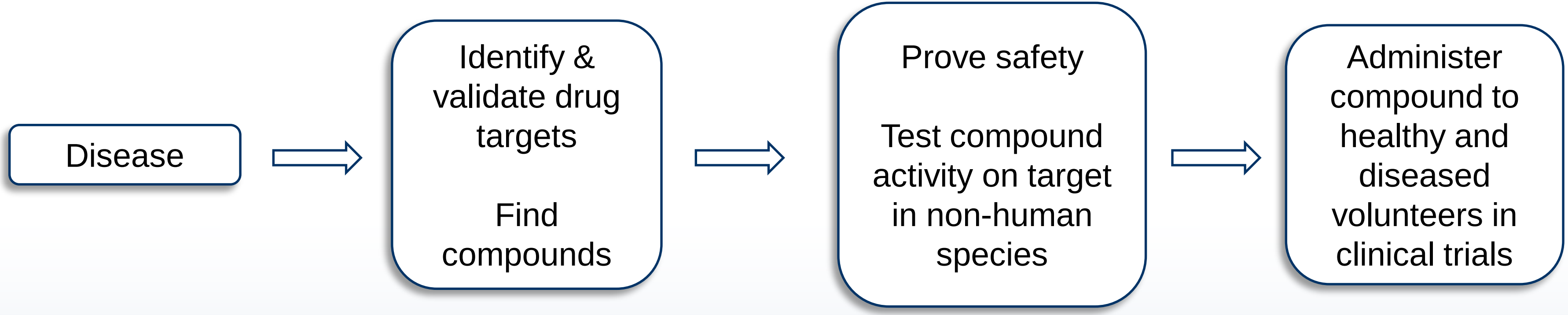
Exploiting orthology and *de novo* transcriptome assembly to refine target sequence information

Julia F. Söllner

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Boehringer Ingelheim

Drug discovery pipeline


Disease

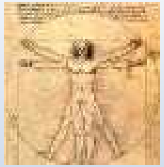
Identify & validate drug targets

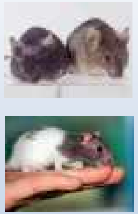
Find compounds

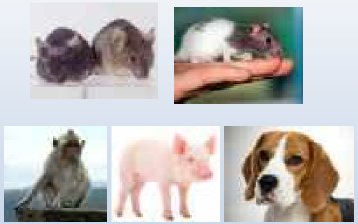
Prove safety

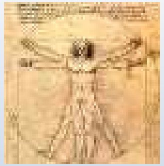
Test compound activity on target in non-human species

Administer compound to healthy and diseased volunteers in clinical trials

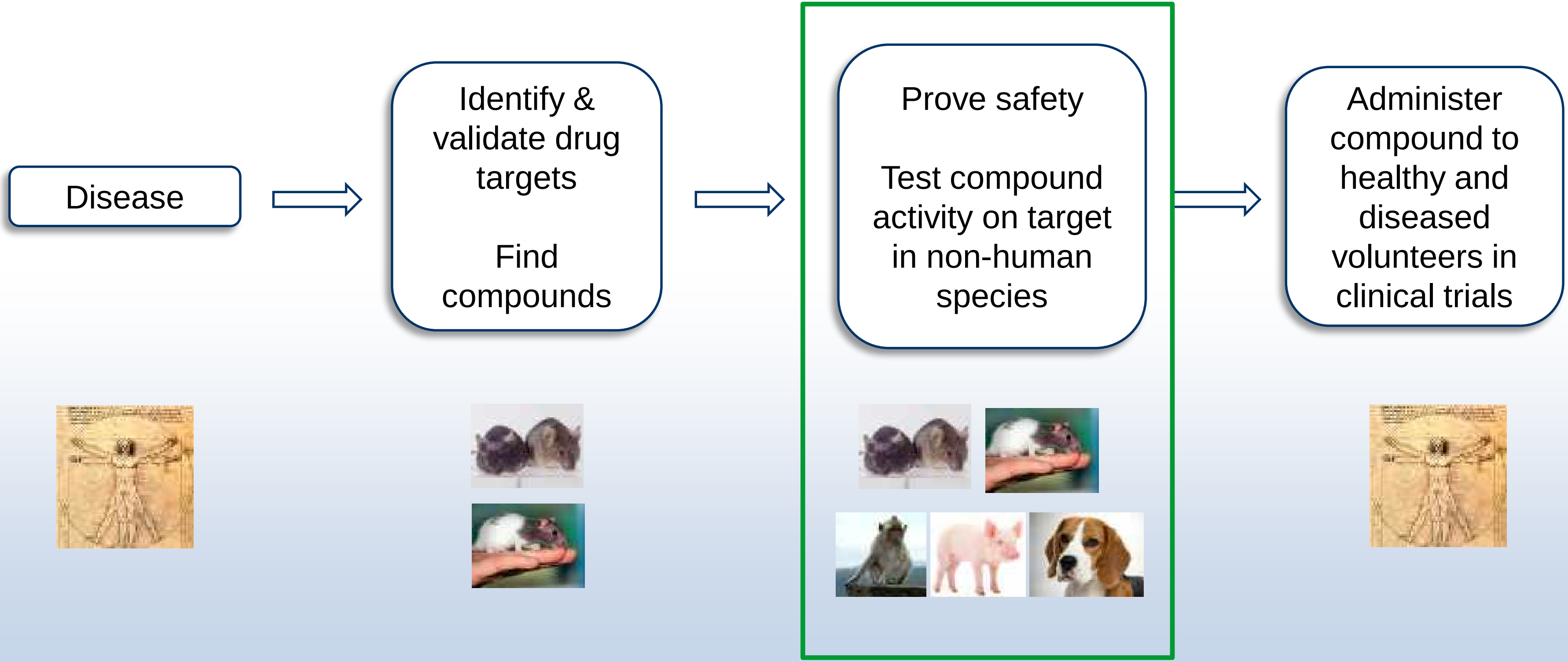








Drug discovery pipeline



Disease

Identify & validate drug targets

Find compounds

Prove safety

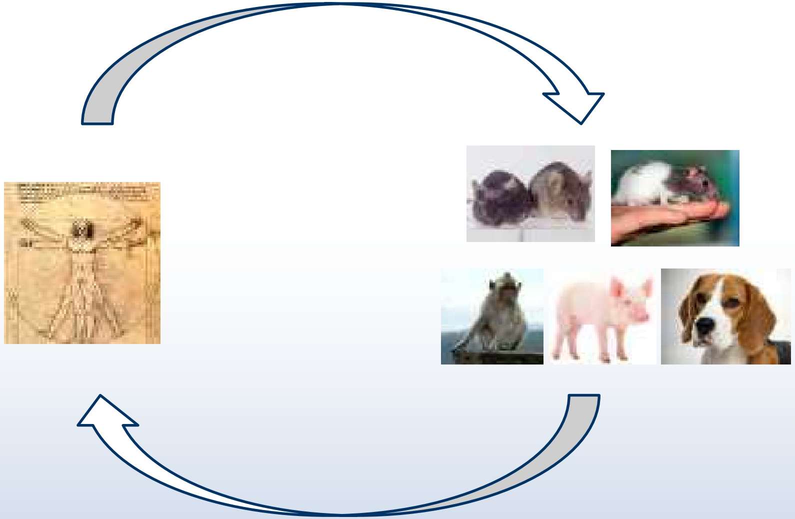
Test compound activity on target in non-human species

Administer compound to healthy and diseased volunteers in clinical trials

Reliable sequences are needed for …

- Correct interpretation of experimental results

- Translatability of results between species



Retrieving sequence information

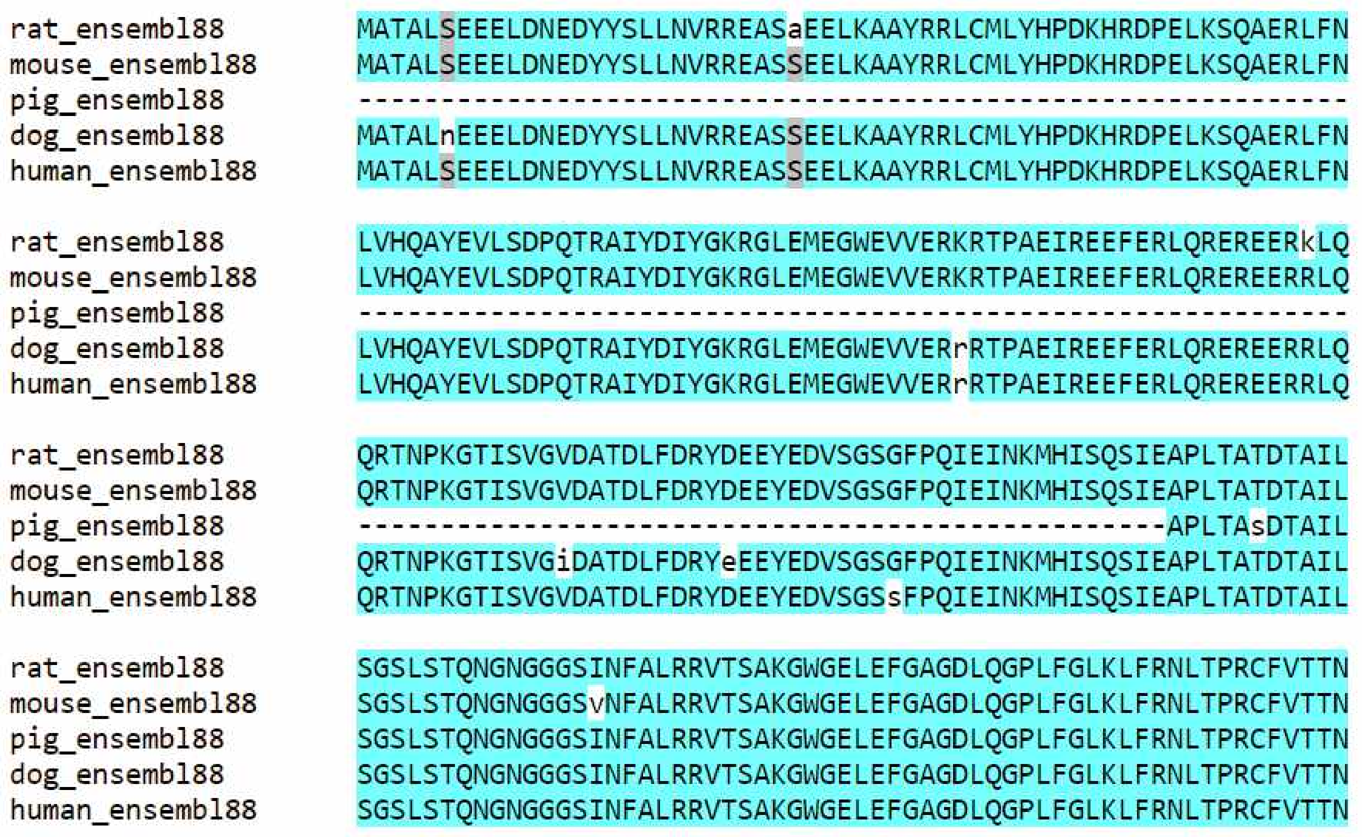
- Public databases:
  - Ensembl
  - UniProt
  - RefSeq

- Manually reviewed sequences from UniProtKB/Swiss-Prot

Example of incomplete pig sequence DNAJC11

orthologues

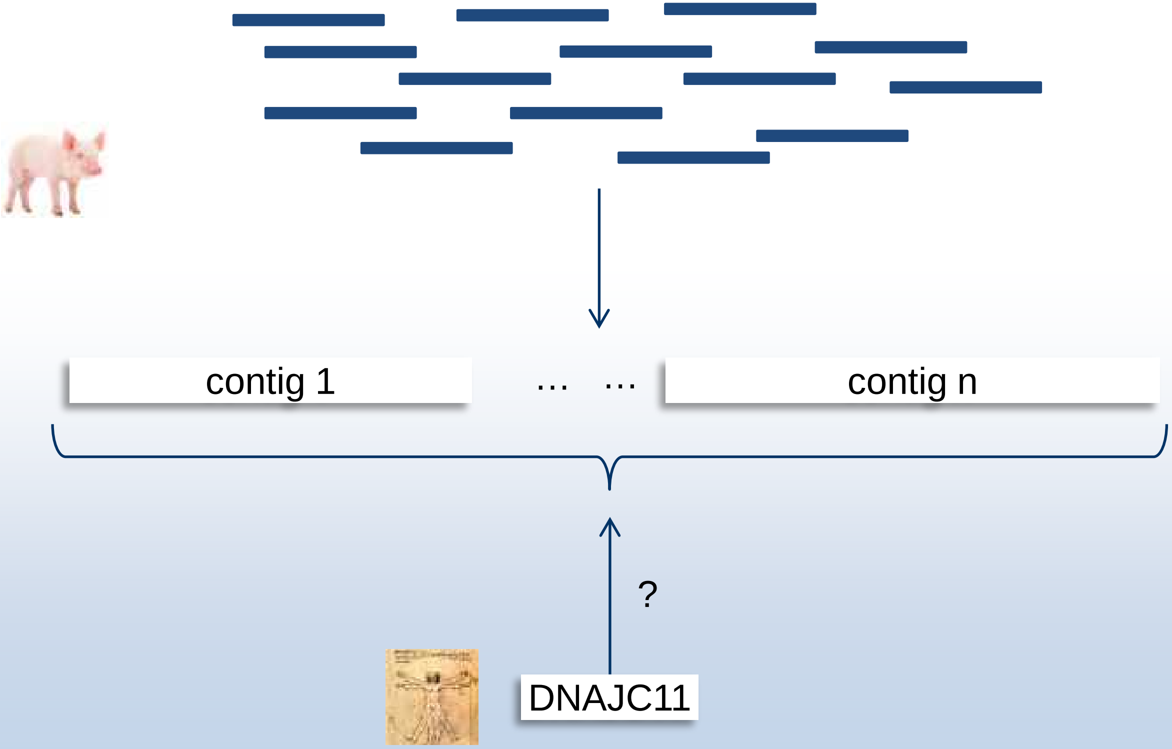
target



```
rat_ensembl88       MATALSEEELDNEDYYSLLNVRREASaEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
mouse_ensembl88     MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
pig_ensembl88       ------------------------------------------------------------
dog_ensembl88       MATALnEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
human_ensembl88     MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN

rat_ensembl88       LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERKRTPAEIREEFERLQREREERkLQ
mouse_ensembl88     LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERKRTPAEIREEFERLQREREERRLQ
pig_ensembl88       ------------------------------------------------------------
dog_ensembl88       LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERrRTPAEIREEFERLQREREERRLQ
human_ensembl88     LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERrRTPAEIREEFERLQREREERRLQ

rat_ensembl88       QRTNPKGTISVGVDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
mouse_ensembl88     QRTNPKGTISVGVDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
pig_ensembl88       -------------------------------------------------APLTAsDTAIL
dog_ensembl88       QRTNPKGTISVGiDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
human_ensembl88     QRTNPKGTISVGVDATDLFDRYDEEYEDVSGSsFPQIEINKMHISQSIEAPLTATDTAIL

rat_ensembl88       SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
mouse_ensembl88     SGSLSTQNGNGGGSvNFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_ensembl88       SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
dog_ensembl88       SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
human_ensembl88     SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
```

Using *de novo* assembly and an orthologous bait sequence for sequence curation

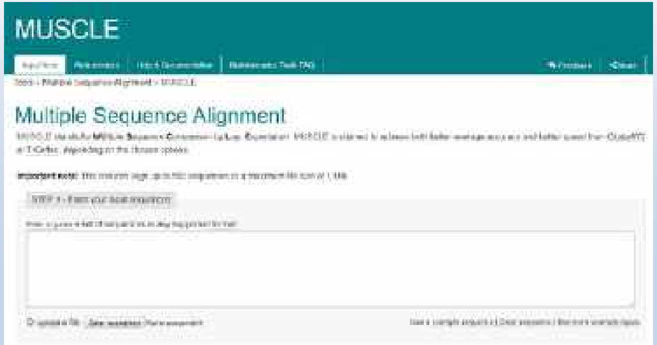
- RNA-Seq reads
- *de novo* transcriptome assembly
- Search with orthologous sequence
- ORF finding & translation



Manual approach for sequence curation

```
BinPacker -s fq -p pair -l left.fastq.gz -r right.fastq.gz

blat BinPacker.fa bait.fa -out=wublast blat.out

samtools faidx BinPacker.fa
samtools faidx BinPacker.fa BINPACKER.100266.1 > myfastafile.fa
```

e!Ensembl



MUSCLE

Tools > Multiple Sequence Alignment > MUSCLE

Multiple Sequence Alignment

MUSCLE stands for MUltiple Sequence Comparison by Log- Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

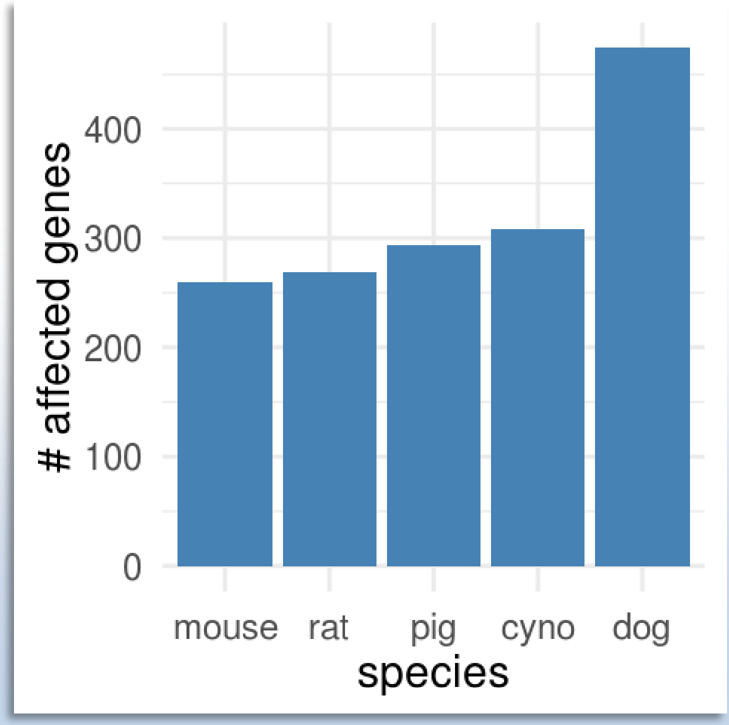
Enter or paste a set of sequences in any supported format:

Or upload a file: Choose file No file chosen

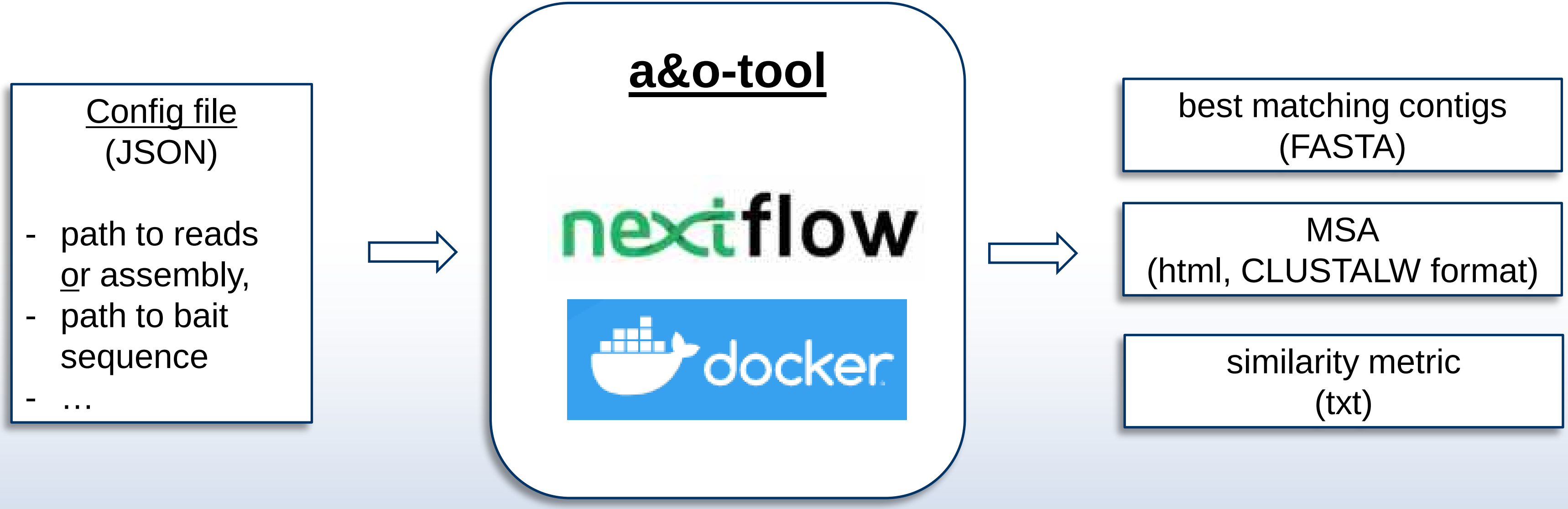
Use a example sequence | Clear sequence | See more example inputs

www.flaticon.com

Number of candidate genes for refinement in 5 model organisms



Automatic pipeline for sequence curation



Config file
(JSON)

- path to reads or assembly,
- path to bait sequence
- …

a&o-tool

best matching contigs
(FASTA)

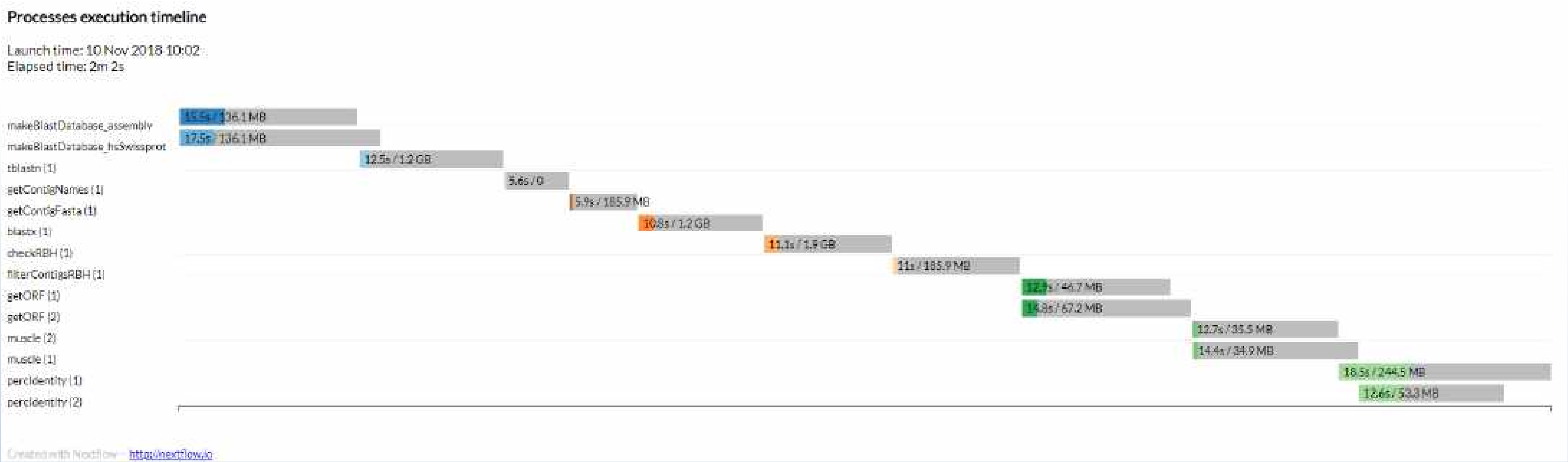
MSA
(html, CLUSTALW format)

similarity metric
(txt)

https://github.com/Julia-F-S/a-o-tool

Runtime and memory consumption for a single target

- With assembly process: ~ 3 h
- With pre-computed assembly: ~ 2 min



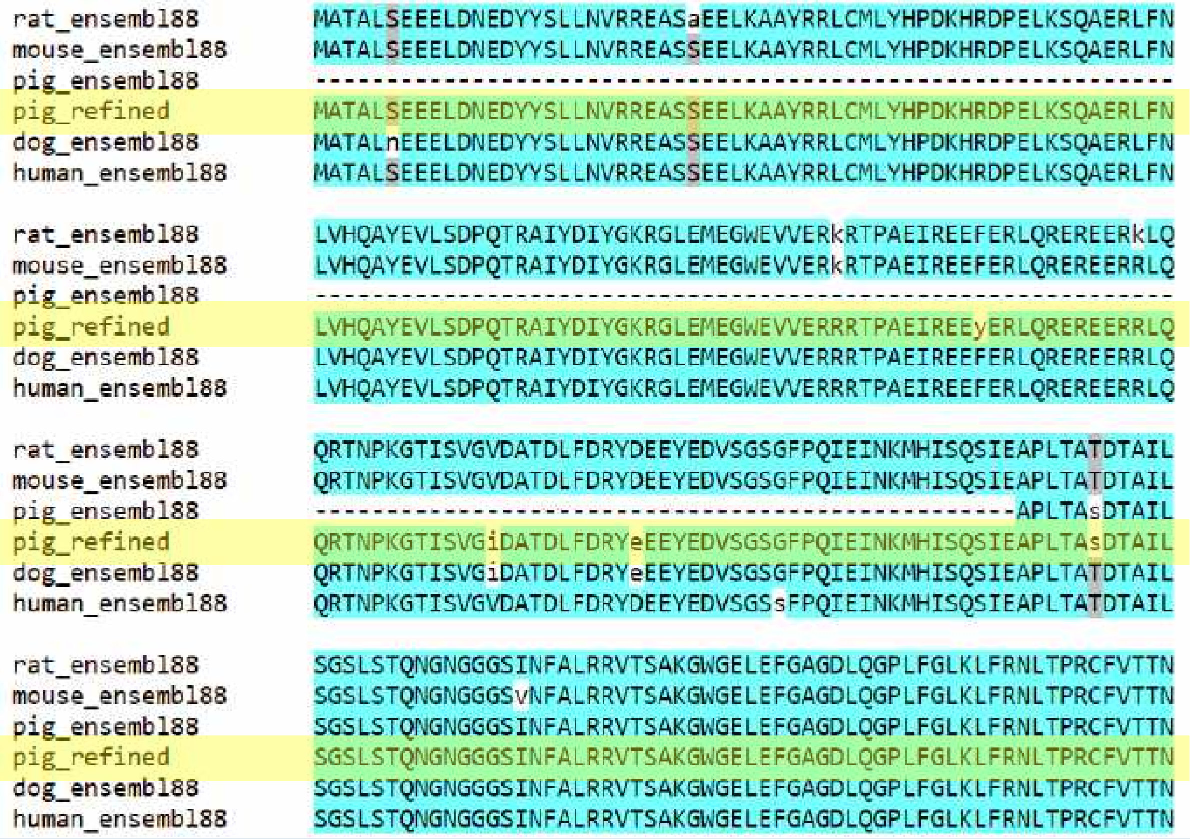
Processes execution timeline

Launch time: 10 Nov 2018 10:02
Elapsed time: 2m 2s

Created with Nextflow – http://nextflow.io

Example of incomplete pig sequence DNAJC11

a&o-tool result

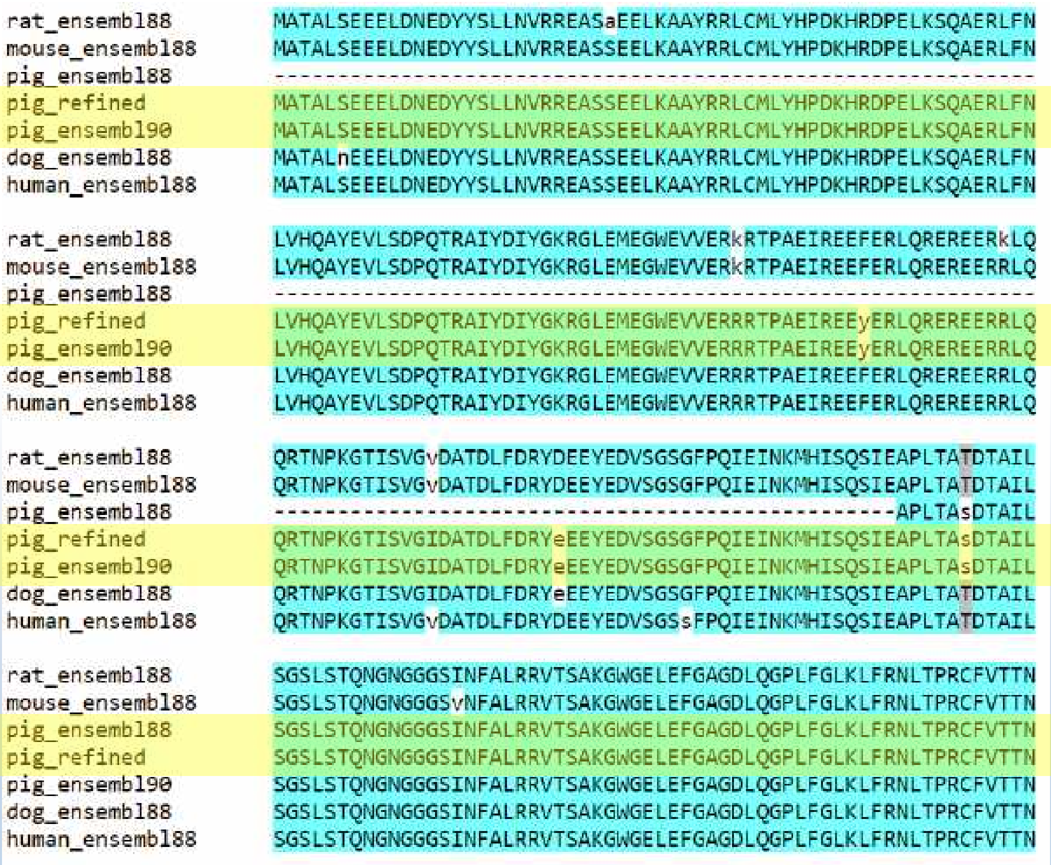


```
rat_ensembl88        MATALSEEELDNEDYYSLLNVRREASaEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
mouse_ensembl88      MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
pig_ensembl88        ------------------------------------------------------------
pig_refined          MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
dog_ensembl88        MATALnEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
human_ensembl88      MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN

rat_ensembl88        LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERkRTPAEIREEFERLQREREERkLQ
mouse_ensembl88      LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERkRTPAEIREEFERLQREREERRLQ
pig_ensembl88        ------------------------------------------------------------
pig_refined          LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEyERLQREREERRLQ
dog_ensembl88        LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEFERLQREREERRLQ
human_ensembl88      LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEFERLQREREERRLQ

rat_ensembl88        QRTNPKGTISVGVDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
mouse_ensembl88      QRTNPKGTISVGVDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
pig_ensembl88        -------------------------------------------------APLTAsDTAIL
pig_refined          QRTNPKGTISVGiDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTAsDTAIL
dog_ensembl88        QRTNPKGTISVGiDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
human_ensembl88      QRTNPKGTISVGVDATDLFDRYDEEYEDVSGSsFPQIEINKMHISQSIEAPLTATDTAIL

rat_ensembl88        SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
mouse_ensembl88      SGSLSTQNGNGGGSvNFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_ensembl88        SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_refined          SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
dog_ensembl88        SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
human_ensembl88      SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
```

Example of incomplete pig sequence DNAJC11

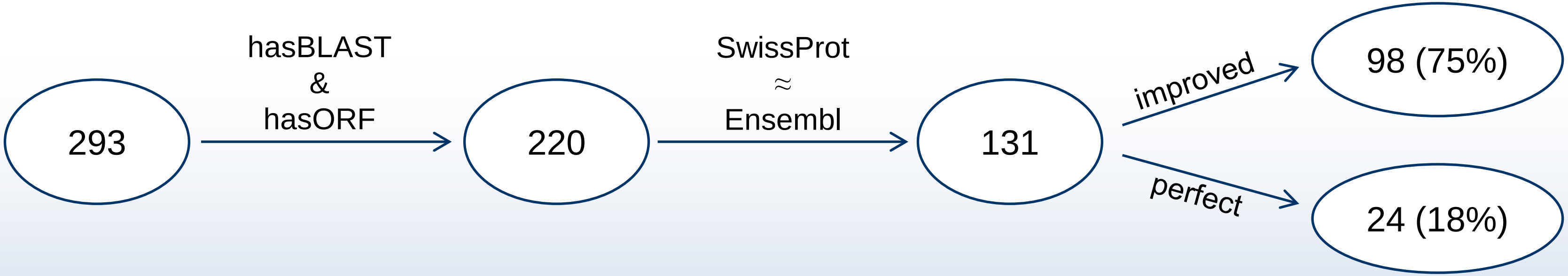
a&o-tool result

Ensembl update



```
rat_ensembl88          MATALSEEELDNEDYYSLLNVRREASaEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
mouse_ensembl88        MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
pig_ensembl88          ------------------------------------------------------------
pig_refined            MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
pig_ensembl90          MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
dog_ensembl88          MATALnEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN
human_ensembl88        MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFN

rat_ensembl88          LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERkRTPAEIREEFERLQREREERkLQ
mouse_ensembl88        LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERkRTPAEIREEFERLQREREERRLQ
pig_ensembl88          ------------------------------------------------------------
pig_refined            LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEyERLQREREERRLQ
pig_ensembl90          LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEyERLQREREERRLQ
dog_ensembl88          LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEFERLQREREERRLQ
human_ensembl88        LVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEFERLQREREERRLQ

rat_ensembl88          QRTNPKGTISVGvDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
mouse_ensembl88        QRTNPKGTISVGvDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
pig_ensembl88          -------------------------------------------------APLTAsDTAIL
pig_refined            QRTNPKGTISVGIDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTAsDTAIL
pig_ensembl90          QRTNPKGTISVGIDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTAsDTAIL
dog_ensembl88          QRTNPKGTISVGIDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAIL
human_ensembl88        QRTNPKGTISVGvDATDLFDRYDEEYEDVSGSsFPQIEINKMHISQSIEAPLTATDTAIL

rat_ensembl88          SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
mouse_ensembl88        SGSLSTQNGNGGGSvNFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_ensembl88          SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_refined            SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_ensembl90          SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
dog_ensembl88          SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
human_ensembl88        SGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
```

Results of a&o refinement for pig



Acknowledgements

- Boehringer Ingelheim
  - Dr. Eric Simon
  - Dr. Germán Leparc
  - Dr. Matthias Zwick
  - Dr. Tanja Schönberger
  - Dr. Tobias Hildebrandt
- University of Tübingen
  - Prof. Dr. Kay Nieselt
- The nextflow team ☺

Thank you for listening!

Questions?

References

1. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucle. 2018;46:754–61.

2. The Uniprot Consortium. UniProt : the universal protein knowledgebase. Nucleic Acids Res. Oxford University Press; 2017;45:158–69.

3. Leary NAO, Wright MW, Brister JR, Ciufo S, Haddad D, Mcveigh R, et al. Reference sequence (RefSeq) database at NCBI : current status , taxonomic expansion , and functional annotation. 2016;44:733–45.

4. Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov A V, Yim SH, et al. Gene expression defines natural changes in mammalian lifespan. Aging Cell. 2015;14:352–65.

5. Uhlén M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28.

6. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80-). 2015;347.

7. Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, et al. BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. PLoS Comput Biol. 2016;12:1–15.

8. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate : reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016;26:1134–44.

9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

10. Edgar RC, Drive RM, Valley M. MUSCLE : multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

11. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol [Internet]. 2017;35:316–9. Available from: http://www.nature.com/doifinder/10.1038/nbt.3820

12. Docker [Internet]. [cited 2018 Nov 5]. Available from: https://www.docker.com/

13. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet : Visualization of Intersecting Sets. IEEE Trans Vis Comput Graph. 2014;20:1983–92.