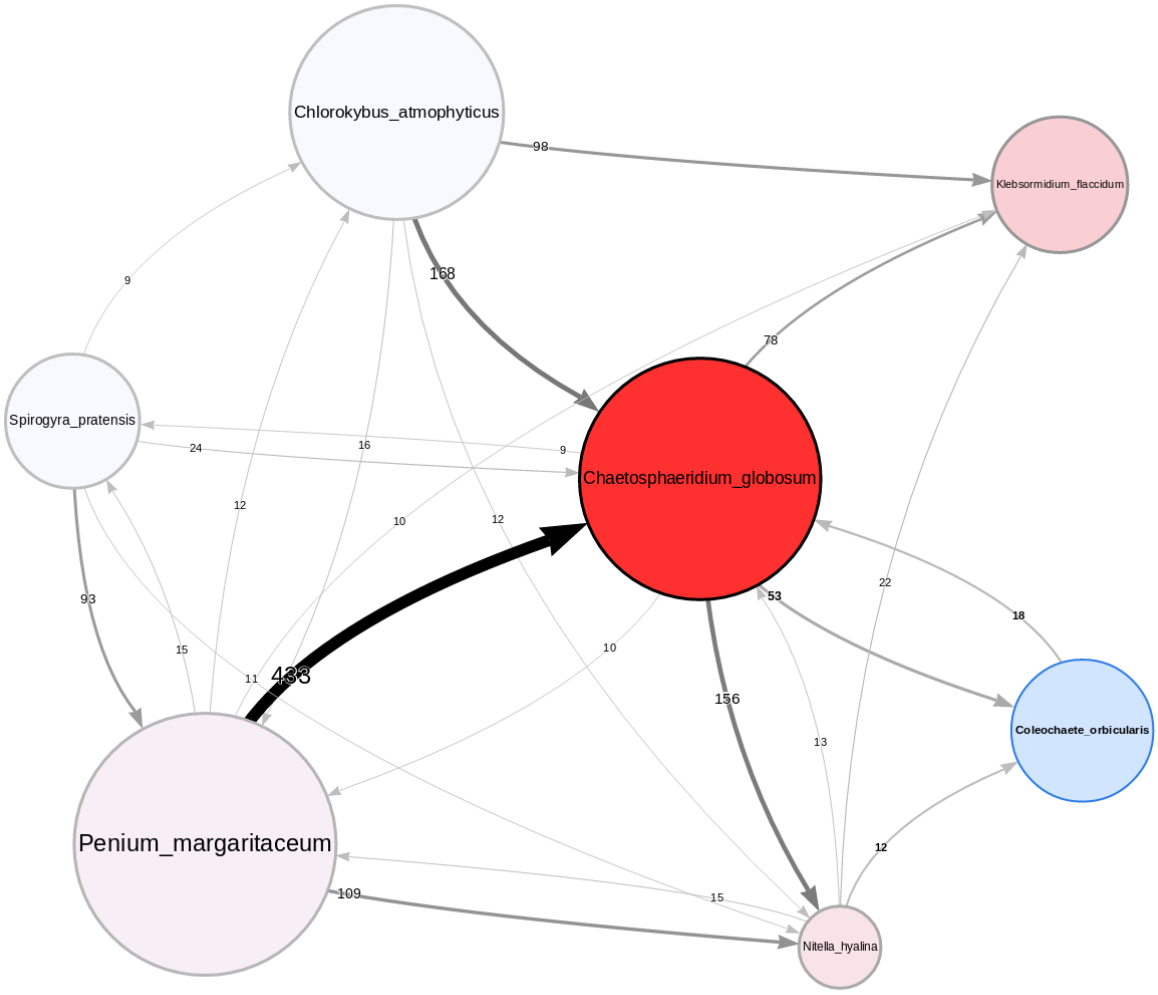
Additional file 1: Figure S4:

Network visualisation of cross contamination patterns in dataset B. Node diameter is proportional to the number of time the sample contaminates another one, node color represent the proportion of its sequence that are contaminated (from white to red), and arrow sizes represent the number of cross contaminations. For clarity, arrows representing less than 2% of the largest cross contamination link are not represented.



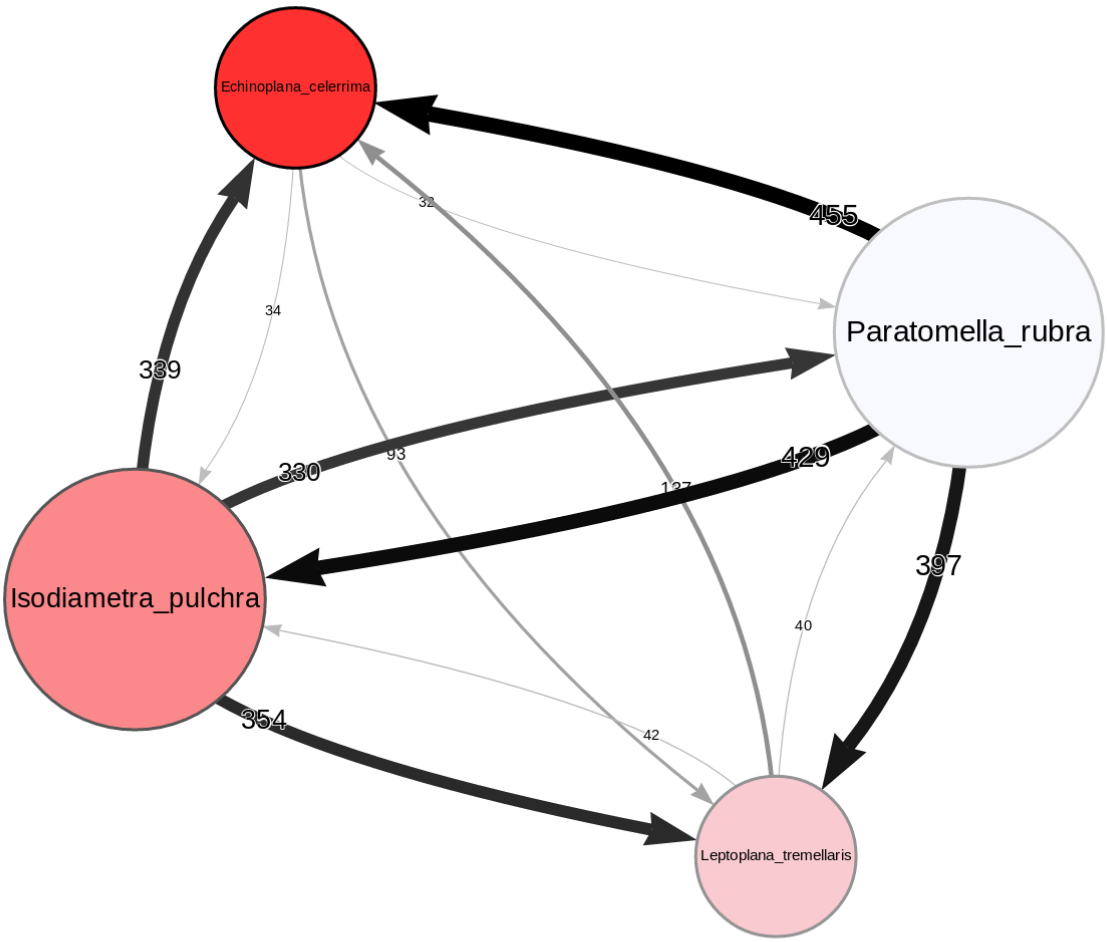
Additional file 1: Figure S5:

Network visualisation of cross contamination patterns in dataset C. Node diameter is proportional to the number of time the sample contaminate another one, node color represent the proportion of its sequence that are contaminated (from white to red), and arrow sizes represent the number of cross contaminations. For clarity, arrows representing less than 2% of the largest cross contamination link are not represented.



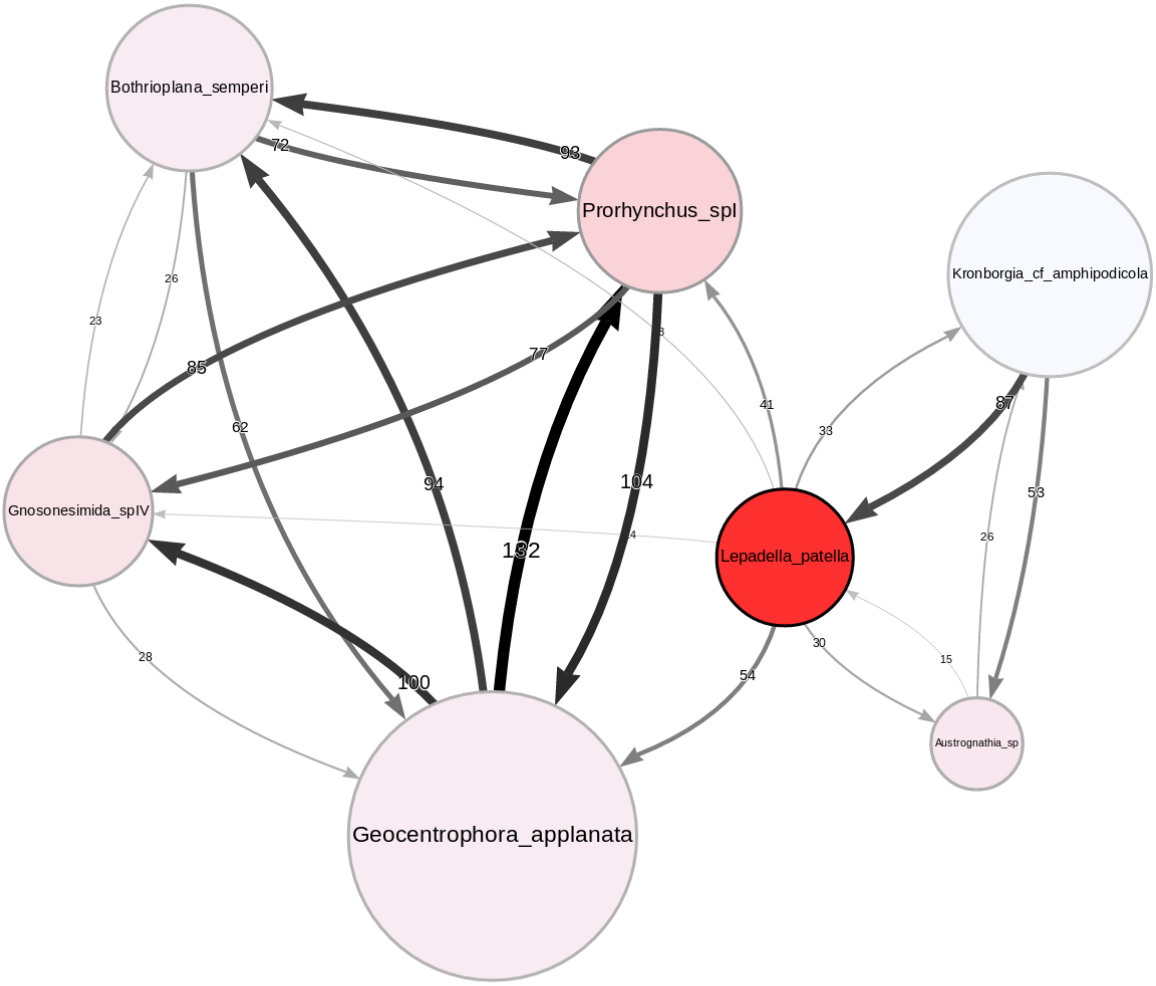
Additional file 1: Figure S6:

Network visualisation of cross contamination patterns in dataset D. Node diameter is proportional to the number of time the sample contaminate another one, node color represent the proportion of its sequence that are contaminated (from white to red), and arrow sizes represent the number of cross contaminations. For clarity, arrows representing less than 2% of the largest cross contamination link are not represented.



Additional file 1: Figure S7:

Network visualisation of cross contamination patterns in dataset E. Node diameter is proportional to the number of time the sample contaminate another one, node color represent the proportion of its sequence that are contaminated (from white to red), and arrow sizes represent the number of cross contaminations. For clarity, arrows representing less than 2% of the largest cross contamination link are not represented.



Additional file 1: Figure S8:

Network visualisation of cross contamination patterns in dataset F. Node diameter is proportional to the number of time the sample contaminate another one, node color represents the proportion of its sequence that are contaminated (from white to red), and arrow sizes represent the number of cross contaminations. For clarity, arrows representing less than 2% of the largest cross contamination link are not represented.

Additional file 1: Table S1:

Datasets from six recent sequencing projects analysed with CroCo. Datasets, species names, taxonomy and accession numbers for sequencing data.

Dataset	Species	Phylum	Accession numbers
A (Moroz *et al.* 2014)	*Beroe abyssicola*	Ctenophora	SRR777787
	Bolinopsis infundibulum	Ctenophora	SRR786491
	Coeloplana astericola	Ctenophora	SRR786490
	Dryodora glandiformis	Ctenophora	SRR777788
	Euplokamis dunlapae	Ctenophora	SRR777663
	Mertensiidae sp.	Ctenophora	SRR786492
	Mnemiopsis leidyi	Ctenophora	SRR789900
	Vallicula multiformis	Ctenophora	SRR786489
B (Simion *et al.* 2017)	*Alcyonium palmatum*	Cnidaria	SRR3407216
	Antipathes caribbeana	Cnidaria	SRR3407160
	Clathrina coriacea	Porifera	SRR3417192
	Coeloplana meteoris	Ctenophora	SRR3407215
	Grantia compressa	Porifera	SRR3417193
	Lampea pancerina	Ctenophora	SRR3407163
	Leuconia nivea	Porifera	SRR3417190
	Liriope tetraphylla	Cnidaria	SRR3407335
	Lucernariopsis campanulata	Cnidaria	SRR3407219
	Pelagia noctiluca	Cnidaria	SRR3407257
	Plakina jani	Porifera	SRR3417194
	Pleraplysilla spinifera	Porifera	SRR3417588
	Plumapathes pennacea	Cnidaria	SRR3407161
	Vallicula multiformis	Ctenophora	SRR3407164
C (This study)	*Dugesia tahitensis*	Platyhelminthes	SRR6436040
	Polycelis felina	Platyhelminthes	SRR6388789
	Polycelis nigra	Platyhelminthes	SRR6379017
	Polycelis tenuis	Platyhelminthes	SRR6388485
D (Finet *et al.* 2010)	*Chaetosphaeridium globosum*	Streptophyta	SRR064327
	Chlorokybus atmophyticus	Streptophyta	SRR064329
	Coleochaete orbicularis	Streptophyta	SRR036732
	Klebsormidium flaccidum	Streptophyta	SRR064330
	Nitella hyalina	Streptophyta	SRR064326
	Penium margaritaceum	Streptophyta	SRR064328
	Spirogyra pratensis	Streptophyta	SRR036731
E (This study)	*Echinoplana celerrima*	Platyhelminthes	SRR1796488
	Isodiametra pulchra	Xenacoelomorpha	PRJNA422347
	Leptoplana tremellaris	Platyhelminthes	SRR1797726
	Paratomella rubra	Xenacoelomorpha	PRJNA422367
F (Laumer *et al.* 2015)	*Austrognathia* sp.	Gnathostomulida	SRR1976176
	Bothrioplana semperi	Platyhelminthes	SRR1955240
	Geocentrophora applanata	Platyhelminthes	SRR1955490
	Gnosonesimida spIV	Platyhelminthes	SRR1976178
	Kronborgia cf. *amphipodicola*	Platyhelminthes	SRR1976457
	Lepadella patella	Rotifera	SRR1976570
	Prorhynchus spI	Platyhelminthes	SRR1980634

Additional file 1: Table S2:

Effect of fold difference parameter value on transcripts categorizations. Transcript catagories, value of the fold difference parameter, number of transcripts in *Mnemiopsis leidyi* and *Vallicula multiformis*.

	fold difference Parameter Value	*Mnemiopsis leidyi*	*Vallicula multiformis*
Clean Transcripts	1.5	117218	5419
	2	114299	5030
	3	98718	4828
Cross Contaminated Transcripts	1.5	2574	52302
	2	2296	47699
	3	2062	33675
Dubious Transcripts	1.5	2229	5384
	2	5426	10376
	3	21241	24602