5

Aspects of Quality and Project Management in Analyses of Large Scale Sequencing Data

Björn M. von Reumont, Sandra Meid and Bernhard Misof
Zoologisches Forschungsmuseum Alexander Koenig,
Adenauerallee 160, 53113 Bonn,
Germany

1. Introduction

We describe step-by-step the outline of a project, in which the evolutionary history of pancrustaceans (crustaceans and hexpods) was revisited using molecular methods. It was part of a larger program, the 'Deep Metazoan Phylogeny' priority program of the Deutsche Forschungsgemeinschaft (DFG), wich aimed to reconstruct the metazoan tree of life involving more than 30 subprojects. This chapter should be understood as a backbone, that clarifies important points to plan and to conduct projects in molecular biology, also using next generation sequencing data. The text is divided in four parts: 1) theoretical aspects to projects in molecular biology, 2) the process from the collection of material in the field to the final sequencing, 3) the process from the sequence to the reconstructed topology with a special emphasis on data quality, and 4) the conclusions to prevent pitfalls.

1.1 Fascination and complexity of molecular evolutionary biology

Working in molecular evolution to reconstruct the evolutionary history of organisms is a very fascinating, but also very complex issue. Per definition evolutionary biology, and respectively molecular evolutionary biology, is the division in science, which overlaps and intersects mostly with other areas of natural sciences, like chemistry, physics, informatics, mathematics, bioinformatics, geography but also philosophy and history. Exactly that complexity and intersection creates the fascination and addiction of many scientists to work in that area.

Being on field excursions and collecting specimens in their natural habitats is like travelling back in time into the century and time of classic field biology, geography and history. If once the laboratory part has started, technicial and laboratory skills are demanded, while in parallel the amount of characterized sequences starts to force one to become a sophisticated software user, partly applying bioinformatics knowledge or (the often much faster alternative) cooperating with bioinformaticians. The analyses, interpretation and discussion of the results represent the climax of the project by some (at least) publications in highly respected journals.

1.2 General management strategies applicable for scientific projects in molecular evolution

In general, scientists are highly educated in their specific disciplines, but are often 'freshmen' in managing projects with all involved aspects.

These eventually less developed soft skills can cause an underestimation of possible volume of work and subsequently lead to a massive lack of time, which finally degrades the results and the quality of the scientific project. A rigorous project management as conducted in economics featuring a global, yet detailed intersected time schedule with 'milestones' as anchor points and deadlines (including buffer-time in reserve) as general frame in a project roadmap is mandatory for a solid project. The 'golden triangle' of project management (e.g. Kerzner, 2009; Litke et al., 2010) illustrates interrelations that affect projects and their quality management: A) goals and qualitative results, B) planned time schedule and C) calculated costs. If one edge of that triangle becomes delicate, all could be at risk, and the quality of the project is affected (see figure 1).

Fig. 1. The golden triangle of project management adapted to molecular projects. The red arrows indicate where the points written outside the second (red) triangle have most impact. However, some points have an impact on more than just one edge. Laboratory difficulties for example cost primarily time, but also stress the budget. If things go wrong (and mostly they unfortunately follow the law of Murphy in the scientific business) goals might also be affected by laboratory difficulties. The core triangle pictures the three main components, which are interwoven. If one edge is affected, the other ones are affected either. A major specification is probably, that A and B generally are more connected with each other in most aspects, while the budget is constant or not directly affected (golden arrows). If e.g. computational analyses of phylogenetic trees do not work or cause difficulties, a delay in the time schedule is created, that primarily affects the results, but not directly the budget

If a larger project is conducted, in which more persons are directly involved or third parties included (e.g. by outsourcing of sequencing to companies, etc.), additional aspects play a veritable role. Who is directly or indirectly involved in and linked to the project? Which interests and influence (negative and positive) have the different persons or parties in the project? All of these involved persons (with different expectations and interests) are stakeholders of the project. In general, a stakeholder analysis in the planning phase is extremely crucial and a standard approach in economics (Weaver, 2007; Freeman, 2010; Litke et al., 2010). Which risks might rise by involved persons? In science, competion between work groups must be considered. Is cooperation possible, which is always to prefer. If no cooperation is feasible, which risks exist subsequently for the project? If third parties are involved by outsourcing of e.g. sequencing, an exact analysis of possible candidate companies and their interests and capability are important (see also additionally paragraph 2.3). Last but not least, if you are a PhD student or postdoc do not forget one very important or even the key stakeholder (Bourne, 2010), the PI or supervisor. What are his interests, which are yours? Is there a risk or conflict you might have to deal with or to solve? What are his expectations? Perhaps an agreement on objectives is necessary. One major factor is an open discussion, regular (scheduled) communication and time for additional, intermediate meetings; also a clearly communicated agreement on objectives avoids difficulties or even disappointment of one or both parties in the project.

The communication strategy is a further key factor (Bourne, 2010), it is important to prevent typical pitfalls like 'just reporting', 'flood of detailed information' and that 'no feedback' is given. See also general principles of communication to transport information (Chapter 1.3.5/1.3.6 in: Wägele, 2005; Bourne, 2010). Communication is quite clearly time consuming, but it pays off. All points of the golden triangle are linked to communication, including budget and quality of results. Communication skills improve the general quality of the project, can save costs and time, and eventually most importantly: control and enhance the motivation of the involved persons.

Several software packages to coordinate communication, interaction and project work exist to provide an effective platform and frame to conduct and coordinate projects. Examples are Teamwork, OpenLab, Italy; Teamlab, Ascensio System (open source); Clarizen (web based); Endeavour software project management, Ezequiel Cuellar (open source). If you are a bioinformatician, the last package might be respectively interesting.

A characteristic of scientific projects is that new open questions and potentially new fields of methodologies are explored. Respectively, if additionally laboratory work is included, the risk to end without any or absolutely unexpected results (latter one might result in the desired nature paper) is part of the scientific business and in general hard to evaluate. That has to be calculated in advance and should be reflected in the time and risk management.

However, there is also a clear difference between projects in economics and science: scientific projects aim in most cases for fundamental and theoretical insights instead for a direct financial benefit of involved parties. Changing and evaluating laboratory methods for example, might be unexpected time consuming, but necessary and can at the same time establish a new state of the art method. Time and space to walk open minded on paths that seem to be ineffective, not suitable or even out of topic at first glance might bring the breakthrough and must be possible. Louis Pasteur (1822-1895) quoted on his accidentally discovery of penicillin, "chance favours the prepared mind", but one condition for this famous quote is, that the scientist needs the (mentally) freedom to meet chance. A too rigid framework and control might hinder that. Contrariwise many scientists focus often too much on details (as being trained for) and loose their track on the overall relations of the

project, which provokes a rather high inefficiency. Consequently a compromise between efficiency and creativity/innovation has to be made. This is easy to write, but hard to transfer and to realize, as personally experienced.

2. Project phases from species collection in the field to sequencing

2.1 Collection and fixation of samples in the field – RNAlater or sooner?

Normally, the planned molecular project starts with the extraction of molecules (DNA or RNA) from specimens (see figure 2) and every true biologist will do his very best to collect and preserve these specimens by himself in the field.

If the specimens or the tissue is preserved in Ethanol for DNA based work, 94% (or higher), ethanol p.a. should be used. This is true for every tissue collected in the field. Despite the rumour, that crustaceans are tricky to sequence in the laboratory, because the aggressive enzymes of the exocrine glands rapidly degrade the DNA, this specific experience was never made working with 94% ethanol p.a.. Working with material collected and sent by colleagues, difficulties appeared and could be linked to the quality (not p.a.) or low concentration of ethanol. Especially material of larger, vessel based expeditions, is obviously often stored in ethanol, which has been diluted due to ethanol shortage during the cruise. If you expect to join an expedition, plan enough quantities of 94% ethanol (and you better hide some of the ethanol in case colleagues did not properly calculate their ethanol contingents, they seem to tend to desperate actions in these situations). Storing the samples in -20 °C probably keeps degradation processes at a low level, but fieldtrip cooling is not obligatory to preserve high quality DNA.

However, cooling plays a veritable role, if you have to collect samples in the field for RNA based analyses. RNA as a single stranded molecule can be degraded very fast (and unfortunately very efficiently) by a group of enzymes, called RNAses. These enzymes are nearly omnipresent in our body including e.g. perspiration liquid. They have to be inhibited by cold temperatures or chemicals (or both) to stop RNA degradation. The best procedure to ensure good quality of RNA samples is consequently to collect the specimen and to extract the RNA immediately. Unfortunately this is in most cases not possible in the field. For example, many groups of crustaceans live in remote habitats.

For example, remipedes live in anchialine cave systems (see figure 2, top right picture) and require cave diving expeditions. They were collected by BMvR on the Yucatan peninsula in Mexico. Even the organization of the cooling chain to freeze the samples directly in the field and to ship them to the laboratory for RNA extraction was not possible: logistic companies that could have shipped the samples in time did not ship dry ice due to regulations of the International Air Transport Association (IATA), In general, the dry ice transportation by airplane is not officially authorized and problematic in some countries. Awareness and integration of such eventual logistic problems are eminent for a realistic project plan and time schedule.

Using RNAlater for RNA isolation is one solution to collect specimens. It is a non toxic, non flammable liquid that can be transported everywhere without any problems (even in airplanes) and it preserves RNA at room temperature at least for 5-7 days (Grotzer et al., 2000; product descriptions of e.g. Qiagen, Applied Biosystems) without loss of quality compared to frozen samples (Grotzer et al., 2000; Mutter et al., 2004; Gorokova, 2005). A closed cooling chain is not mandatory. For preservation of microcrustaceans of zooplankton like copepods, up to a month of storage time is possible without any losses of RNA quality if RNAlater is used (Gorokhova, 2005). Own experiences corroborate this study with samples

Fig. 2. Overview of the typical phases within a molecular project that start with material collection in the field and end with the final sequences. The two pictures in the left on top [1] show a dissected house gecko (*Hemidactylus frenatus*), which was parasitized by tongue worms (Pentastomida, small picture) in his lunge tract. On the right, a remote anchialine cave system in Mexico is shown. Within these caves live the enigmatic Remipedia (*Speleonectes cf. tulumensis*) that were collected by cave diving

of different sizes like copepods, ostracods, remipedes, and leptostracans, which were stored at room temperature for up to 14 days after collection (including transportation and shipping time). High temperatures may harm the sample quality despite RNAlater preservation, depending on the general temperature conditions of the expedition area. Good experiences were made with standard fridges (about 4°C), they are easy to organize and the sample is cooled, but not frozen.

RNAlater should have room temperature for preservation of tissue samples to enable a thorough penetration, and the liquid should not be cooled before and directly after preservation of material. Before preservation, tissue has to be cut into little fragments, additionally use a pestle (even some smaller crustaceans have a carapace that has to be cracked) to ensure a fast diffusion of the liquid into the tissue. After a few hours or a day, RNAlater can be moderately cooled. If frozen away after one day, a cooling chain must be guaranteed.

For marine organisms a careful sorting or sample preparation is crucial before the preservation of tissue to prevent larger amounts of salt water to dilute and affect the preservation liquid. In general, RNAlater should be sufficiently added to the sample, about 1:5-10 between sample and RNAlater (according to manufacturer protocols) turned out to be insufficient. Even for smaller specimens 15-25 ml tubes were at least used, depending on the collected numbers.

However, contrary to own good experience with RNAlater, other projects using RNAlater to preserve representatives of evolutionary early hexapod lineages report frustrating results, gaining degraded RNA or only very few EST sequences. As stated, the best method has to be tested for each species group. In that special case the best choice was liquid nitrogen, with all subsequent difficulties in the field. An interesting effect is, that RNAlater perfectly preserves DNA (Gorokova, 2005; Vink et al., 2005), which makes it an ideal alternative to ethanol preservation.

The main goal of many projects in molecular biology is the reconstruction of the evolutionary history of species. In this context so called large-scale next generation sequencing approaches have recently been used applying RNA based sequencing (see paragraph 2.3). The approach aims to randomly sequence expressed genes of a specimen when the tissue or specimen was collected and preserved ('transcriptome shot'). One quality criterion to achieve a good coverage of different genes is, how fast the specimen was preserved. If the stress level of the specimen was high, a relatively high level of stress response proteins are the consequence, biasing the quantity but also quality of finally sequenced genes. Always ensure that stress is kept to a minimum level for organisms before preservation to guarantee a maximum number of represented genes. Another important method to achieve a maximum intersection of expressed genes is the collection of different larval and/or development stages of an organism to cover possibly different gene expression patterns. If parasitic forms are sampled, like tongue worms, that parasitize the respiratory tract of vertebrates (Pentastomida, see figure 2 top left picture), a careful preparation of the tissue is necessary to prevent contamination by the host tissue.

Collected specimens should carefully be determined before preservation. Additionally, collected and stored voucher species might enable a second identification after sequencing, if unexpected results or difficulties occur. This specific point is often forgotten. An approach to centralize the storage of voucher specimens and DNA including the linked collection and laboratory data is the DNA bank network (Gemeinholzer et al., 2011). This platform provides an efficient and practical solution to access and exchange data and tissue in an extended form, compared to classical accession sheets like in GenBank. This storage allows a

general traceability of DNA sequences, and their quality concerning specimen identification and the DNA itself, like concentration, signal strength, electropherogram etc. In most cases this information is missing in published NCBI data (see figure 4).

2.2 Extracting DNA, RNA and subsequent amplification of the molecules

The extraction of DNA or RNA from tissue follows standard protocols and available kits (e.g. Mülhardt, 2008; Sambrook & Russel, 2000). Eventually it is reasonable, to test different kits and protocols to be time efficient.

A fast and specifically tested method is needed to isolate RNA from tissue. Only few studies mainly from the medical/clinical field are published, which show that quality and quantity of RNA yields are dependent on used preservation/isolation method and extraction kits; additionally both parameters can improve using RNAlater (Forster et al., 2008; Hemmrich et al., 2010, see also Gorokhova, 2005). One serious consideration should be outsourcing of RNA extraction and subsequent sequencing. Time is saved if one party or company provides service from extraction to the final sequences, also in cases of difficulties with the samples.

The PCR method is an established method and several specific adaptations exist to ensure the maximum sensitivity to amplify the desired fragments (e.g. Mülhardt, 2008; Palumbi in: Hillis et al., 1996).

Everyone who works in a molecular lab performing PCR knows that this step is the most sensitive and delicate one for possible contamination. Consequently, a rigorous management should be conducted to maintain high standards in working procedures (Mülhardt, 2008; Sambrook & Russel, 2000). The awareness that contamination can happen despite all efforts is important. If that is considered and influences a general risk management, in consequence all sequences, which are finally included in analyses are blasted in a standard procedure. Exactly this step is the last bastion to guarantee as first step the quality of phylogenetic analyses. If a contamination occurred, the contaminated sequences must be identified and excluded (see figure 4 and paragraph 3.1).

2.3 The sequencing process – A typical case for outsourcing

The term phylogenomics was coined by Eisen (1998) and is recently used for analyses including large scale sequencing data and large numbers of genes derived from cDNA libraries (see also Philippe et al., 2005). A new strategy is the sequencing of the 'transcriptome', which represents the set of expressed genes in an organism, that are encoded by mRNA molecules. Most mRNA molecules are tagged by Poly-A tails and thus easily to fish by specific adaptors if total RNA was isolated. These fished mRNA molecules are reverse transcribed in cDNA and finally libraries are reconstructed that represent ideally all expressed genes in an organism. These mRNA fragments are called expressed sequence tags (ESTs) because of their poly-A tail 'tag' (excellent reviews on that topic: Jongeneel, 2000; Bouck & Vision, 2007). With the new technology of pyrosequencing the possibility arose to directly sequence cDNA molecules in a large scale sequencing approach. Pyrosequencing is not based on the principles of the Sanger sequencing with chain termination reactions, but instead on an enzyme cascade, which generates light if deoxynucleotides are added and pyrophosphate is separated. This difference enables a highly miniaturized and parallelized procedure and technique (see figure 3). For more details see Ronaghi, 2001; Shendure et al., 2004; Ellegren, 2008; Hudson, 2008; Petterson et al., 2009; Voelkerding et al., 2009.

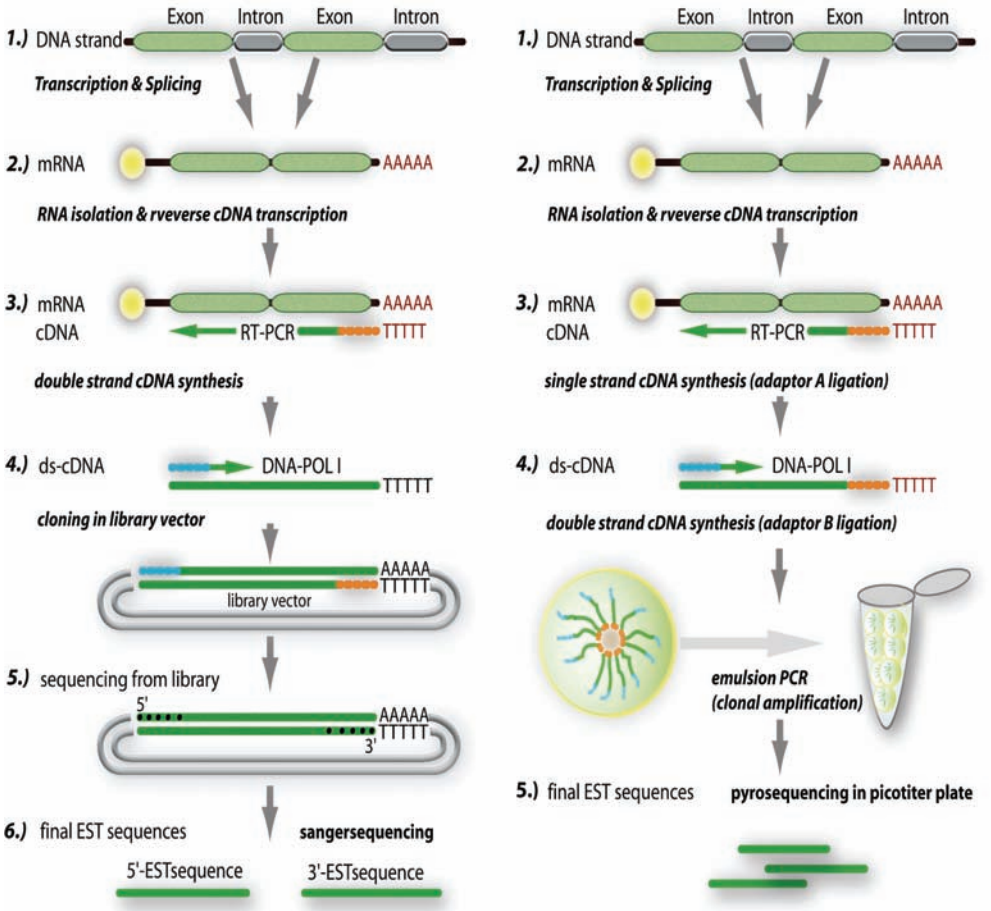

Fig. 3. Differences between standard sanger-sequencing (on the left) and the new pyrosequencing technology (on the right) of next generation sequencing (NGS). Both technologies use mRNA specific target sequences to extract mRNA form the total RNA, which is isolated from tissue. The main difference is that the time and cost intensive step of fragment cloning and sequencing from a subsequently picked library is skipped for pyrosequencing. Depending on the precise technology, double stranded cDNA is generated by an emulsion PCR, in which fragments are amplified in micro compartiments. The sequence fragments are finally transferred on picotiter plates for a massive parallel sequencing

Sequencing is frequently outsourced, which offers a price level that is hard to beat by do-it-yourself sequencing at universities or other research institutions. Focused on large scale or next generation sequencing, some points should be considered. In most companies laboratory procedures and steps are ISO certified ensuring a guaranteed high level of quality and reproducability.

It is a specific quality of molecular biological studies that often unique samples of species with rather unknown evolutionary history are analysed. The collection of these specimens is

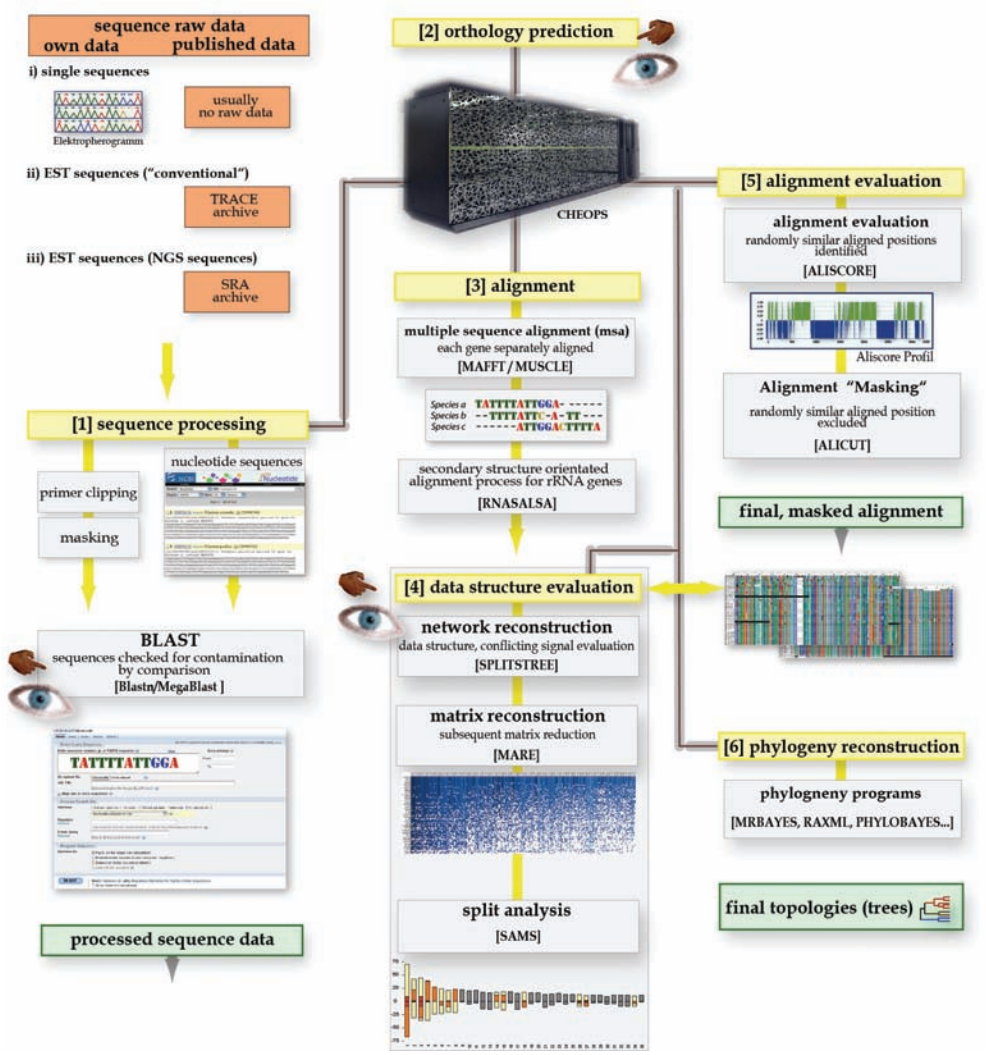

Fig. 4. Working flow of a typical phylogenetic analysis, which starts from scratch with the raw data (gained sequences) and ends with the final topology. Finger and eye symbols pinpoint crucial points to control not only the quality of the process, but also the data quality in the meaning of potential information or conflicts within gene sequences (data structure). A major aspect is, that large scale sequencing and phylogenomic data requires enormous computational power. Supercomputers (in this case CHEOPS: Cologne High Efficiency Operating Platform for Science, RRZK University of Cologne) or large cluster systems (ZFMK Bonn) are an essential requisite in the conducted analyses. Bold bars shaded in grey with internal brown lines symbolize circuit paths and represent steps that are constraint by computational limitations. Own sequence raw data and published data (orange) are processed and quality controlled

often difficult and dependent on single favourable unpredictable conditions. Thus, if anything goes wrong during sequencing, the loss may be irreversible. The second aspect is that samples must not be contaminated by other samples before and after sequencing. If contamination happens, it might not be detectable at all with desastrous consequences. This aspect must be integrated in process flows of sequencing facilities, for example by using tagging techniques applied on each library prior to sequencing to identify immediately eventual contamination. BLAST procedures against other processed project samples or libraries must be a second manadatory strategy.

3. Quality management during molecular analyses

For phylogenomic data the presented figure 4 illustrates only a rough scheme or framework of analysis. Depending on applied techniques and the choice of different software packages an adaptation is needed. Detailed descriptions of the working process to analyse rRNA and phylogenomic data with an emphasis on data quality are given in: von Reumont et al., (2009), von Reumont, (2010) and Meusemann et al., (2010).

[1] Sequences from different sources are processed in software pipelines, quality checked and controlled. It is problematic, that normally electropherograms are not available for published single sequences selected from public databases i). Therefore sequence errors cannot be discovered in these data. ii) EST sequences are normally stored in the TRACE archive in NCBI including the trace files. These represent the raw data and are in general not quality checked. iii) NGS raw data is stored in the Short Read Archive (SRA), which accounts for the difference of sequences from next generation sequencing to the 'conventional' EST sequences. [2] Respectively for the phylogenomic data the prediction of putative ortholog genes is eminent important. This step is computationally intensive and different approaches can be used, see paragraph 3.2. [3] Processed sequence data is aligned applying multiple sequence alignment programs. In case of rRNA genes a secondary structure-based alignment optimization is suggested. [4] A first impression of the data structure is gained by phylogenetic network reconstructions. That point becomes problematic with phylogenomic datasets comprising hundreds of genes and alignment sizes larger than 100 MB! Consequently, a method to evaluate the structure for these datasets could be the software MARE that reconstructs graphics of the data matrix based on the tree-likeness of single genes for each taxon (Misof & Meyer, 2011). Subsequently, a matrix reduction is possible after the alignment evaluation. [5] The final alignment evaluation and processing is applied for each gene with ALISCORE (Misof & Misof, 2009) to identify randomly similar aligned positions and those positions are subsequently excluded (=masking) by ALICUT (www.utilities.zfmk.de). Single, masked alignments are concatenated to the final alignment or supermatrix. A matrix reduction for phylogenomic datasets is performed applying MARE to enlarge the relative informativeness and to exclude genes that are uninformative (Misof & Meyer, 2001; www.mare.zfmk.de). For most analyses it could be useful to compare data structure before and after the alignment process in a network reconstruction or unreduced matrix [4]. Information content in respect of signal that supports different splits in the alignment can be visualized by SAMS (Wägele & Mayer, 2007). [6] After this the phylogenetic tree reconstruction is performed with several software packages.

3.1 The processed sequences and their quality

Most phylogenetic studies use own and published sequences in their analyses. However, in both cases a rigorous control of the quality of the sequence is crucial. This is conducted in

the steps of sequence processing (see figure 4, [1]). Different software tools guarantee quality by threshold value settings. A completely different aspect of quality is that the finally included sequence is indeed linked to the supposed species. Either misidentification of the specimen or the sequence can evoke serious bias in a subsequent analysis. If reaction in the laboratory were contaminated, the sequence is linked to the wrong species depending on the source of contamination. Both kinds of misidentification can be identified in general by careful BLAST procedures (Altschul et al., 1997, Kuiken & Corber, 1998). Yet, they are time intensive and in some cases difficult to interpret. For example, if you work with closely related species. In this case, the misidentification or contamination is rather impossible to detect, in particular if one species is unknown or only few or no sequences have been published. Other sources of data (like morphology) can also help to identify contamination (Wiens, 2004).
Several studies report that possible contaminations of taxa played a veritable role in studies, which proposed new evolutionary scenarios, but were actually based on contaminated sequences (von Reumont, 2010; Waegele et al., 2009; Koenemann et al., 2010). A careful control of sequence quality or a more critical interpretation of the reconstructed topologies could have prevented the (eventually repeated) inclusion of the contaminated sequences and subsequent publication of such suspicious phylogenetic trees. If contaminated sequences of older studies from rarely sequenced species are tacitly included into new analyses, this indeed can obscure phylogenetic implications. That is probably the case with the Mystacocarida, a crustacaean group with an still unclear phylogenetic position. They are rarely sequenced and the first and only published 18S rRNA sequence by Spears and Abele (1998) is very likely a contamination (von Reumont, 2010; Koenemann et al., 2010), which was impossible to identify for the authors in that study of 1998, which constituted the first larger analysis of crustaceans at all. A new study with completely sequenced 18S rRNA genes (von Reumont et al., 2009) including a new 18S rRNA gene sequence of the Mystacocarida revealed the contamination of the published sequence (von Reumont, 2010).
The search for contamination reaches a new dimension in phylogenomic data. A recent study (Longo et al., 2011) describes, that some non-primate genome databases, like the NCBI trace archive, provide sequences with human DNA contaminations, which can be traced back to pre-sequencing errors and/or low quality standards. Consequently, cross checking with published data might not help to be 100 percent sure about your own sequences. If you read the last sentence think about your own laboratory routines. Are they sufficient? If you outsource EST sequencing to an external company, which quality standard do they have and which risk management to handle possible contaminations?
This is respectively worrisome in cases of cross species analyses and genome analyses and indicates, that a better screening is generally needed (Phillips, 2011). The response of NCBI was, that trace archive data represents the raw data, which is not quality checked (http://www.ncbi.nlm.nih.gov/About/news/18feb2011.html). A careful processing of these sequences is obligate before analyses, including the control for possible contamination. An important conclusion is that every sequence from public databases should be treated suspiciously and a careful processing procedure is necessary to prevent errors by contamination. Do not trust your own data, but also do not trust public data.

3.2 Orthology prediction

Only homologous genes can be used in molecular phylogenetic studies. Homologous genes are further distinguished in two different classes: i) ortholog genes which originate in a single speciation event, and ii) paralog genes that originated from gene duplications

independently of speciation events (Fitch, 1970; Sonnhammer & Koonin, 2002; see review: Koonin, 2005). The prediction of ortholog genes in the era of large scale and next generation sequencing is a very delicate and computationally intensive process. An overview of commonly used methods for prediction of putative ortholog genes and their efficiency assessment is given in Roth et al. (2008) and Altenhoff and Dessimoz (2009).
A difficulty for phylogenetic reconstructions within arthropods is that only few data bases include sufficient numbers of complete arthropod genomes (Altenhoof & Dessimoz, 2009). INPARANOID and OMA are the two leading projects concerning the number of included arthropods. For that reason the orthology prediction for an arthropod dataset (Meusemann et al., 2010; von Reumont, 2010) and a further pancrustacean dataset (von Reumont et al., 2011) were based on INPARANOID 6 and 7 (Ostlund et al., 2010). Identified ortholog gene sets were extended using the HaMStR approach (Ebersberger et al., 2009) relying on the INPARANOID project. A set of orthologous genes was constructed using the InParanoid transitive closure (TC) approach in HaMStR described by Ebersberger et al. (2009). This set based on proteome data of so called 'primer taxa', which are completely sequenced genome species. Sequences of primer taxa were aligned within the set of orthologs and used to infer profile hidden Markov models (pHMMs). Subsequently, the pHMMs were used to search for putative orthologs among the translated ESTs of all taxa in the data set.
For the pancrustacean dataset pre-analyses were performed to compare the influence of using the OMA or INPARANOID projects with the same settings in HaMStR and the previous processing pipeline. For both analyses the same five primer taxa (*Aedes aegypti, Apis mellifera, Daphnia pulex, Ixodes scapulatis, Capitella* sp.) were used in HaMStR to train hidden markov models to extent the putative orthologs for all included taxa. Relying on OMA, 344 putative ortholog genes were identified in contrast to 1886 genes using INPARANOID. The resulting, reduced topologies (RAXML, -f, a, PROTCATWAG, 1000 BS) differ clearly in their resolution: the OMA based topology shows less resolution.
However, these results demonstrate the importance of further, more detailed studies on the impact of ortholog gene prediction. The quality of the trees might be severely influenced in this step of the analysis. A problem is the enormous computational power needed for comparative analysis of phylogenomic datasets.

3.3 Evaluation of data structure and data quality

All steps described so far are important to obtain in a standardized, rigorous processing high quality of the data and finally gene sequences, which are subsequently aligned and used for phylogenetic analyses.
The term *data quality*, however, addresses a different level of quality. A given multiple sequence alignment (MSA, synonymously often named data matrix) can include processed genes that are finally (after the processing procedure) of high quality, but for the phylogenetic goal to reconstruct a specific evolutionary history maybe not usable, if not informative. *Data quality* indeed refers to the scale of information or signal within the alignment. The term *data structure* is sometimes used synonymously to the term *data quality*.
Multiple substitution processes generally change sequences with time caused by random substitution processes, however, the extent of substitutions differs for parts of the DNA. In some parts of the DNA this substitution process erodes the former phylogenetic signal by multiple exchanges of nucleotides. After a long time nucleotides that represented synapomorphic characters to a sister taxon are by chance multiple substituted in the process

of signal erosion (Wägele & Mayer, 2007). By this process a different, random signal (noise) can arise, that in most cases is in conflict (and obscures) the historical, phylogenetic signal. In contrast, other genes are extremely conservative and nucleotides barely change with time. In this case a phylogenetic signal is hardly to detect either, caused by too few substitutions or synapomorphic characters. The mathematical substitution models, which are applied to reconstruct phylogenetic trees from multiple sequence alignments, try to implement several aspects of the briefly described processes. However, they are always an approximation and respectively are unable to differ between phylogenetic signal and noise. For further details see (Felsenstein, 1988; Wägele, 2005; Wägele & Mayer, 2007).

A first and fast evaluation of the structure in a dataset is feasible with network reconstructions, in which conflicts are visualized that are not illustrated by the (forced) bifurcations in phylogenetic trees (Holland et al., 2004; Huson & Bryant, 2006). It was the first time proposed by Bandelt and Dress (1992) to combine every phylogenetic analysis with a non-approximative method, which allows not compatible, alternative groupings contrary to bifurcting phylogenetic trees. One approach, the method of split decompositon, was developed by Bandelt and Dress (Bandelt & Dress, 1992). Hendy, Penny and Steel published a second method, the split analysis (Hendy & Penny, 1993; Hendy et al., 1994). Both methods work with so called bifurcations or splits.

A split is a couple of two groups of taxa, which are distinct subsets of the whole taxaset. Within the molecular phylogenetic context splits are distinguished by the occurence of nucleotide bases within sites. For a set of n taxa, exist 2^{n-1} possible bipartitions, in real datasets occur normally fewer splits. If there is only split signal for one unique dichotomous tree within a dataset, the number of splits is of the same value as the edges of a possible phylogeny. Given a taxon quartet (A, B), (C, D) few synapomophies between B and C can cause a split for second, alternatively supported topology (A, D) (B, C). This split migth not be visualized in a reconstructed tree-topology. Software packages offering non-approximate methods are SplitsTree (Huson & Bryant, 2006), Spectrum (Charleston, 1998), Spectronet (Huber et al., 2002) and SAMS (Wägele & Mayer, 2007).

SAMS is a software approach that was developed by Wägele and Mayer (2007) to perform a split analysis on the alignment. It accounts for all states of bases but analyses the columns of an alignment for occurring splits in a efficient way. Hence you can generate a split spectrum showing conflicting signal simultaneously obtaining a good overview on the data quality. Real splits are additionally differentiated from the conflicting ones. The method is currently under development, at the moment large datasets are difficult to analyze. Additionally, only nucleotide data is possible as input format. Further development is necessary and in progress to establish a new system, which evaluates all sites of an alignment and weights them according to contrast and homogeneity aspects to address these aspects.

Yet, network reconstruction and split analysis is limited by the size of a dataset and with larger or phylogenomic datasets still beyond abilities of available programs. Additionally, networks give only a rough overview and illustrate the present data structure, answering the question if a conflict or noise exists. More details are often not to analyze, for example which single genes or partitions create a conflict within an alignment. This part becomes additionally delicate handling 'supermatrices' that are composed of phylogenomic data.

Several strategies exist to handle 'supermatrices', which mostly are data sets with a large number of taxa and genes, but also missing information or gaps. Often, concatenated 'supermatrices' are filtered and reduced using predefined thresholds of data availability

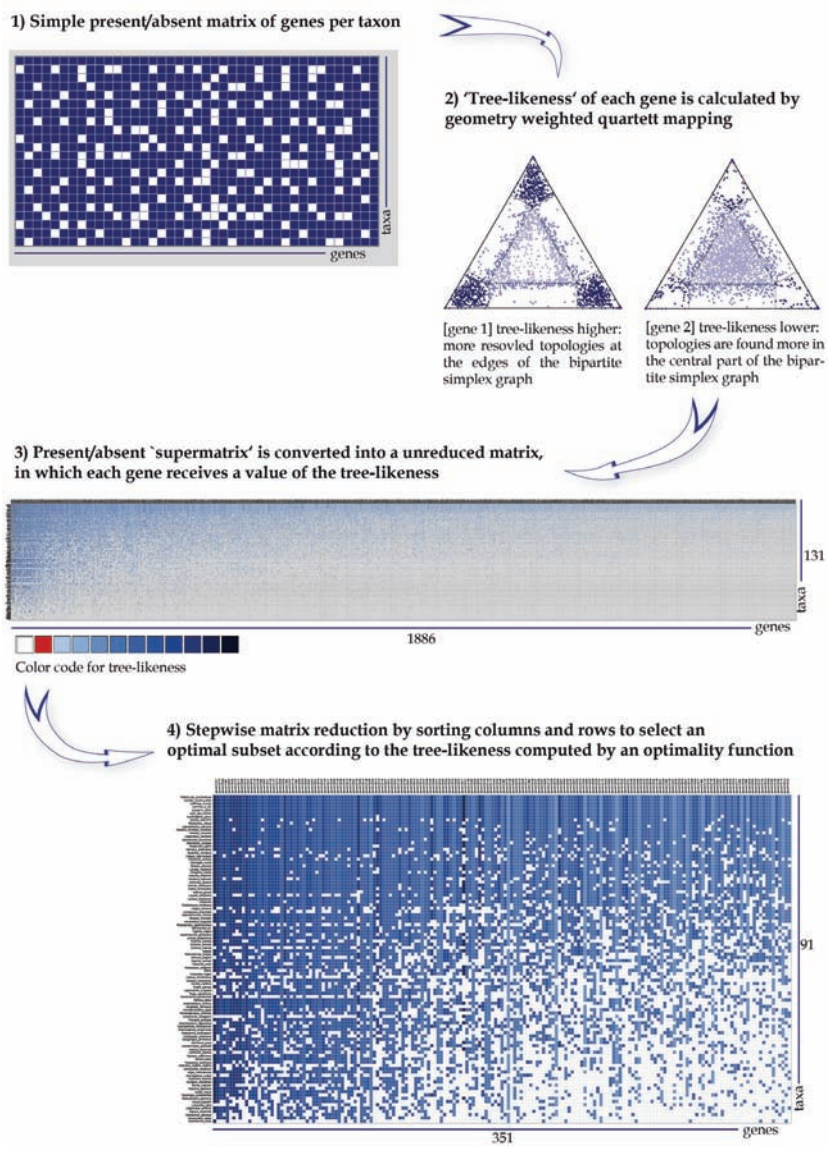

Fig. 5. Work flow of the MARE software. All genes are concatenated to a supermatrix, which is transformed into a `supermatrix' composed of all genes that are represented by tree-likness value. A tree-likeness is calculated in the step before via geometry weighteed quartet mapping. This supermatrix` is reduced by selecting an optimal subset of genes and taxa relying on the calculated value of the tree-likeness. The reduction is stepwise performed using an optimality function. The matrices composed of the tree-likeness values for each gene are colour coded. White symbolizes an absent gene, red a value of 0. From light to dark blue the value increases, dark blue represents a value of 0.9 -1.0

(Dunn et al., 2008; Philippe et al., 2009) depending on the relational number of present genes for a taxon. Taxa are excluded, if they are represented by less genes than accepted with the defined threshold value. Software tools like MARE are a first step to evaluate the data more detailed and enable an objective reduction of 'supermatrices' (large MSA´s of phylogenomic data), by selecting subsets of genes. MARE utilizes an alternative approach to data reduction selecting a subset of genes and taxa from a supermatrix based on information content and data availability (Meyer & Misof, 2010; http://mare.zfmk.de; Meusemann et al., 2010; von Reumont et al., 2011). The approach yields a condensed data set of larger information content by maximizing the ratio of signal to noise, and reducing uninformative genes or poorly sampled taxa.

MARE evaluates in a first step the 'tree-likeness' of each single gene. Tree-likeness reflects the relative number of resolved quartets for all possible (but not more than 20,000) quartets of a given sequence alignment or alignment partitions. The process is based on geometry-weighted quartet mapping (Nieselt-Struwe & von Haeseler, 2001), extended to amino acid data. For each gene a value for the tree-likeness is calculated by summarizing the support values for each of the three possible topologies during the quartet mapping procedure. After this step the previous present/absent matrix is changed to a matrix that contains values of tree-likeness for each gene per taxon. In the second step the matrix reduction is performed. The connectivity of the matrix (the gene and taxa overlap) is monitored during this step: two genes must have connection with at least three taxa. The matrix is reduced stepwise, with each reduction a new matrix is generated. Within each reduction step the column or row with the lowest information content (sum of values for tree-likeness) is excluded. The procedure is guided by an optimality function, which represents a trade off between matrix density and retained taxa and genes. For further details on the procedure and the algorithm, see: (Meyer & Misof, 2011; http://mare.zfmk.de).

4. Conclusions

When conducting or managing a project in molecular evolution use the available elements of project managing to prevent mistakes at this basic level. Important are the time schedule and milestones with sufficient backup time. A careful stakeholder analysis provides a detailed risk analysis, which is important in general, respectively if many persons or working groups are involved. Fieldtrips and appropriate preservation methods of the collected species must be carefully planned either, to start the molecular analysis with qualitative successful isolated material.

A process flow with a rigorous concept of quality control contributes to the quality of the gained sequences or data. The final sequences should have been checked for contamination. If techniques of next generation sequencing or expressed sequence tags are used, pay sufficient attention to select the best strategy for the prediction of ortholog genes. The aligned sequences should always be processed in the multiple sequence alignment for each gene or partition. Software like ALISCORE identifies randomly aligned alignment positions. Before the reconstruction of phylogenetic trees the *data quality* should be evaluated applying software to visualize the data structure and potential conflicts. Software for a more specific split analysis capable of larger data is e.g. SAMS, which is still under development. Assessing the data structure and quality is an essential strategy to identify conflict in phylogenetic trees or their eventual inability to reflect the 'real' evolutionary history of a species group.

Large data matrices or MSAs should be reduced to subsets, which were selected by the tree-likeness of each gene applying the software MARE. The software MARE is a first step to utilize objective criteria to select informative subsets of genes from a partially 'supermatrix'. However, several aspects are still to address further in future. Procedures of orthology prediction and matrix reduction need for example further investigation.

5. Acknowledgement

BMvR and SAM thank J-W Wägele for the chance and support to conduct the projects within the DMP framework. We would like to thank all colleagues who have been involved in the priority program SPP 1174 'Deep Metazoan Phylogeny' of the Deutsche Forschungsgemeinschaft (DFG) and the members of the molecular lab and Zentrum für molekulare Biodiversität (zmb) at the Zooloogischen Forschungsmuseum Alexander Koenig (ZFMK), Bonn. Respectively cooperation with Karen Meusemann was prosperous. Open discussions and exchange of experiences was extremely fruitful in all fields, not only the molecular area. Michael Kube from the Max Planck Institute of Molecular Biology and Genetics, Berlin, Germany gave eminent help and tips for the work with RNA. For detailed explanations and answers regarding the NGS projects we would like to thank colleagues from following companies: GATC, Konstanz, Germany and LGC, Berlin, Germany. The work for this manuscript is granted by the DFG proposals WA530/34, WA530/33.

6. References

Altschul, S. F.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids and Research,* 25, 3389-3402

Altenhoff, A. M. & Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods, *PLoS Computational Biology*, 5, 1

Bandelt, H. J. & Dress, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1:242-252.

Bouck, A. & Vision, T. (2007). The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology,* 16, 907-924

Bourne, L. (2010). Beyond reporting. The communication strategy, *PMI Global Congress Proceedings*, Melbourne, Australia

Budd, G.E & Telford, M.J. (2009). The origin and evolution of arthropods, *Nature*, 457, pp. 812-817

Charleston M. (1998). Spectrum: spectral analysis of phylogenetic data, *Bioinformatics (Oxford, England)* 14, 1, 98-9

Forster, J.L.; Harkin, V.B.; Graham, D.A. & McCullough, S.J. (2008). The effect of sample type, temperature and RNAlater (TM) on the stability of avian influenza virus RNA, *Journal of Virological Methods*, 149, pp. 190-194

Ebersberger, I.; Strauss, S. & Von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs, *BMC Evolutionary Biology*, 9, 157

Edgecombe, G.D. (2010). Arthropod phylogeny: An overview from the perspectives of morphology, molecular data and the fossil record, *Arthropod Structure and Development,* 39, pp. 74-87

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis, *Genome Research,* 8, 163-7

Ellegren, H. (2008). Sequencing goes 454 and takes large-scale genomics into the wild, *Molecular Ecology,* 17, 1629-1631.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22:521-565.

Fitch, W. M. (1970). Further improvements in the method of testing for evolutionary homology among proteins, *Journal of Molecular Biology,* 49, 1-14.

Freeman, E.R. (2010). *Strategic management: a stakeholder approach.* ISBN 978-0521151740, Cambridge University Press (first published by Pitman Publishing, 1984)

Gemeinholzer, B.; Droege, G.; Zetzsche, H.; Knebelsberger, T.; Raupach, M.; Borsch, T.; Klenk, H.-P.; Haszprunar, G. & Waegele; J.-W. (2011). The DNA Bank Network: the start from a German initiative. Biopreservation and Biobanking. April 2011, 9 (1):51-55, available at http://www.dnabank-network.org

Gorokhova, E. (2005). Effects on preservation and storage of microcrustacenas in RNAlater™ on RNA and DNA degradation, *Limnology and Oceanography: Methods*, 3, 143-148

Grotzer, M.A.; Pati, R.; Georger, B.; Eggert, A.; Chou, T.T. & Philips, P.C. (2000), Biological stability of RNA isolated from RNAlater™-treated brain tumor and neuroblastoma xenografts, *Medical Pediatric Oncology,* 34:438-442

Hemmrich, K.; Denecke, B.; Paul, N.E.; Hoffmeister, D. & Pallua, N., (2010). RNA Isolation from Adipose Tissue: An Optimized Procedure for High RNA Yield and Integrity, *Labmedicine*, 41 (2), pp 104-106

Hendy, M. & Penny, D., (1993). Spectral analysis of phylogenetic data. Journal of Classification, 10, 1, 5-24

Hendy, M., Penny, D. & Steel, M., (1994). A discrete Fourier analysis for evolutionary trees. Proceedings of the National Academy of Sciences of the United States of America, 91, 8, 3339-43

Holland, B. R.; Huber, K. T.; Moulton, V. & Lockhart, P. J. (2004). Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny, *Molecular Biology and Evolution,* 21, 1459-1461

Huber, K, Langton M, Penny D, Moulton V, & Hendy M., (2002). Spectronet: a package for computing spectra and median networks., Applied bioinformatics 1, 3, 159-61

Hudson, M. E., (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8, 3-17

Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution,* 23, 254-267

Jongeneel, C. V. (2000). Searching the expressed sequence tag (EST) databases: panning for genes. *Briefings in Bioinformatics* 1, 76-92.

Kerzner, H. (2009). *Project management: a systems approach to planning, scheduling and controlling*, ISBN 978-0470278703, John Wiley & Sons, 10th edition

Koenemann, S.; Jenner, R. A.; Hoenemann, M.; Stemme, T. & Von Reumont, B. M. (2010). Arthropod phylogeny revisited, with a focus on crustacean relationships, *Arthropod Structure and Development*, 39, 88-110

Koonin, E. (2005). Orthologs, paralogs and evolutionary genomics, *Annual Reviews of Genetics*, 39, 1, 209-338

Kuiken, C. & Korber, B. (1998). Sequence quality control, Los Alamos National Laboratory *HIV Compendium*, III, pp. 80-90

Litke, H.-D.; Kunow, I. & Schulz-Wimmer, H. (2010). *Projektmanagment,* ISBN 978-3-448-09949-2, Haufe-Lexware GmbH & Co. KG, Freiburg

Longo, M. S.; Longo, M. J.; O'Neill, R. J. & O'Neill (2011). Abundant Human DNA Contamination Identified in Non-Primate Genome Databases, *PLoS ONE,* 6, 2, e16410. doi:10.1371/journal.pone.0016410

Meusemann, K.; Von Reumont, B. M.; Simon, S.; Roeding, F.; Strauss, S.; Kuck, P.; Ebersberger, I.; Walzl, M.; Pass, G.; Breuers, S.; Achter, V.; Von Haeseler, A.; Burmester, T.; Hadrys, H.; Wagele, J. W. & Misof, B. (2010). A phylogenomic approach to resolve the arthropod tree of life. *Molecular Biology and Evolution* 27, 2451-64.

Meyer B. & Misof, B. (2011). MARE: Matrix Reduction – A tool to select optimized data subsets from supermatrices for phylogenetic inference. Zentrum für molekulare Biodiversitätsforschung (zmb) am ZFMK, Adenauerallee 160, 53113 Bonn, Germany, http://mare.zfmk.de

Misof, B. & Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion, *Systematic Biology*, 58, 1

Mülhardt, C. (2008). *Der Experimentator: Molekularbiologie/Genomics*, Spektrum Akademischer Verlag, 6. Auflage. ISBN-10: 9783827420367

Mutter, G.L.; Zahrieh; D., Liu; C.M.; Neuberg, D.; Finkelstein, D.; Baker, H.E. & Warrington, J.A. (2004). Comparison of frozen and RNAlater™ solid tissue storage methods for use in RNA expression microarrays, *BMC Genomics*, 5:88

Nieselt-Struwe K. & Von Haeseler A. (2001). Quartet-mapping, a generalization of the likelihood-mapping procedure. *Molecular Biology and Evolution* 18:1204-1219

Ostlund, G.; Schmitt, T.; Forslund, K.; Köstler, T.; Messina, D. N.; Roopra, S.; Frings, O. & Sonnhammer, E. L. L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, Nucleid Acid Research, 38

Palumbi, S. R. (1996). Nucleic acids II: The Polymerase Chain Reaction, in: Molecular Systematics, Hillis, D. M., Moritz, C., Mable, B. K. 2nd edition, Sinauer Associates, ISBN 978-0878932825

Petterson, E.; Ludneber, J. & Ahmadian, A. (2009). Generations of sequencing technologies, *Genomics*, 93, pp. 105-111

Philippe, H.; Delsuc, F.; Brinkmann, H. & Lartillot, N. (2005). Phylogenomics, *Annual Review of Ecology and Evolutionary Systematics,* 36, 541-562

Philippe H; Derelle R; Lopez P; Pick, K.; Borchiellini, C.; Boury-Esnault, N.; Vacelet, J.; Renard, E.; Houliston, E.; Quéinnec, E.; Da Silva, C.; Wincker, P.; Le Guyader, H.; Leys, S.; Jackson, D. J.; Schreiber, F.; Erpenbeck, D.; Morgenstern, B.; Wörheide, G.

& Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Curr Biol*. 19:706-712.

Phillips, M.L. (2011). Contamination of non-primate DNA archives with human sequences indicates that better screening is needed, *nature news*, doi:10.1038/news.2011.99

Ronaghi, M. (2001). Pyrosequencing Sheds Light on DNA Sequencing, *Genome Research,* 11, pp. 3-11

Sambrook, J. & Russel, D. W. (2000). Molecular Cloning: A laboratory manual, 3rd reprint, ISBN 978-0879695774

Shendure, J.; Mitra, R.; Varma, C. & Church, G. (2004). Advanced sequencing technologies: methods and goals, *Nature Reviews in Genetics,* 5, pp. 335-344.

Sonnhammer, E. L. L. & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes, Trends in Genetics, 18, 12, 619-620

Spears, T. & Abele, L. G. (1998). Crustacean phylogeny inferred from 18S rDNA, In *Arthropod Relationships*, editors: R. A. Fortey and R. H. Thomas, ISBN 978-0412754203, Chapman and Hall, pp. 169-187, London

Thornton, J. W. & Desalle, R. (2000). Gene family evolution and homology: genomics meets phylogenetics, *Annual Reviews of Genomics and Human Genetics,* 1, 41-73

Vink, C.J.; Thomas, S.M.; Paquin, P.; Hayashi, C.Y. & Hedin, M. (2005). The effects of preservatives and temperatures on arachnid DNA, *Invertebrate Systematics*, 19, pp. 99-104

Voelkerding, K. V.; Dames, S. A. & Durtschi, J. D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics, *Clinical Chemestry*, 55, pp. 641-658

Von Reumont, B. M.; Meusemann, K.; Szucsich, N.; Dell'ampio, E.; Gowri-Shankar, V.; Bartel, D.; Simon, S.; Letsch, H. O.; Stocsits, R. R.; Luan, Y. X.; Wägele, J. W.; Pass, G.; Hadrys, H. & Misof, B. (2009). Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships, *BMC Evolutionary Biology* 9, 119.

Von Reumont, B. M. (2010). *Molecular insights to crustaecan phylogeny. A status quo of past, present and perspective prospects also covering phylogenomics,* ISBN 978-3-8381-1770-6, Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken, Germany.

Von Reumont, B. M.; Jenner, R. A.; Wills, M. A.; Dell´Ampio, E.; Pass, G.; Ebersberger, I.; Meusemann, K.; Meyer, B.; Koenemann, S.; Iliffe, T. I.; Stamatakis, A.; Niehuis, O. & Misof, B. (2011). Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as a sister group to Hexapoda, accepted with minor revisions, in re-prep for MBE

Weaver, P. (2007). A Simple View of Complexity in Project Management, *Proceedings of the 4th World Project Management Week, Singapore*

Wiens, J. (2004). The Role of Morphological Data in Phylogeny Reconstruction, *Systematic Biology*, 53, 653-661

Wägele, J.-W. (2005). *Foundations of phylogenetic systematics*, ISBN-13: 9783899370560, Friedrich Pfeil Verlag, München

Wägele. J.-W. & Mayer, C. (2007). Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects, *BMC Evolutionary Biology*, 7, 147

Wägele, J. W.; Letsch, H.; Klussmann-Kolb, A.; Mayer, C.; Misof, B. & Wagele, H. (2009). Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny), *Frontiers in Zoology,* 6, 12