

Молекулярные методы исследования биологической эволюции и их информационное обеспечение

Д.Ю.Щербаков

February 18, 2016

1 teaser

2 Введение: Определения

- Сравнение последовательностей: Статистика

3 Некодирующая ДНК

- Общая характеристика
- кодирующая и нкДНК в эволюции
- Отбор и некодирующие ДНК
 - Тесты на нейтральность, применимые к нкДНК
- Консервативная нкДНК
- Структура РНК и эволюция нкДНК
- Пример
- Перспективы

4 Оценка достоверности филогенетических гипотез

- ABC
- Скитания полихет
- Гастроподы: перемена ниш и межвидовая гибридизация

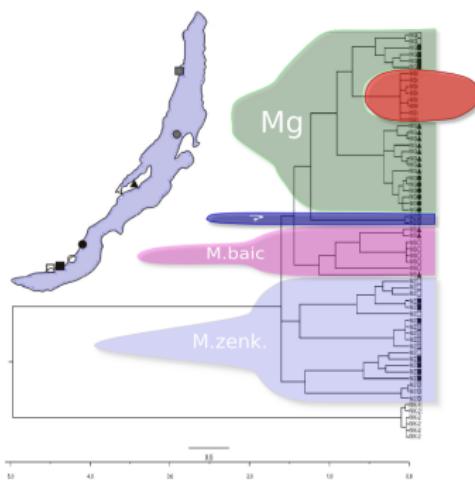
5 NGS и поп. генетика

- Мотивация
- Сцена

Пресноводные полихеты рода *Manayunkia*



Manayunkia baicalensis



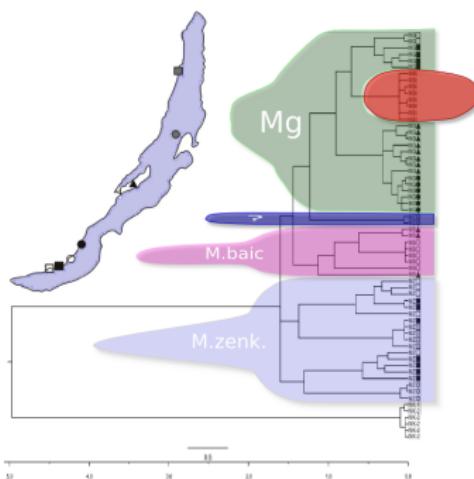
Филогения на основании мит. ДНК:

Байкальские полихеты распадаются как минимум на 3 (может быть - четыре) монофилетические клады, соответствующие трем морфологическим видам *Manayunkia*, описанным ранее.

Пресноводные полихеты рода *Manayunkia*



Manayunkia baicalensis



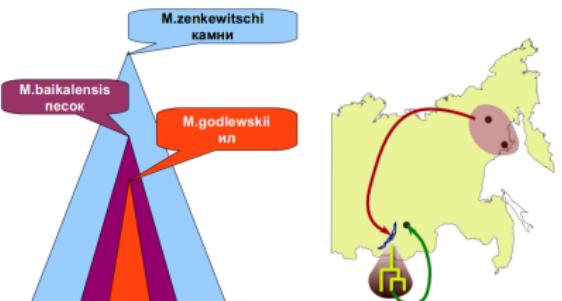
Филогения на основании мит. ДНК:

Байкальские полихеты распадаются как минимум на 3 (может быть - четыре) монофилетические клады, соответствующие трем морфологическим видам *Manayunkia*, описанным ранее.

Три вида или три экологические формы?

Различия между байкальскими манаюнками

Основные различия между байкальскими видами манаюнкий состоят в субстратных предпочтениях. Поэтому можно предположить: **вилообразование у полихет произошло путем адаптации к разным субстратам**



Возможные механизмы видообразования *Manayunkia*

- **H1**Переключение субстратов происходило единожды
- **H2**Преключение носило множественный характер и параллельно происходило в разных частях Байкала

Результаты байесовского сравнения гипотез с использованием митохондриальных последовательностей

H1 Митохондриальная филогения содержит сильную поддержку гипотезы о монофилии - т.е. о единственном переключении.

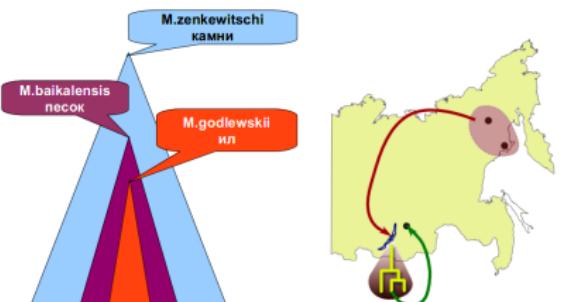
Три вида или три экологические формы?

Различия между байкальскими манаюнками

Основные различия между байкальскими видами манаюнкий состоят в субстратных предпочтениях. Поэтому можно предположить: **вилообразование у полихет произошло путем адаптации к разным субстратам**

Возможные механизмы видообразования *Manayunkia*

- **H1**Переключение субстратов происходило единожды
- **H2**Преключение носило множественный характер и параллельно происходило в разных частях Байкала



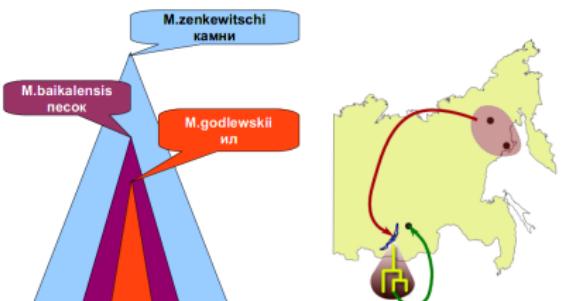
Результаты байесовского сравнения гипотез с использованием митохондриальных последовательностей

H1 Митохондриальная филогения содержит сильную поддержку гипотезы о монофилии - т.е. о единственном переключении.

Три вида или три экологические формы?

Различия между байкальскими манаюнками

Основные различия между байкальскими видами манаюнкий состоят в субстратных предпочтениях. Поэтому можно предположить: **вилообразование у полихет произошло путем адаптации к разным субстратам**



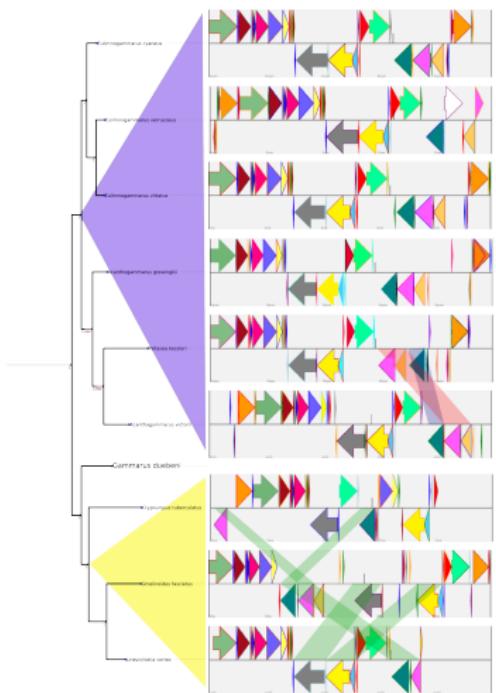
Возможные механизмы видообразования *Manayunkia*

- **H1**Переключение субстратов происходило единожды
- **H2**Преключение носило множественный характер и параллельно происходило в разных частях Байкала

Результаты байесовского сравнения гипотез с использованием митохондриальных последовательностей

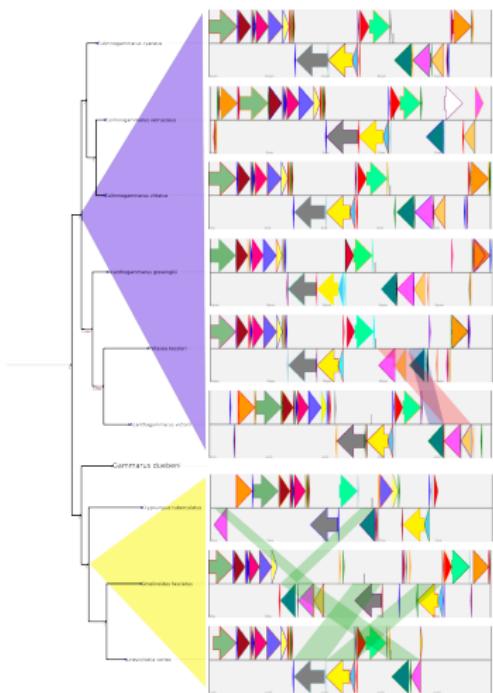
H1 Митохондриальная филогения содержит сильную поддержку гипотезы о монофилии - т.е. о единственном переключении.

Gene order



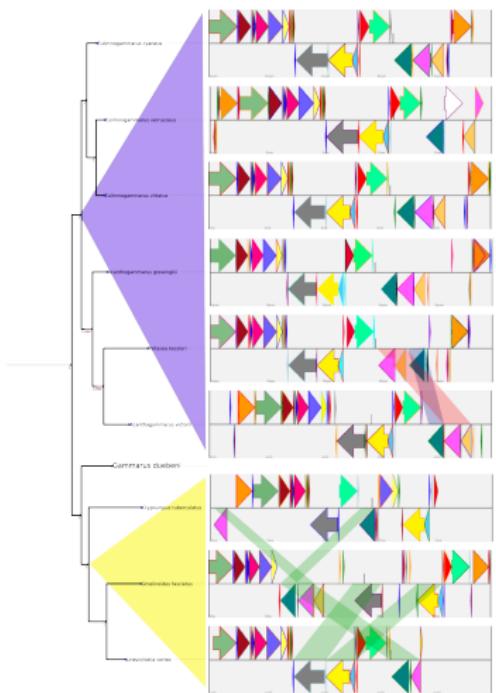
- Gene order rearranged in both major lineages;
- Rearranged parts of genome always keep their orientation. Not a single violation was found;
- Every reshuffled genome differs from the base "Gammarus" pattern by many steps;
- All reshuffled genomes belong to the shallow water dwellers
- The range of genome re-arrangements is much higher in Baikal in comparison to other amphipods where mostly tRNA genes are involved

Gene order



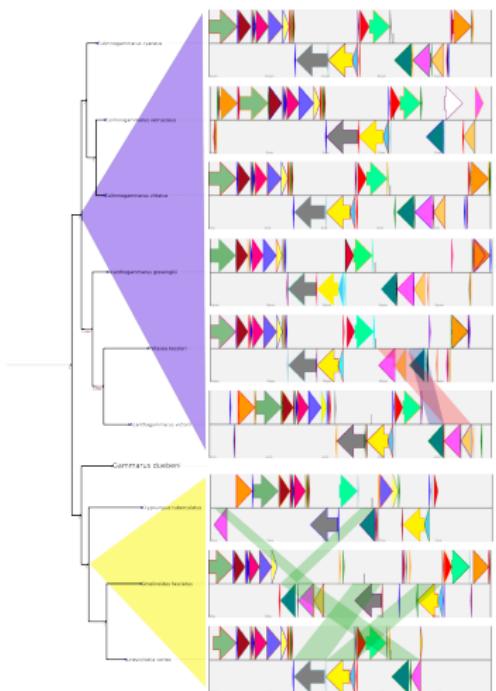
- Gene order rearranged in both major lineages;
- Rearranged parts of genome always keep their orientation. Not a single violation was found;
- Every reshuffled genome differs from the base "Gammarus" pattern by many steps;
- All reshuffled genomes belong to the shallow water dwellers
- The range of genome re-arrangements is much higher in Baikal in comparison to other amphipods where mostly tRNA genes are involved

Gene order



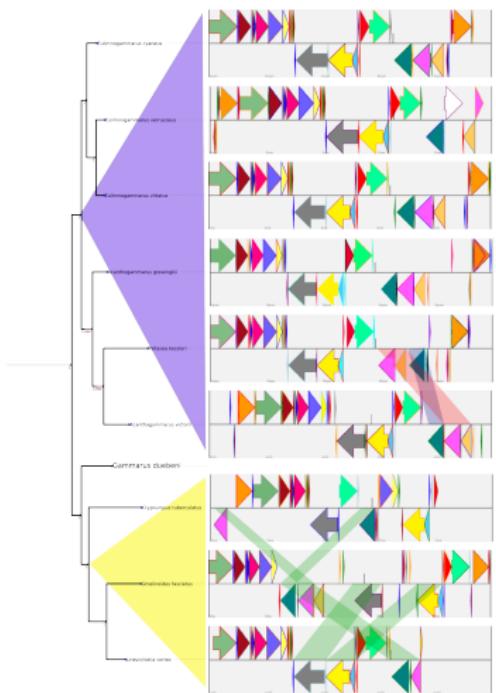
- Gene order rearranged in both major lineages;
- Rearranged parts of genome always keep their orientation. Not a single violation was found;
- Every reshuffled genome differs from the base "Gammarus" pattern by many steps;
- All reshuffled genomes belong to the shallow water dwellers
- The range of genome re-arrangements is much higher in Baikal in comparison to other amphipods where mostly tRNA genes are involved

Gene order



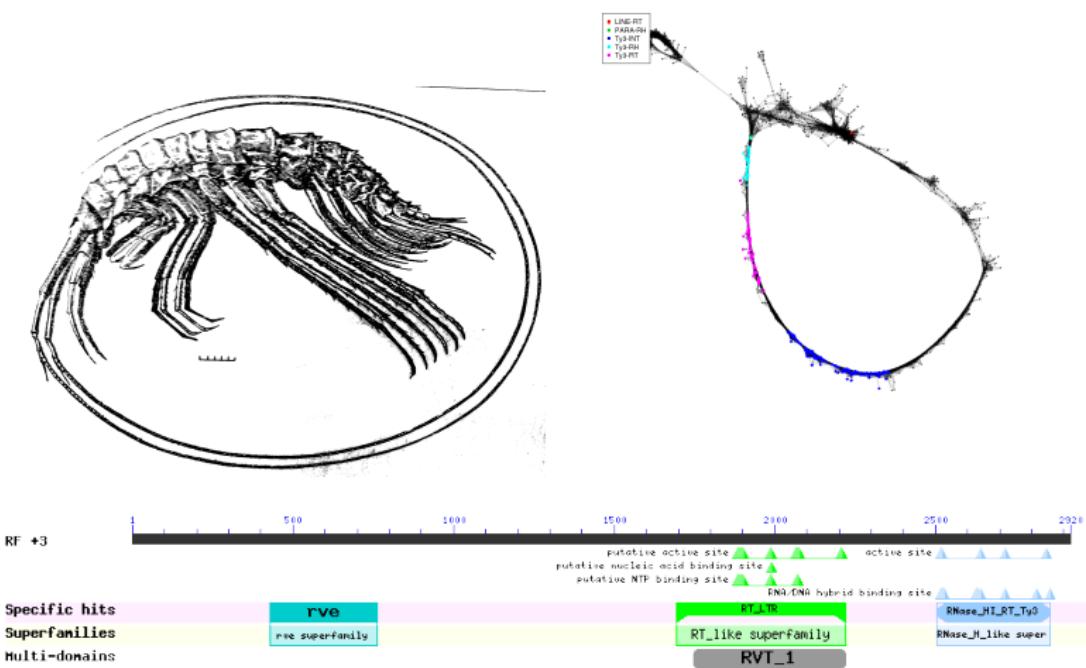
- Gene order rearranged in both major lineages;
- Rearranged parts of genome always keep their orientation. Not a single violation was found;
- Every reshuffled genome differs from the base "Gammarus" pattern by many steps;
- All reshuffled genomes belong to the shallow water dwellers
- The range of genome re-arrangements is much higher in Baikal in comparison to other amphipods where mostly tRNA genes are involved

Gene order



- Gene order rearranged in both major lineages;
- Rearranged parts of genome always keep their orientation. Not a single violation was found;
- Every reshuffled genome differs from the base "Gammarus" pattern by many steps;
- All reshuffled genomes belong to the shallow water dwellers
- The range of genome re-arrangements is much higher in Baikal in comparison to other amphipods where mostly tRNA genes are involved

LTR retro-element example



Филогенетика

Филогенетика

Исследование эволюционных взаимоотношений между группами организмов.

Термин происходит от греческих слов $\varphiυλη$ (племя, клан или раса) и $\gammaενεσις$ (происхождение от). **Объекты филогенетического анализа называются ОТЕ - операционными таксономическими единицами**

Метод

Методом филогенетики является филогенетический анализ на основе сравнения либо матриц морфологических признаков, либо - последовательностей нуклеиновых кислот или белков.

Результат

Результатом филогенетического анализа является филогенетическая гипотеза, предполагающая эволюционный сценарий, приведший к формированию исследуемой группы организмов.

Филогенетика

Филогенетика

Исследование эволюционных взаимоотношений между группами организмов.

Термин происходит от греческих слов $\varphiυλη$ (племя, клан или раса) и $\gammaενεσις$ (происхождение от). **Объекты филогенетического анализа называются ОТЕ - операционными таксономическими единицами**

Метод

Методом филогенетики является филогенетический анализ на основе сравнения либо матриц морфологических признаков, либо - последовательностей нуклеиновых кислот или белков.

Результат

Результатом филогенетического анализа является филогенетическая гипотеза, предполагающая эволюционный сценарий, приведший к формированию исследуемой группы организмов.

Филогенетика

Филогенетика

Исследование эволюционных взаимоотношений между группами организмов.

Термин происходит от греческих слов *φυλη* (племя, клан или раса) и *γενεσις* (происхождение от). Объекты филогенетического анализа называются ОТЕ - операционными таксономическими единицами

Метод

Методом филогенетики является филогенетический анализ на основе сравнения либо матриц морфологических признаков, либо - последовательностей нуклеиновых кислот или белков.

Результат

Результатом филогенетического анализа является филогенетическая гипотеза, предполагающая эволюционный сценарий, приведший к формированию исследуемой группы организмов.

Еще несколько определений

Ниша

потребности вида для обеспечения положительной скорости роста численности

Фундаментальная ниша

набор абиотических факторов, обеспечивающих положительную скорость роста численности особей данного вида

Консерватизм ниши

стремление родственных видов занимать сходные ниши; а также - долговременная стабильность ниши, ее стремление оставаться стабильной в течение длительных промежутков времени (т.е. высокая временная автокорреляция).

Климатическая ниша

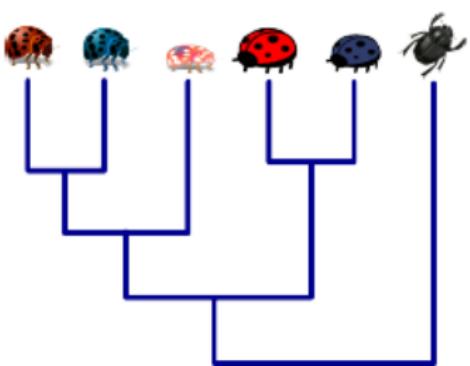
Аспект средовой ниши, который определяется пределами климатических вариаций Вне пределов этой ниши популяция не может поддерживать положительную скорость роста своей численности

Стазис ниши

Отсутствие каких-либо изменений параметров ниши



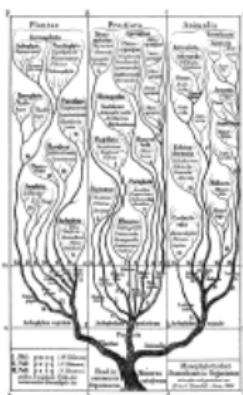
Эволюционное дерево



Обычно филогенетическую гипотезу представляют в виде эволюционного дерева. Его можно использовать для классификации, но оно не обязательно является таксономическим деревом

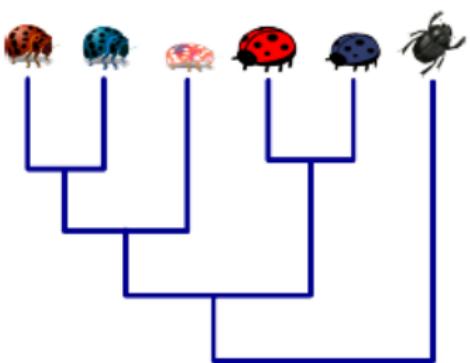
Важно помнить:

На филогенетическом дереве современные ОТЕ находятся только на терминальных ветвях. Они никогда не бывают во внутренних узлах!



Дерево-классификация живого мира, предложенное Геккелем. Основано на интуитивном представлении об эволюции. Отражает скорее сходство между организмами, история их происхождения друг от друга в данном случае вторична по отношению к систематике

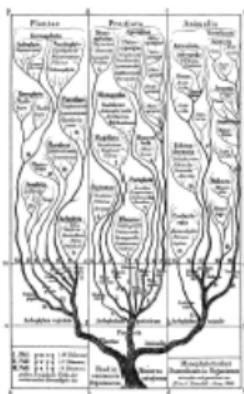
Эволюционное дерево



Обычно филогенетическую гипотезу представляют в виде эволюционного дерева. Его можно использовать для классификации, но оно не обязательно является таксономическим деревом

Важно помнить

На филогенетическом дереве современные ОТЕ находятся только на терминальных ветвях. Они никогда не бывают во внутренних узлах!

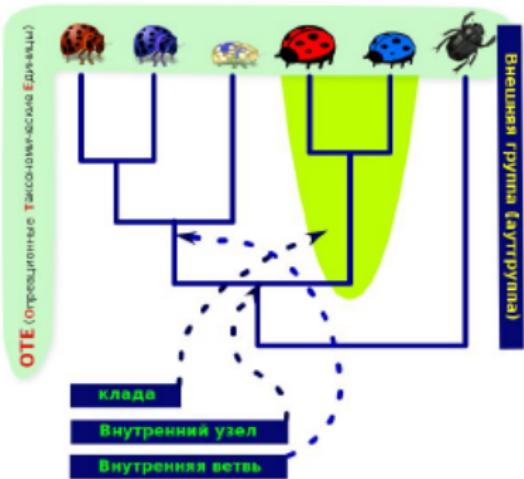


Дерево-классификация живого мира, предложенное Геккелем. Основано на интуитивном представлении об эволюции. Отражает скорее сходство между организмами, история их происхождения друг от друга в данном случае вторична по отношению к систематике

Эволюционное дерево

Эволюционное дерево

Граф, соединяющий объекты и отражающий историю их появления от общего предка



Эволюционное дерево

полностью разрешенное дерево

дерево, у которого каждый из внутренних узлов имеет не только три ветви, две из которых ведут к потомкам , одна - к предку. Такое дерево называется также **бинарным**

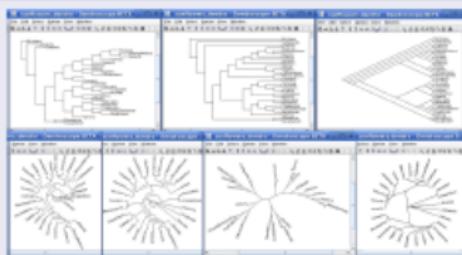
ПОЛИТОМИЯ

это явление, когда из одного узла выходят более двух ветвей. Дерево, содержащее как минимум одну политомию называется **не полностью разрешенным**

типы политомий

- ❶ **мягкая политомия** возникает в том случае, когда не хватает признаков для того, чтобы на полностью разрешенном дереве каждая из ветвей поддерживалась изменением состояния хотя бы одного информативного признака;
- ❷ **жесткая политомия** не зависит от количества используемых в анализе признаков и отражает биологическую природу явлений.

типы представления деревьев

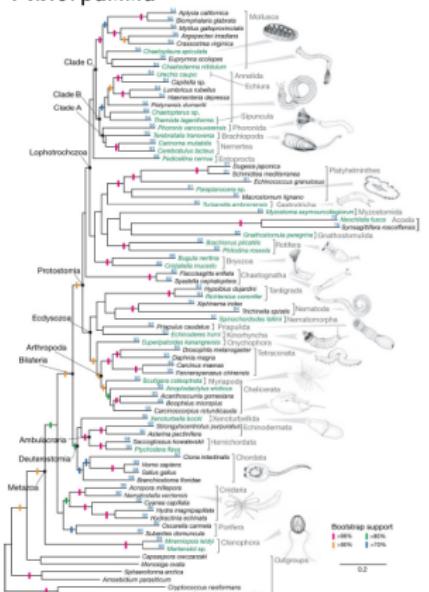


- ❶ прям угольная филограмма
- ❷ прям угольная кладограмма
- ❸ наклонная кладограмма
- ❹ круговая филограмма
- ❺ круговая кладограмма
- ❻ радиальная филограмма
- ❼ радиальная кладограмма

(<http://ab.inf.uni-tuebingen.de/software/dendroscope/>)

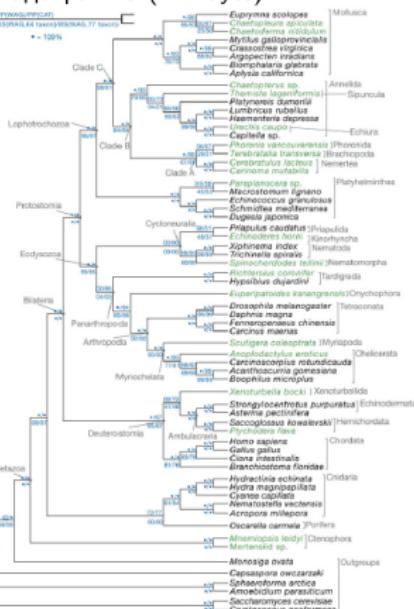
Эволюционное дерево

Филограмма

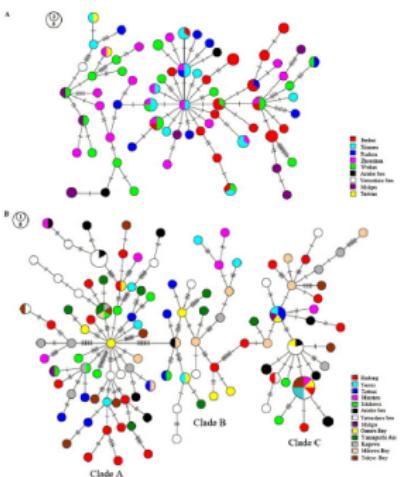


иа "Broad phylogenomic sampling improves resolution of the animal tree of life", CW Dunn, A Hejnol, DQ Matus, K Pang, WE Browne, SA Smith, E Seaver, GW Rouse, M Obst, GD Edgecombe, MV Sørensen, SHD Haddock, A Schmidt-Rhaesa, A Okusu, RM Kristensen, WC Wheeler, MQ Martindale & G Giribet *Nature* 452, 745-749 (10 April 2008)

Кладограмма (mrBayes)



Эволюционное дерево



Jin-Xian Liu, Tian-Xiang Gao, Koji Yokogawa, Ya-Ping Zhang. Differential population structuring and demographic history of two closely related fish species, Japanese sea bass (*Lateolabrax japonicus*) and spotted sea bass (*Lateolabrax maculatus*) in Northwestern Pacific (2006) Mol Phylog Evol 39: 799-811

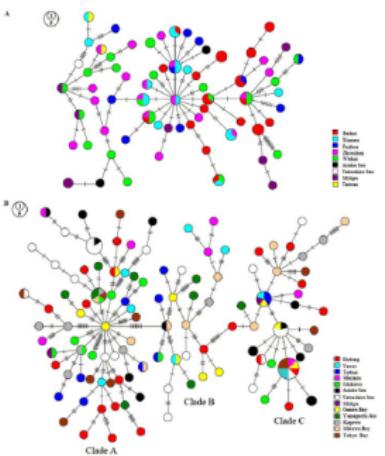
Простирающееся дерево

или минимальное простирающееся дерево (*minimal spanning tree*) - граф минимальной длины, связывающий ОТЕ. Длина ветвей простирающегося дерева измеряется в мутационных шагах.

- предки сосуществуют с потомками и полярность их взаимоотношений можно только предполагать. Например - вирусы, бактерии
 - при исследованиях либо внутривидового полиморфизма, либо - полиморфизма в рамках близкородственных видов, когда велика вероятность эффектов неполного разделения предковых линий и неясны взаимоотношения ОТЕ предков-потомкам.

В отличие от филогенетического анализа, где все современные ОТЕ обязаны располагаться на терминальных ветвях, при построении простирающихся деревьев задача состоит в том, чтобы максимальное количество ОТЕ расположить во внутренних узлах дерева

Эволюционное дерево



Jin-Xian Liu, Tian-Xiang Gao, Koji Yokogawa, Ya-Ping Zhang. Differential population structuring and demographic history of two closely related fish species, Japanese sea bass (*Lateolabrax japonicus*) and spotted sea bass (*Lateolabrax maculatus*) in Northwestern Pacific (2006) Mol. Phyl. Evol. 39: 799–811

Простирающееся дерево

или минимальное простирающееся дерево (minimal spanning tree) - граф минимальной длины, связывающий ОТЕ. Длина ветвей простирающегося дерева измеряется в мутационных шагах.

Применение

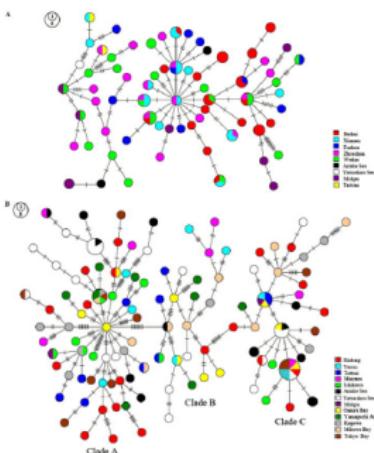
Простирающиеся деревья используются в случаях, когда

- предки сосуществуют с потомками и полярность их взаимоотношений можно только предполагать.
Например - вирусы, бактерии
- при исследованиях либо внутривидового полиморфизма, либо - полиморфизма в рамках близкородственных видов, когда велика вероятность эффектов неполного разделения предковых линий и неясны взаимоотношения ОТЕ предок-потомок.

Важно помнить!

В отличие от филогенетического анализа, где все современные ОТЕ обязаны располагаться на терминальных ветвях, при построении простирающихся деревьев задача состоит в том, чтобы максимальное количество ОТЕ расположить во внутренних узлах дерева

Эволюционное дерево



Jin-Xian Liu, Tian-Xiang Gao, Koji Yokogawa, Ya-Ping Zhang. Differential population structuring and demographic history of two closely related fish species, Japanese sea bass (*Lateolabrax japonicus*) and spotted sea bass (*Lateolabrax maculatus*) in Northwestern Pacific (2006) Mol. Phyl. Evol. 39: 799–811

Простирающееся дерево

или минимальное простирающееся дерево (minimal spanning tree) - граф минимальной длины, связывающий ОТЕ. Длина ветвей простирающегося дерева измеряется в мутационных шагах.

Применение

Простирающиеся деревья используются в случаях, когда

- предки сосуществуют с потомками и полярность их взаимоотношений можно только предполагать. Например - вирусы, бактерии
- при исследованиях либо внутривидаового полиморфизма, либо - полиморфизма в рамках близкородственных видов, когда велика вероятность эффектов неполного разделения предковых линий и неясны взаимоотношения ОТЕ предок-потомок.

Важно помнить!

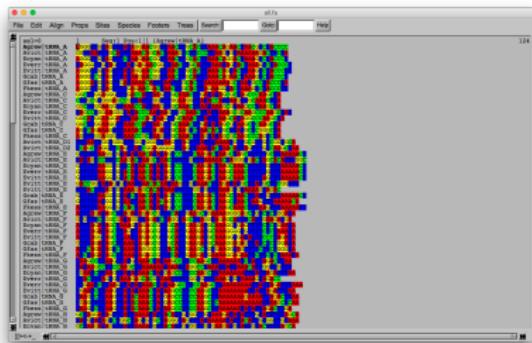
В отличие от филогенетического анализа, где все современные ОТЕ обязаны располагаться на терминальных ветвях, при построении простирающихся деревьев задача состоит в том, чтобы максимальное количество ОТЕ расположить во внутренних узлах дерева

Задача

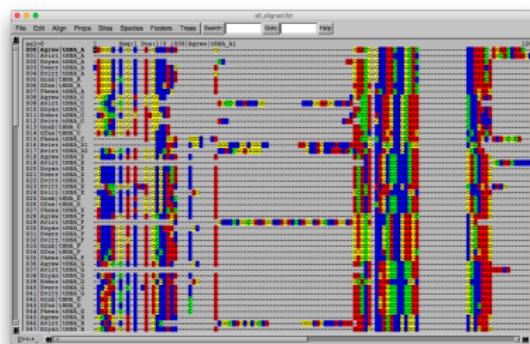
Выравнивание последовательностей

Имеет целью предложить наиболее вероятную схему последовательностей, в которой гомологичные основания (аминокислотные остатки) располагались бы друг под другом. Для этой цели в последовательности вносят минимальное количество делеций или вставок (инделей)

Невыровненные последовательности



Выровненные последовательности



E-value

Для двух достаточно длинных последовательностей нуклеиновых кислот или полипептидов можно показать, что

$$E = Kmne^{-\lambda S} \quad (1)$$

где m и n - длины сравниваемых последовательностей, K и λ - параметры, S - «счет» сходства. В простейшем случае это – число совпадающих позиций. E - вероятность получить степень сходства равную или превосходящую S случайно. Именно эта величина и называется **E-value** и приводится для оценки, например, результатов поиска гомологичных последовательностей в Генбанке

Без знания механизма определения сходства разных позиций последовательности (то есть значений K и λ) значение E не имеет особого смысла. Поэтому была предложена **побитовая мера сходства последовательностей**, которая определяется как

$$S' = \frac{\lambda S - \ln K}{\ln 2} \quad (2)$$

и тогда уравнение (1) превращается в

$$E = mn2^{-S'} \quad (3)$$

Некоторые функции некодирующей ДНК

некодирующая ДНК

это вся ДНК, которая не кодирует белков

Разновидности некодирующей ДНК

- Транскрибируемая некодирующая ДНК:
 - служит матрицей для рРНК, тРНК и различных малых РНК;
 - интроны и различные цис-регулирующие элементы
 - По сообщениям нескольких групп, исследующих транскрипцию человеческого генома, процент транскрибируемой ДНК может существенно превышать 60%.
 - Нетранскрибируемая ДНК включает теломеры, некоторые высокоповторенные последовательности и т.п.



Некоторые функции некодирующей ДНК

некодирующая ДНК

это вся ДНК, которая не кодирует белков

Разновидности некодирующей ДНК

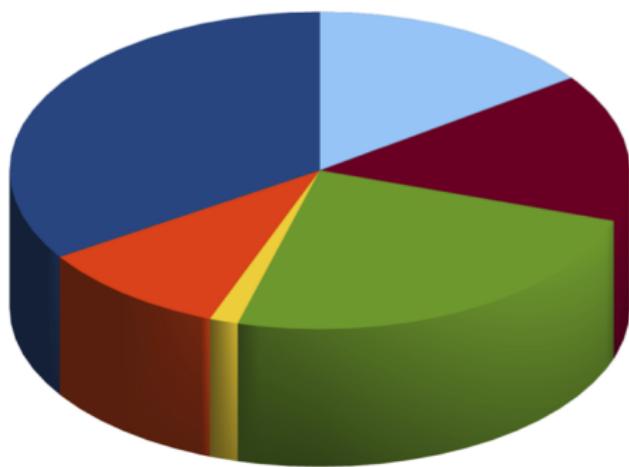
- Транскрибируемая некодирующая ДНК:
 - служит матрицей для рРНК, тРНК и различных малых РНК;
 - интроны и различные цис-регулирующие элементы
 - По сообщениям нескольких групп, исследующих транскрипцию человеческого генома, процент транскрибируемой ДНК может существенно превышать 60%.
- Нетранскрибируемая ДНК включает теломеры, некоторые высокоповторенные последовательности и т.п.

Следовательно:

Оправдано предположение о том, что большая часть нкДНК проявляет активность через РНК

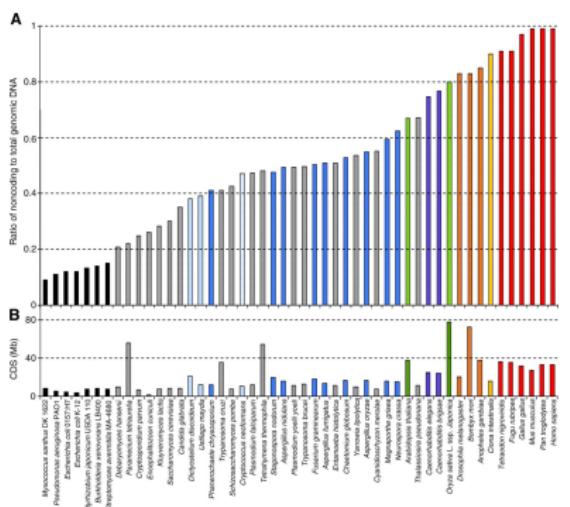
Кодирующая ДНК составляет около 2% генома человека

Состав человеческого генома



- повторяющиеся ДНК, включая транспозоны и похожие на них элементы
 - Alu повторы
 - Экзоны (включая рРНК и тРНК)
 - Интроны и регуляторные последовательности
 - Уникальная некодирующая ДНК
 - Повторяющаяся ДНК не связанная с транспозонами

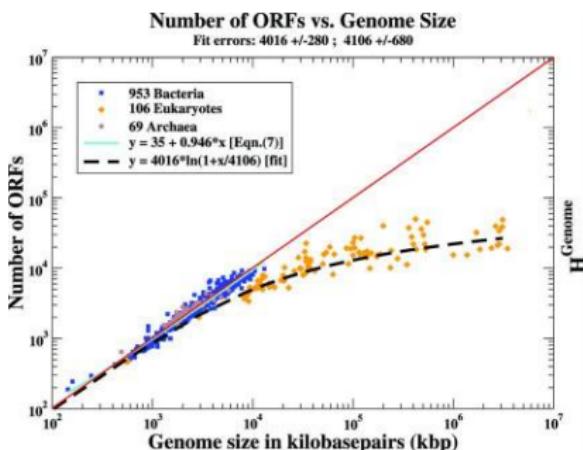
Доля некодирующей ДНК в геномах



(John S. Mattick 2007. A new paradigm for developmental biology J Exp Biol 210, 1526-1547)

Функция большей части некодирующей РНК

Управление активностью генов



Friar JL, Goldman T, Pérez-Mercader J. 2012 Genome sizes and the Benford distribution. PLoS One. 7(5):e36624.

$$y_{ORF}(G) = \{y_{ORF}^{(min)} - A \ln(1 + \frac{G^{(min)}}{B})\} + A \ln 1 + \frac{G}{B}$$

где $YORF$ – число открытых рамок ч-считывания, $G^{(min)}$ – минимальный возможный размер генома, G – размер генома

Мутации некодирующей ДНК и жизнеспособность

Последствия появления не-нейтральных аллелей в популяции

- В пределах одной популяции
 - снижение степени генетического разнообразия
 - сдвиг спектра частот аллелей
 - неравновесие по сцеплению
- В сравнении с другими популяциями
 - повышенное значение F_{ST}
 - повышенное число фиксированных замен

Тесты нейтральности

Во всех тестах нуль-гипотеза состоит в нейтральности аллеля, а альтернативная

- в том, что он находится под давлением отбора

- однолокусные
- многолокусные

количественные характеристики генетического разнообразия

Нуклеотидное разнообразие

$$\Pi = \frac{n}{n-1} \sum_{ij} x_i x_j \pi_{ij} \quad (4)$$

где x_i и x_j - количества последовательностей i и j , а π_{ij} - мутационное расстояние между этими последовательностями, оцененное в соответствии с принятой моделью молекулярной эволюции

Гаплотипическое разнообразие S

Число различных гаплотипов в популяции (независимо от степени их различия). Синоним – число сегрегирующих аллелей (гаплотипов)

Тесты

$$\theta = 4N_e\mu \quad (5)$$

Величины N_e и μ определить по отдельности очень сложно, но популяционно-генетический параметр θ можно оценить, зная либо нуклеотидную изменчивость, либо гаплотипическое разнообразие. Для того, чтобы результаты вычислений совпали (то есть $\theta(\Pi) = \theta(S)$), должны соблюдаться следующие условия:

- Все признаки нейтральны
- Бесконечно большое число сайтов
- Панмиксия
- Отсутствие рекомбинации
- Равновесие
- Постоянный размер популяции

D критерий Таджимы

Критерий Таджимы:

$$D = \frac{\theta(\Pi) - \theta(S)}{\sqrt{var(\theta(\Pi) - \theta(S))}} \quad (6)$$

- В случае соблюдения всех условий $-2.5 < D < 2.5$ (приблизительно)
- Достоверные отличия D от 0 можно объяснить многими способами, поэтому требуются дополнительные исследования. Особенно это относится к постоянству численности
- $D \gg 0$ означает дефицит редких аллелей
- $D \ll 0$ указывают на недостаток гетерозигот, возможно в результате бутылочного горлышка или селективного обеднения (selective sweep)

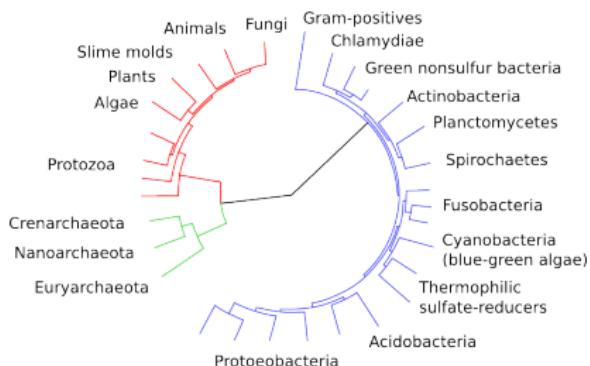
Некодирующая ДНК и стабилизирующий отбор

- Результаты оценки доли находящихся под действием отбора участков генома с помощью полногеномных данных:

У мышей 1.5% экзонов, но 5% всего генома находятся в зоне стабилизирующего отбора

- Примерно такое же соотношение CNS и нейтрально эволюционирующих последовательностей обнаружено у человека. .
- В более компактных геномах (*Drosophila melanogaster*, *Caenorhabditis elegans*) - CNS (консервативные некодирующие последовательности) могут составлять до половины генома.

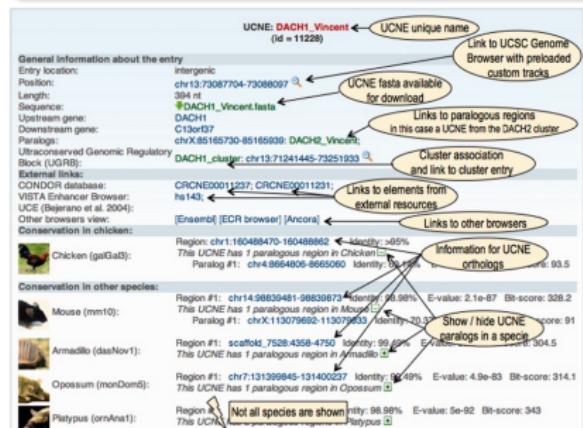
Самые консервативные из нкДНК



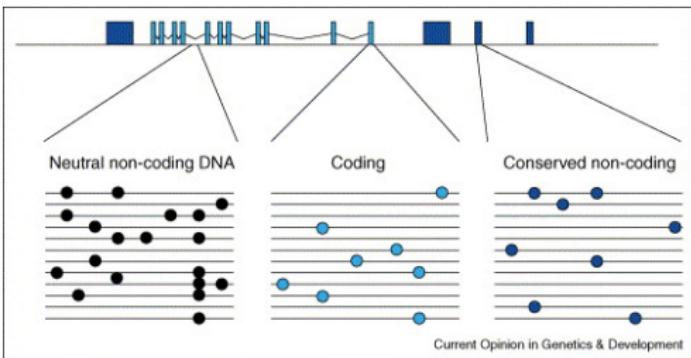
Некоторые нкДНК длиной 100-200 пн настолько консервативны, что их рассматривают как свидетельство в пользу гипотезы о монофилической функции этих участков ничего пока узнать не удалось

источники информации об ультраконсервативных нкДНК

База UCNEbase (<http://ccg.vital-it.ch/UCNEbase>) содержит информацию о ≈ 20000 обнаруженных к настоящему времени ультра-консервативных нкДНК.



из Dimitrieva S, Bucher P, **UCNEbase—database of ultraconserved non-coding elements and genomic regulatory blocks**. Nucleic Acids Res. 2013 Jan;41:D101-9.

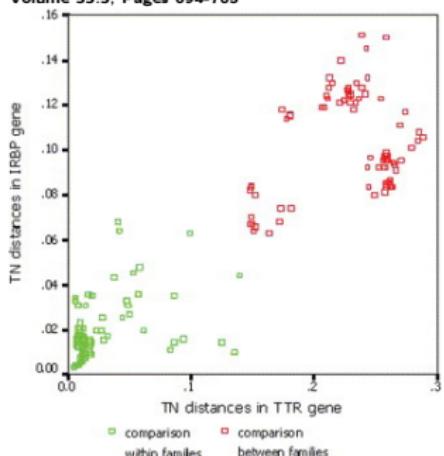


CP Bird, BE Stranger, ET Dermitzakis, [Functional variation and evolution of non-coding DNA](#), Current Opinion in Genetics & Development,

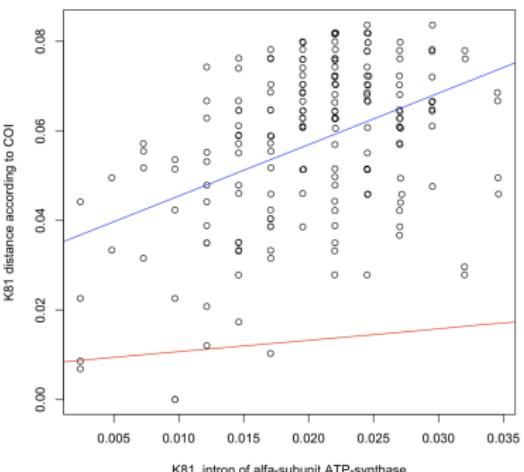
Volume 16, Issue 6, December 2006, Pages 559-564

Сравнение кодирующей и некДНК

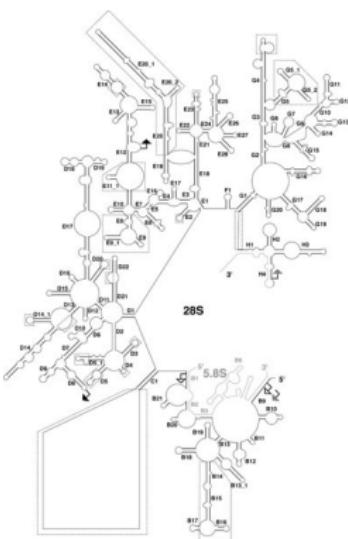
引文: Li Yu, Qing-wei Li, O.A. Ryder, Ya-ping Zhang, 2003 Phylogenetic relationships within mammalian order Carnivora indicated by sequences of two nuclear DNA genes, Molecular Phylogenetics and Evolution, Volume 33-3, Pages 694-705



JRPB - экзоп. ТТР-интрапн

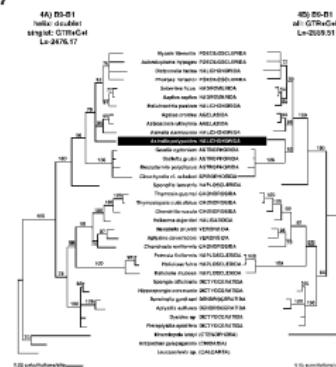


Учет структурных ограничений эволюции нкДНК



Структура большой субъединицы рРНК (Schnare MN, Damberger SH, Gray MW, Gutell RR (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA. J Mol Biol 256:701-719). Сохранение этой структуры в эволюции необходимо для работы пептидилтрансферазы

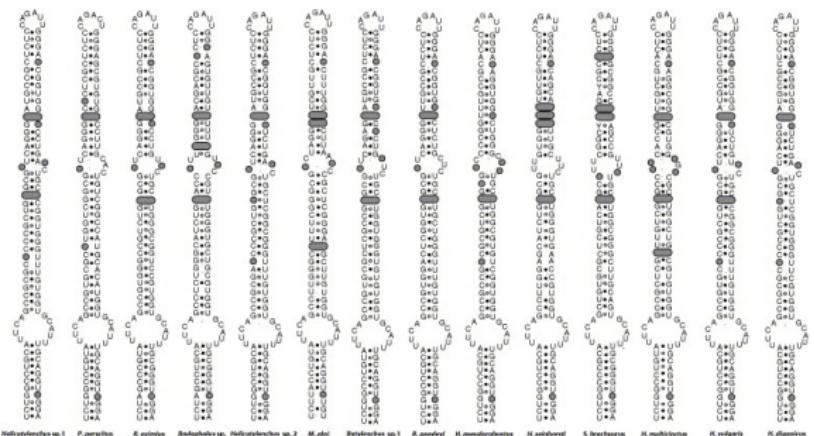
Erpenbeck D, Nichols SA, Voigt O, Dohrmann M, Degnan BM, Hooper JN, Wörheide G. Phylogenetic analyses under secondary structure-specific substitution models outperform traditional approaches: case studies with diploblast LSU. 2007 J Mol Evol. :64(5):543-57



На нескольких примерах показали, что учет шпилечной структуры влияет на топологию дерева и улучшает статистическую поддержку

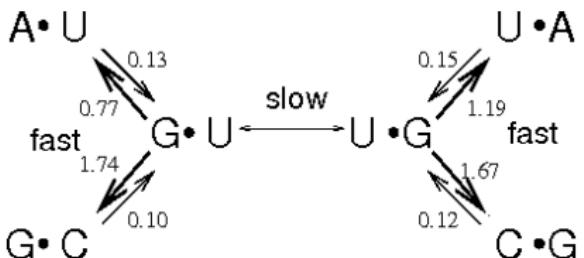


Сохранение вторичной структуры в эволюции

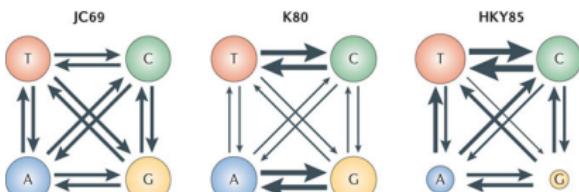


B Lamontagne, G Ghazal, I Lebars, S Yoshizawa, D Fourmy, SA Elela. JMB:327(5), 2003, 985–1000

Дублетная модель эволюции НК



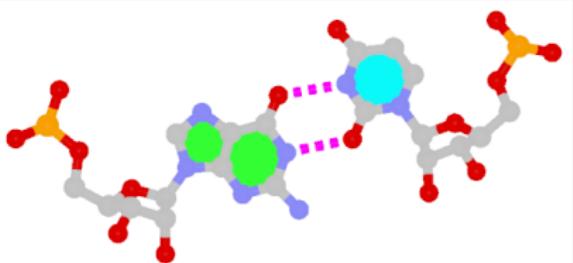
Higgs PG 2000, [RNA secondary structure: physical and computational aspects](#). Q Rev Biophys. 33(3): 199-253



Nature Reviews | Genetics

Ziheng Yang & Bruce Rannala 2012. [Molecular phylogenetics: principles and practice](#) Nature Reviews Genetics 13, 303-314

Посредник в дублетной модели



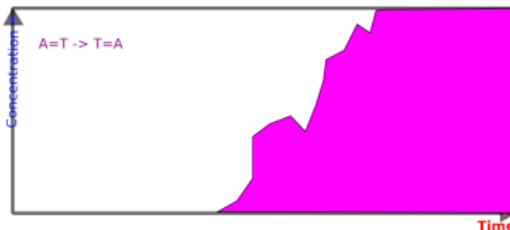
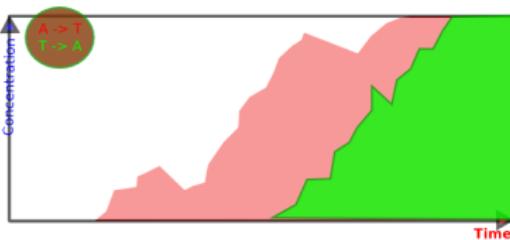
Замена одного дублета на другой без нарушения структуры
шпилки возможна через неканоническую пару **G•U**

Общее свойство однонуклеотидных моделей

Все замены происходят совершенно независимо друг от друга.



Популяции и дублетная эволюция



основное отличие

Дублет может фиксироваться либо в результате дрейфа, либо в результате отбора, но – как единое целое. Мононуклеотидный механизм предполагает:

- отбор против первой мутации, нарушающей структуру (в результате её фиксация возможна только в условиях бутылочного горлышка)
- отбор в пользу второй мутации, которая восстанавливает равновесие

Подходит ли дублетная модель?

Проблема

Требуется сравнить лучшую из однонуклеотидных моделей молекулярной эволюции с дублетной и оценить количественно преимущество предпочтительной модели.

Трудности

- Различное число параметров и принципиальные различия между моделями.
- Неоднозначность моделей укладки РНК.
- Эволюция может протекать в соответствии с промежуточной моделью
- Для достаточно длинных фрагментов требуется много раз менять рамку, но даже для одного набора данных приходится проводить очень объёмные вычисления

Подходит ли дублетная модель? (продолжение)

байесовское отношение шансов и ABC

$$K = \frac{\text{prob}(D|M_1)}{\text{prob}(D|M_2)} \quad (7)$$

числитель и знаменатель, соответственно, - вероятность наблюдаемых данных при условии модели №1 и при условии модели №2

Поскольку аналитически эти величины вычислить невозможно, используются их приближенные значения, полученные, например, с помощью программы MrBayes или соответствующих библиотек функций на языках R, Python и т.п.



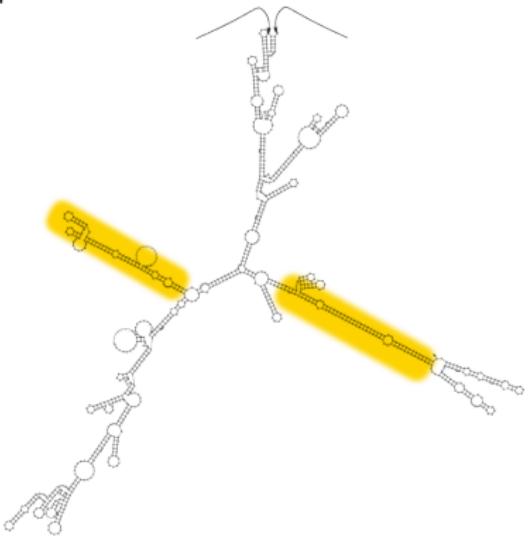
2-й инtron рибосомного белка S7 цихлид



Учет вторичной структуры влияет на длины ветвей, но не на топологию деревьев



Вторичная структура РНК интрана, желтым выделены участки, комплементарные фрагментам SSU РНК



Байесовское отношение шансов > 6 в пользу дублетной модели

Модели эволюции

зависимые мутации

- дублетная модель – частный случай контекстно-зависимых моделей
- наиболее распространенные контекстно-зависимые модели - кодонные
- появляются сосед-зависимые модели

Проблемы сложных моделей

- Рост числа параметров затрудняет поиск наборов их оптимальных значений и приводит к росту объёмов экспериментальных данных, которые для этого нужны
- Затруднено сравнение принципиально разных моделей с разным числом степеней свободы

Для того, чтобы оценить достоверность различий между последовательностями, требуется:

- Насколько вероятно наблюдаемое сходство для случайно взятых последовательностей?
- Насколько вероятно наблюдаемое сходство для родственных последовательностей, которые случайно перемешали, сохранив таким образом неравновесные частоты нуклеотидов (аминокислот)
- Насколько вероятно наблюдаемое сходство для случайных последовательностей, генерированных с использованием реальной модели эволюции?

Новые возможности

- Исследование особенностей эволюции нкДНК позволит найти функционально важные участки генома, о функции которых мы не имеем понятия;
- Уже первые результаты принесли парадоксы, нуждающиеся в объяснении, что будет стимулом развития биологии
- Новый уровень понимания механизмов функционирования наследственного аппарата клетки не может не принести практической пользы
- Не исключено, что понимание функций и механизмов, связанных с функционированием нкДНК потребует переосмыслиния многих кажущихся сегодня незыблемых принципов биологии в целом

Условная вероятность

Пример

Игральная кость подбрасывается один раз. Известно, что выпало более трёх очков. Какова вероятность того, что выпало чётное число очков?

Зная, что выпало более трёх очков, мы можем сузить множество всех возможных элементарных исходов до трёх одинаково вероятных исходов: $\Omega = \{4, 5, 6\}$, из которых событию $A = \{\text{выпало чётное число очков}\}$ благоприятствуют ровно два: $A = \{4, 6\}$. Поэтому $P(A) = 2/3$.

Посмотрим на вопрос с точки зрения первоначального эксперимента. Пространство элементарных исходов при одном подбрасывании кубика состоит из шести точек: $\Omega = \{1, \dots, 6\}$. Слова «известно, что выпало более трёх очков» означают, что в эксперименте произошло событие $B = \{4, 5, 6\}$. Слова «какова при этом вероятность того, что выпало чётное число очков?» означают, что нас интересует, в какой доле случаев при осуществлении B происходит и A . Вероятность события A , вычисленную в предположении, что о результате эксперимента уже что-то известно (событие B произошло), мы будем обозначать через $P(A|B)$.

Условная вероятность

Пример

Игральная кость подбрасывается один раз. Известно, что выпало более трёх очков. Какова вероятность того, что выпало чётное число очков?

Зная, что выпало более трёх очков, мы можем сузить множество всех возможных элементарных исходов до трёх одинаково вероятных исходов:

$\Omega = \{4, 5, 6\}$, из которых событию $A = \{\text{выпало чётное число очков}\}$ благоприятствуют ровно два: $A = \{4, 6\}$. Поэтому $P(A) = 2/3$.

Посмотрим на вопрос с точки зрения первоначального эксперимента. Пространство элементарных исходов при одном подбрасывании кубика состоит из шести точек: $\Omega = \{1, \dots, 6\}$. Слова «известно, что выпало более трёх очков» означают, что в эксперименте произошло событие $B = \{4, 5, 6\}$. Слова «какова при этом вероятность того, что выпало чётное число очков?» означают, что нас интересует, в какой доле случаев при осуществлении B происходит и A . Вероятность события A , вычисленную в предположении, что о результате эксперимента уже что-то известно (событие B произошло), мы будем обозначать через $P(A|B)$.

Условная вероятность

Пример

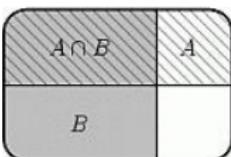
Игральная кость подбрасывается один раз. Известно, что выпало более трёх очков. Какова вероятность того, что выпало чётное число очков?

Зная, что выпало более трёх очков, мы можем сузить множество всех возможных элементарных исходов до трёх одинаково вероятных исходов:
 $\Omega = \{4, 5, 6\}$, из которых событию $A = \{\text{выпало чётное число очков}\}$ благоприятствуют ровно два: $A = \{4, 6\}$. Поэтому $P(A) = 2/3$.

Посмотрим на вопрос с точки зрения первоначального эксперимента. Пространство элементарных исходов при одном подбрасывании кубика состоит из шести точек:
 $\Omega = \{1, \dots, 6\}$. Слова «известно, что выпало более трёх очков» означают, что в эксперименте произошло событие $B = \{4, 5, 6\}$. Слова «какова при этом вероятность того, что выпало чётное число очков?» означают, что нас интересует, в какой доле случаев при осуществлении B происходит и A . Вероятность события A , вычисленную в предположении, что о результате эксперимента уже что-то известно (событие B произошло), мы будем обозначать через $P(A|B)$.

Условная вероятность - продолжение

Мы хотим найти, какую часть составляют исходы, благоприятствующие A внутри B (т.е. одновременно A и B), среди исходов, благоприятствующих B .



$$P(A|B) = \frac{2}{3} = \frac{2/6}{3/6} = \frac{P(A \cap B)}{P(B)}$$

определение

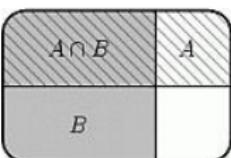
Условной вероятностью события A при условии, что произошло событие B , называется число

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Условная вероятность определена только в случае, когда $P(B) > 0$

Условная вероятность - продолжение

Мы хотим найти, какую часть составляют исходы, благоприятствующие A внутри B (т.е. одновременно A и B), среди исходов, благоприятствующих B .



$$P(A|B) = \frac{2}{3} = \frac{2/6}{3/6} = \frac{P(A \cap B)}{P(B)}$$

определение

Условной вероятностью события A при условии, что произошло событие B , называется число

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Условная вероятность определена только в случае, когда $P(B) > 0$

Теорема Байеса

В простейшей форме теорема Байеса выглядит так:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (8)$$

где

$P(B|A)$ - условная вероятность B , или вероятность B при условии A .
Иначе этот член называется **правдоподобием** по-английски – **likelihood**

$P(A)$ -априорная вероятность A . Другие названия - маргинальная или безусловная вероятность B . Эта вероятность называется “априорной” или “предварительной” потому, что она никак не зависит от события B - оно может вообще не наступить.

$P(A|B)$ - условная вероятность B при условии A . Иначе называется **постериорной** вероятностью. “Постериорной” эта величина называется потому, что прямо зависит от B .

Данные и гипотезы

Данные

Предположим, что в нашем распоряжении есть некоторый набор данных **D**. В случае молекулярной филогенетики это, как правило, набор нуклеотидных последовательностей:

| | |
|---------|---|
| 8648084 | CACCTCTACAATGGATGCCGACAGGATTGTATTCAAGAGCTAATAATCAGGTGGTCTCTTT |
| 9627197 | CACCTCTACAATGGATGCCGACAAGATTGTATTCAAAGTCATAATCAGGTGGTCTCTTT |
| BR-Pfx1 | CACCCCTACAATGGATGCCGACAAGATTGTGTTCAAAGTCATAATCAGGTGGTCTCTTT |
| CQ92 | CACCCCTACAATGGATGCCGACAAGATTGTGTTCAAAGTCATAATCAGGTGGTCTCTTT |
| SH06 | CACCCCTACAATGGATGCCGACAAGATTGTGTTCAAAGTCGACAATCAGGTGGTCTCTTT |
| CVS-11 | CACCCCTACAATGGATGCCGACAAGATTGTGTTCAAAGTCATAATCAGGTGGTCTCTTT |
| 1DRV | CACCCGTACAATGGATGCCGACAAGATTGTATTCAAAGTCATAATCAGGTGGTCTCTTT |
| 1ERA | CACCCCTACAATGGATGCCGACAAGATTGTATTCAAAGTCATAATCAGGTGGTCTCTTT |

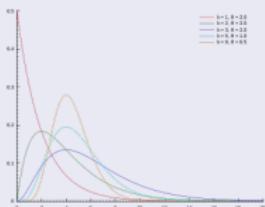
Задача

Как правило, можно предложить несколько гипотез, предлагающих сценарии появления этих данных: H_1, H_2, \dots, H_n .

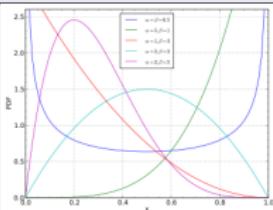
Цель применения байесовского фактора состоит в том, чтобы выбрать одну из этих гипотез

priors 1

две гипотезы



предположим, параметры модели распределены в соответствии с Гамма-функцией



Другая гипотеза состоит в предположении, что мы имеем дело с бета-распределением

В данном примере рассматриваются всего две гипотезы - H_1 и H_2 , однако можно одновременно рассматривать большее их число. В этом случае будет производиться попарное сравнение

priors 2

Поскольку обе гипотезы - H_1 и H_2 - относятся к **одному и тому же** набору данных \mathbf{D} , то в первом случае предположение состоит в том, что наблюдаемые значения \mathbf{D} были получены в соответствии с гипотезой H_1 из соответствующего распределения $pr(\mathbf{D}|H_1)$, а во втором случае - из $pr(\mathbf{D}|H_2)$. Имея априорные (предварительные) вероятности $p(H_1)$ и $p(H_2) = 1 - p(H_1)$, анализ реальных данных приводит к постериорным вероятностям $p(H_1|\mathbf{D})$ и $p(H_2|\mathbf{D})$.

Роль экспериментальных данных

Роль экспериментальных данных, таким образом, состоит в том, что на их основе априорные вероятности превращаются в постериорные, и позволяют оценить "качество" сравниваемых гипотез.

Сравнивая постериорные вероятности, можно предпочесть одну из гипотез другой, а также оценить, насколько обосновано это предпочтение.

Для сравнения постериорных вероятностей принято использовать **отношение шансов, или odds ratio**

Отношение шансов

Отношение шансов

Отношением шансов называется отношение

$$o = \frac{p}{1-p}$$

где p - вероятность события, для которого рассчитывают эту величину

по теореме Байеса получаем:

$$pr(H_k|\mathbf{D}) = \frac{pr(\mathbf{D}|H_k)pr(H_k)}{pr(\mathbf{D}|H_1) + pr(\mathbf{D}|H_2)} \quad (9)$$

где $k = (1, 2)$

тогда

$$\frac{pr(H_1|\mathbf{D})}{pr(H_2|\mathbf{D})} = \frac{pr(\mathbf{D}|H_1)}{pr(\mathbf{D}|H_2)} \frac{pr(H_1)}{pr(H_2)} \quad (10)$$

Байесовским фактором называется отношение

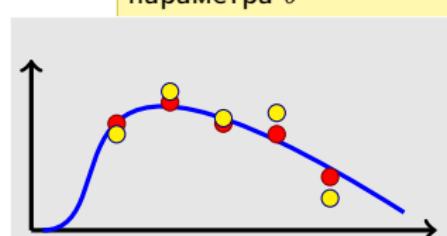
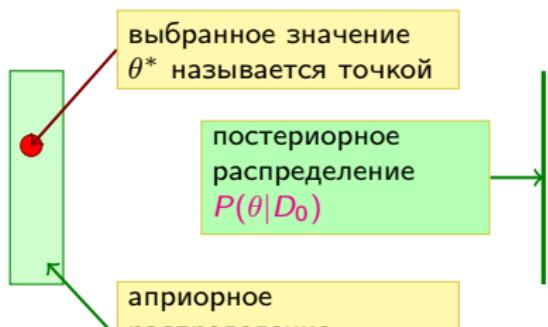
$$B_{12} = \frac{pr(\mathbf{D}|H_1)}{pr(\mathbf{D}|H_2)}$$

Часто невозможно или слишком сложно вычислить функцию правдоподобия (likelihood) $pr(B|A)$. Тогда вычисление правдоподобия заменяют сравнением между наблюдаемыми и симулированными данными.

ABC: оценка параметра. Продолжение

Задача

На основании априорного распределения параметра $P(\theta)$ и данных D_0 требуется определить постериорное распределение $P(\theta|D_0)$



На первом этапе случайным образом из априорного распределения $P(\theta)$ выбираем случайным образом некое значение θ^* .

Затем с помощью моделирующей процедуры $f(D|\theta^*)$ несколько раз используем данные D и параметр θ^* для того, чтобы получить результат моделирования (симуляции) D^* .

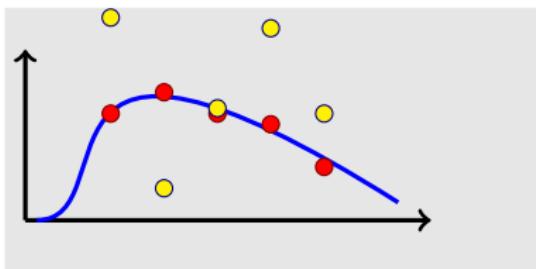
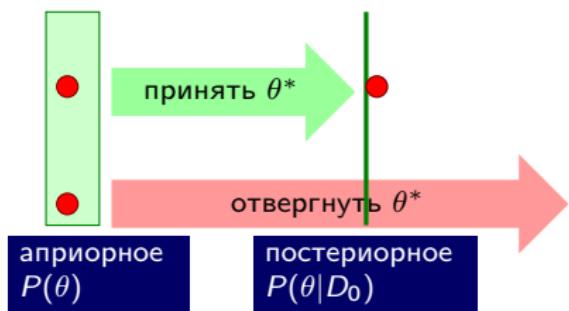
Процедура симуляции может быть любой - от уравнения до алгоритма вычисления топологии эволюционного дерева.

Симулированные данные D^* сравнивают с исходными D_0 с вычисляя дистанцию между ними с помощью функции d и "уровня толерантности" ϵ . Значение θ^* принимается в том случае, если $d(D_0, D^*) \leq \epsilon$, то это значение принимается.

ABC: оценка параметра

Задача

На основании априорного распределения параметра $P(\theta)$ и данных D_0 требуется определить постериорное распределение $P(\theta|D_0)$

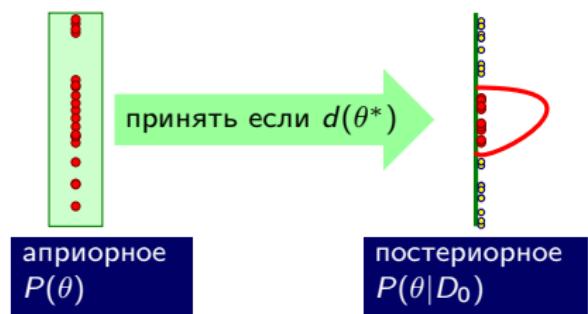


В первом случае (верхняя точка) θ^* принимается поскольку соблюдается $d \leq \epsilon$, то есть D_0 и D^* достаточно близки. Во втором случае это оказывается не так. D_0 и D^* показаны на графике. $d > \epsilon$, соответствующее значение θ^* отвергается.

ABC: оценка параметра. Окончание

Задача

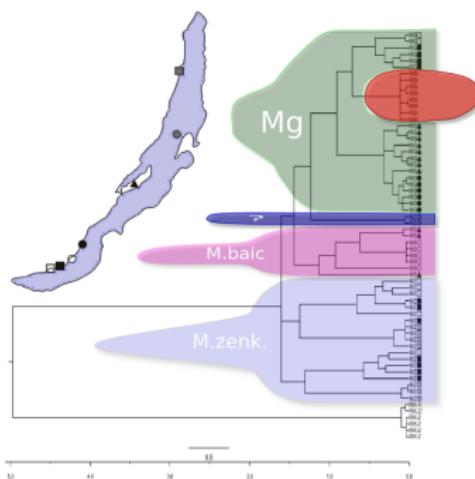
На основании априорного распределения параметра $P(\theta)$ и данных D_0 требуется определить постериорное распределение $P(\theta|D_0)$



Пресноводные полихеты рода *Manayunkia*



Manayunkia baicalensis



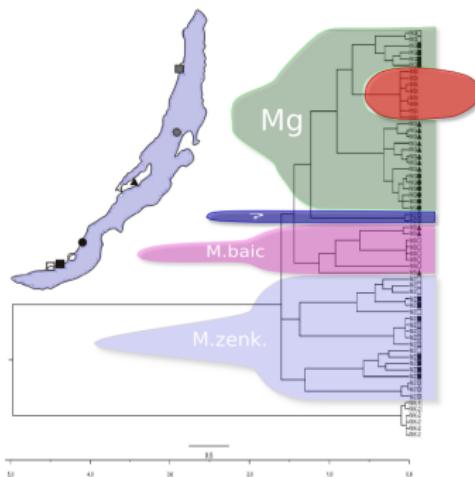
Филогения на основании мит. ДНК:

Байкальские полихеты распадаются как минимум на 3 (может быть – четыре) монофилетические клады, соответствующие трем морфологическим видам *Manayunkia*, описанным ранее.

Пресноводные полихеты рода *Manayunkia*



Manayunkia baicalensis



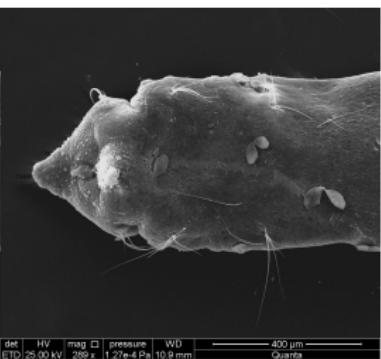
Филогения на основании мит. ДНК:

Байкальские полихеты распадаются как минимум на 3 (может быть - четыре) монофилетические клады, соответствующие трем морфологическим видам *Manayunkia*, описанным ранее.

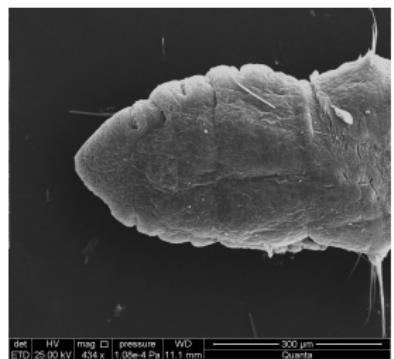
Морфологические различия между байкальскими полихетами



M.baicalensis



M.baicalensis



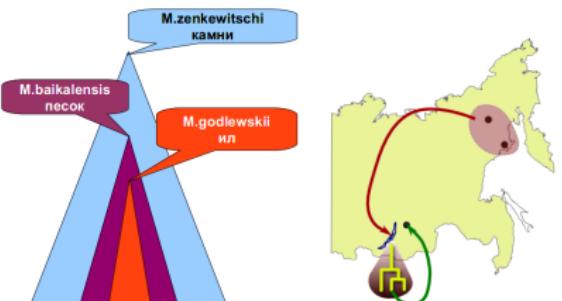
M.zenkewitschii

Фото Т.Я. Ситникова, Т.А.Пудовкина

Три вида или три экологические формы?

Различия между байкальскими манаюнками

Основные различия между байкальскими видами манаюнкий состоят в субстратных предпочтениях. Поэтому можно предположить: **вилообразование у полихет произошло путем адаптации к разным субстратам**



Возможные механизмы видообразования *Manayunkia*

- **H1**Переключение субстратов происходило единожды
- **H2**Преключение носило множественный характер и параллельно происходило в разных частях Байкала

Результаты байесовского сравнения гипотез с использованием митохондриальных последовательностей

H1 Митохондриальная филогенетика содержит сильную поддержку гипотезы о монофилии - т.е. о единственном переключении.

Три вида или три экологические формы?

Различия между байкальскими манаюнками

Основные различия между байкальскими видами манаюнкий состоят в субстратных предпочтениях. Поэтому можно предположить: **вилообразование у полихет произошло путем адаптации к разным субстратам**



Возможные механизмы видообразования *Manayunkia*

- **H1**Переключение субстратов происходило единожды
- **H2**Преключение носило множественный характер и параллельно происходило в разных частях Байкала

Результаты байесовского сравнения гипотез с использованием митохондриальных последовательностей

H1 Митохондриальная филогения содержит сильную поддержку гипотезы о монофилии - т.е. о единственном переключении.

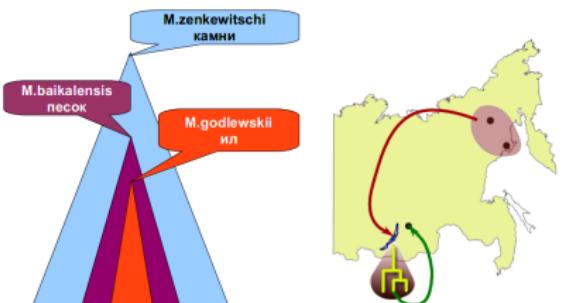
Три вида или три экологические формы?

Различия между байкальскими манаюнками

Основные различия между байкальскими видами манаюнки состоят в субстратных предпочтениях. Поэтому можно предположить: **вилообразование у полихет произошло путем адаптации к разным субстратам**

Возможные механизмы видообразования *Manayunkia*

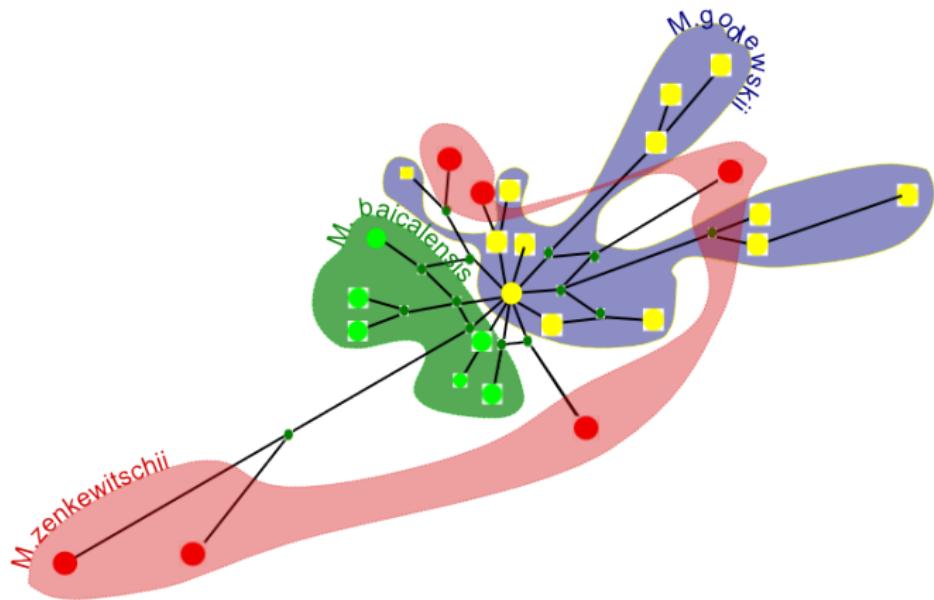
- **H1**Переключение субстратов происходило единожды
- **H2**Преключение носило множественный характер и параллельно происходило в разных частях Байкала



Результаты байесовского сравнения гипотез с использованием митохондриальных последовательностей

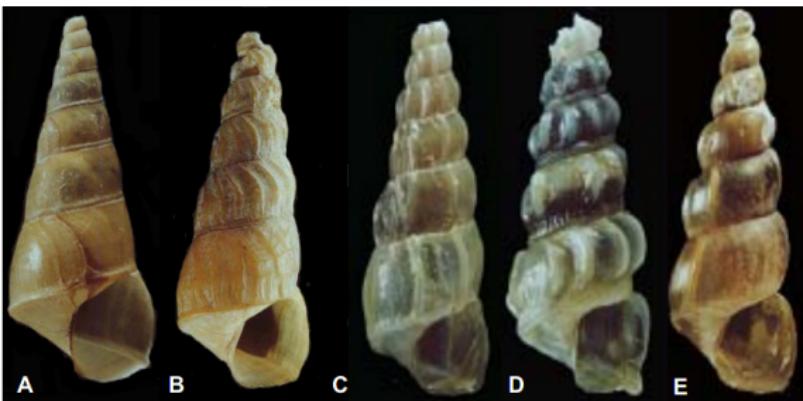
H1 Митохондриальная филогения содержит сильную поддержку гипотезы о монофилии - т.е. о единственном переключении.

последовательности интрана α -субъединицы АТФазы и "кластеризация" полихет



Простирающееся дерево последовательностей интрана α -субъединицы АТФазы и ITS (не показано) согласуется с гипотезой о множественном происхождении байкальских видов полихет

Baicaliidae

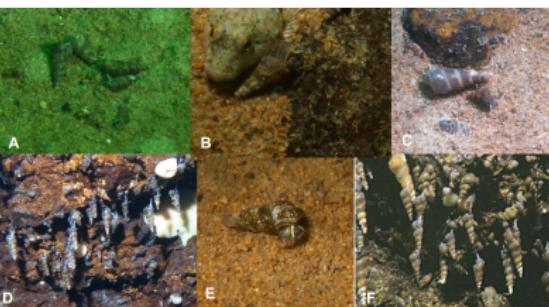


A – *B. carinata*, B – *B. carinata rugosa*, C – *B. carinatocostata*, D – *B. dybowskiana*, E – *B. turriformis*

Фото: Т. Sitnikova

Ниши видов рода *Baicalia*

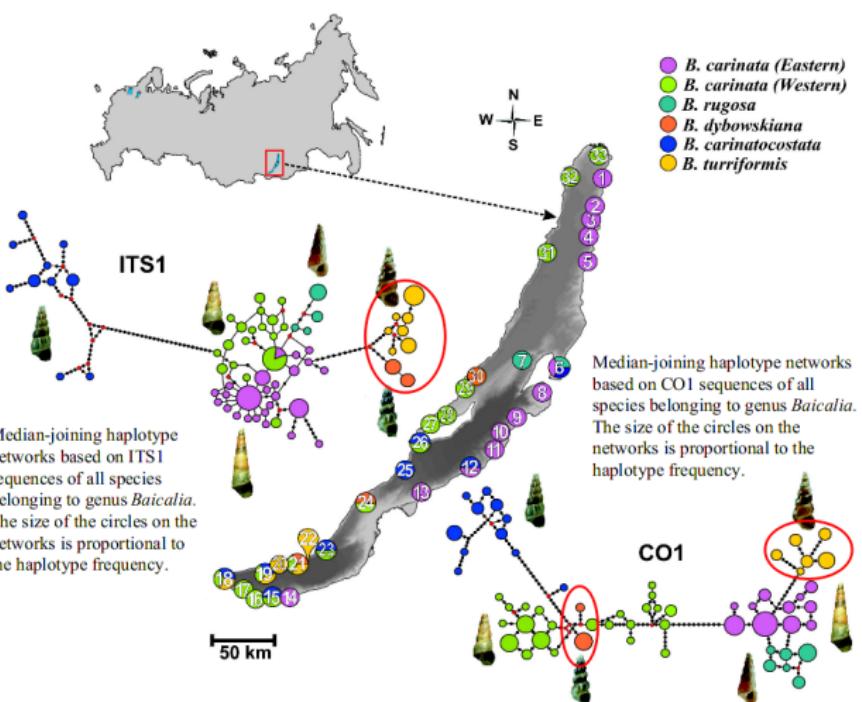
| Species | Substrate | Breeding | West coast | | East coast |
|---------------------------|-----------|--------------------|------------|---|------------|
| | | | | | |
| <i>B. carinata</i> | Sand | Conspecific shells | + | + | |
| <i>B. carinatocostata</i> | Sand | Sand | + | + | |
| <i>B. dybowskiana</i> | Sand | Stones | + | - | |
| <i>B. rugosa</i> | Sand | Stones | - | + | |
| <i>B. turiformis</i> | Stone | Stone | + | - | |



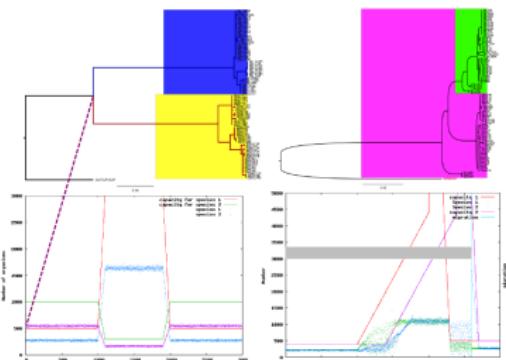
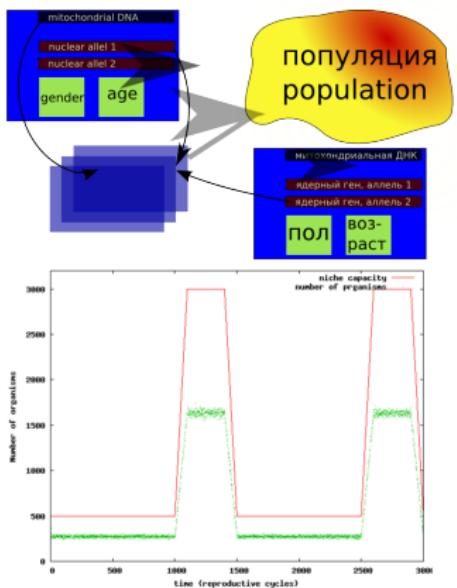
• A, C – *B. carinata*, B – *B. rugosa*, D, F – *B. turiformis*

Фото: И.Ханаев, К.Иванов

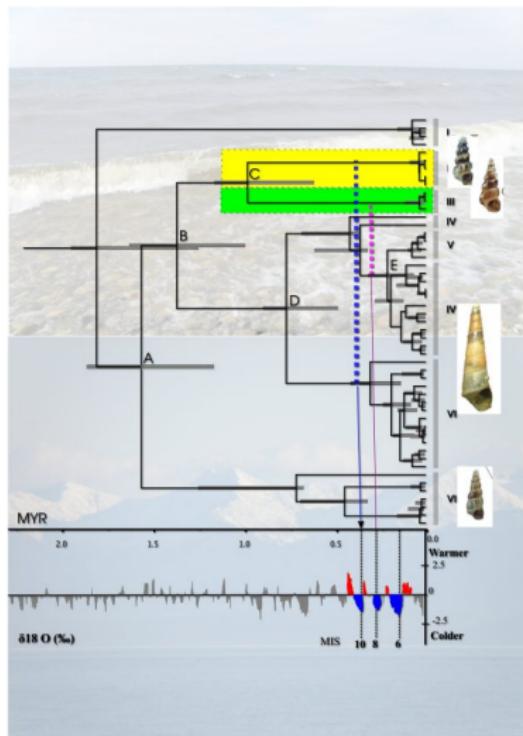
ЭВОЛЮЦИЯ ЯДЕРНЫХ И МИТОХОНДРИАЛЬНЫХ ГЕНОВ у *Baicalia*



индивидуально-ориентированное моделирование



Baicalia резюме



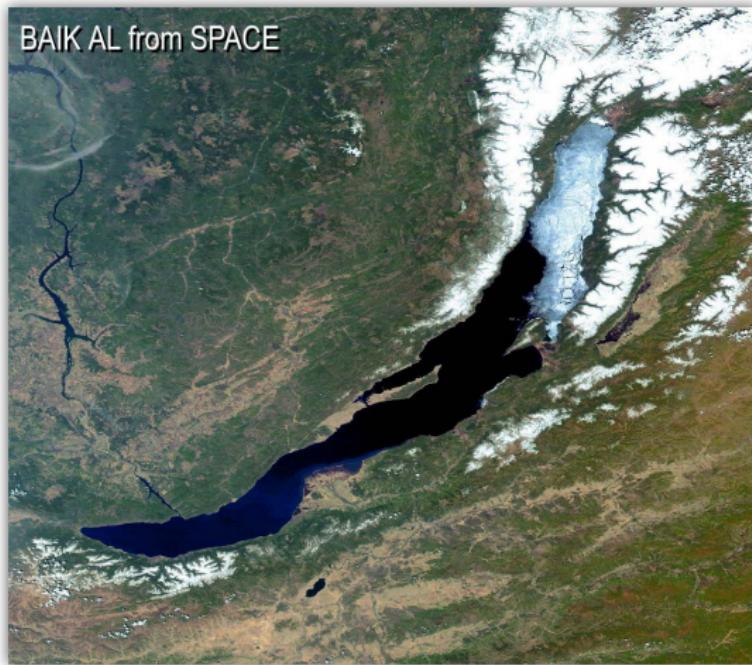
Филогения рода *Baicalia*

Использование гипотезы молекулярных часов показывает, что оба случая митохондриальной трансгрессии приходятся на холодные и сухие периоды истории байкальской экосистемы. В обоих случаях доступность песчаных биотопов должна была существенно снизиться.

Возможные причины эволюционного успеха рода *Baicalia*

Виды рода *Baicalia* в процессе эволюции приспособились к различным средовым нишам. Несогласованность между филогениями, построенными по ядерным и митохондриальным генам можно объяснить чередованиями благополучных периодов и периодов, когда для выживания одним видам *Baicalia* приходилось занимать средовые ниши других видов. Это могло сопровождаться межвидовой гибридизацией.

Lake Baikal



- The largest fresh water reservoir of the planet – 23000 km^3
- depth – 1637 m
- width - 30-70 km
- length – 620 km
- age of continuously existing large lake- 25 million years
- area 30500 km^2

Ancient lake full of young species

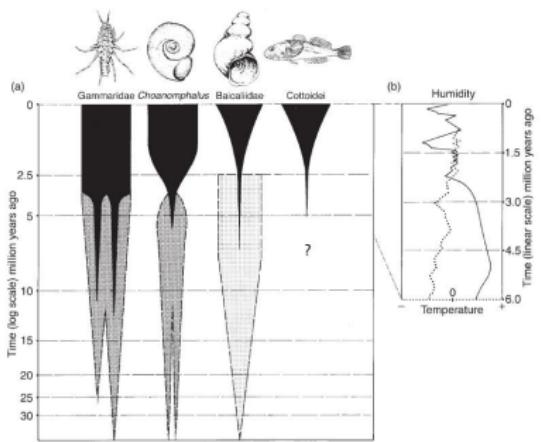


Fig. 1. (a) Summary of evolutionary histories of four Baikalian invertebrate species flocks: amphipods (Gammaridae, Crustacea), subendemic Baikalian genus *Choanophalus* (Pulmonata, Mollusca), endemic family Baicalidae (Prostomatoidea, Monotocardia), and sculpins (Scorpaeniformes). The inferred phylogenetic trees are based on molecular phylogenetic studies⁴¹. The width of the black areas correspond to the number of lineages in the species flocks at a particular point in time. The grey areas represent phylogenetic lineages that have no known descendants. In the case of Baicalidae, ancient fossils are known but their taxonomic relationship to contemporary species is doubtful⁴², thus the area is shaded a lighter grey. The question mark indicates that, owing to a lack of fossils, it is unknown whether sculpins existed in Lake Baikal before approximately 2.5 million years ago. (b) Trends in annual temperature (unbroken line) and humidity (broken line) for the past 5 million years⁴² with which to compare the critical periods in the development of the lake's biodiversity.

«Ancient» groups

Groups with MRCA existing 25+ MYA:

- Amphipods
- Chironomids (*Sergentia*)
- Lumbriculidae
- some groups of turbellarians

«Young» groups

Groups with MRCA existing less than 2.5 MYA

- Baicalidae
- *Choanophalus*
- few groups of turbellarians
- isopods (?)
- chironomids except *Sergentia*
- caddisflies
- sculpins
- polychaets *Manayunkia*

Why NGS?

Ancient flocks:

Problem: Number of characters

Already excessive while using usual "barcoding" markers.

types of variable characters

NGS data may be a source for microsatellite markers suitable for population-level studies

Expected:

- Increased number of variable sites will remove conflicting signals from different genes
- Comparison between evolutionary fates of different genes will shed new light on adaptation mechanisms

Young flocks:

Problem: Number of characters

Insufficient while using usual "barcoding" markers. NGS may bring new molecular traits required for phylogenetic inferences

types of variable characters

NGS data may be a good source for microsatellite markers suitable for population-level studies comparison between sister species

Expected:

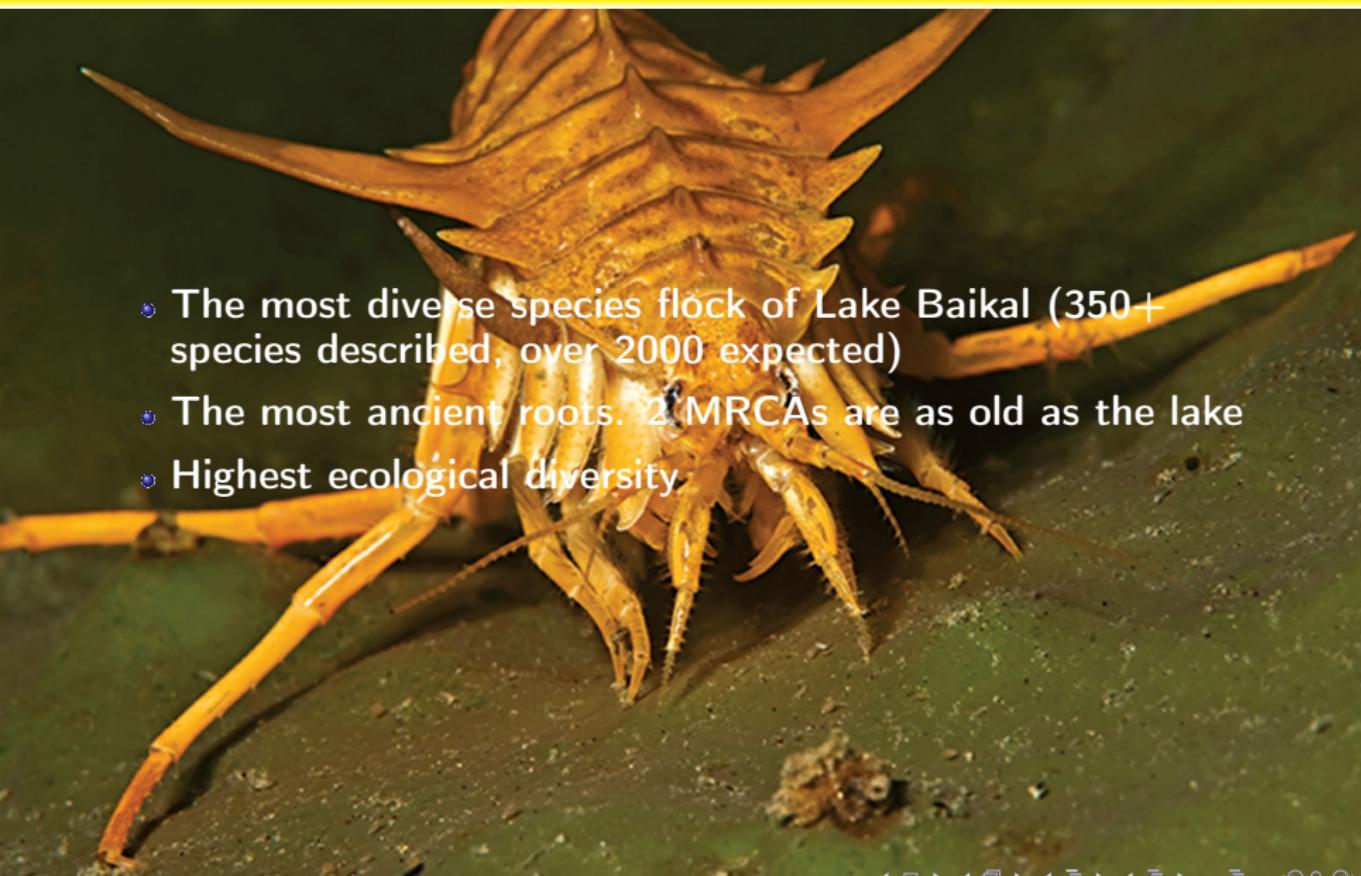
- Sufficient number of variation for well resolved phylogenies and better datings
- Comparison of fast nuclear markers to the mitochondrial ones will elucidate the discrepancies observed

Amphipods



Amphipods

- The most diverse species flock of Lake Baikal (350+ species described, over 2000 expected)
- The most ancient roots. 2 MRCAs are as old as the lake
- Highest ecological diversity



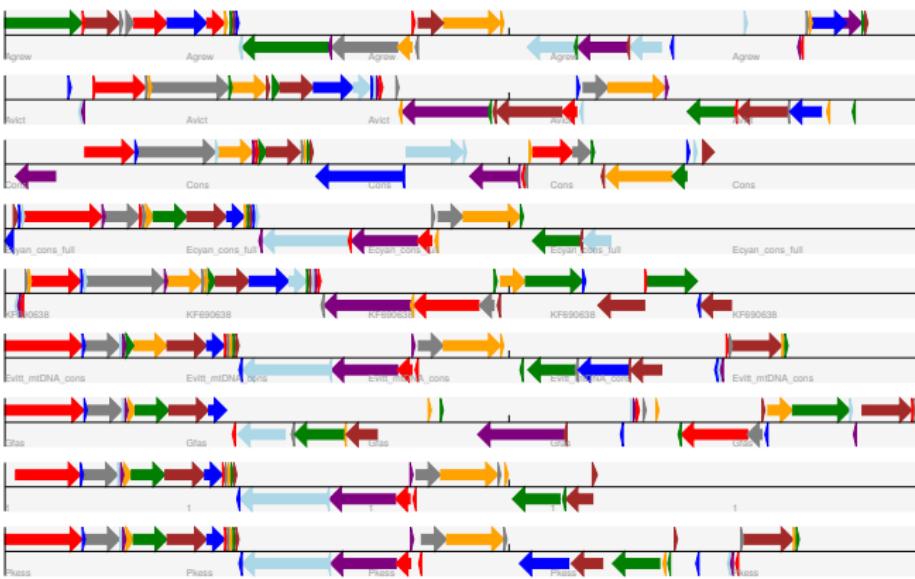
Amphipods

- The most diverse species flock of Lake Baikal (350+ species described, over 2000 expected)
- The most ancient roots. 2 MRCAs are as old as the lake
- Highest ecological diversity

25 years of molecular evolutionary studies did not result yet in a good phylogeny



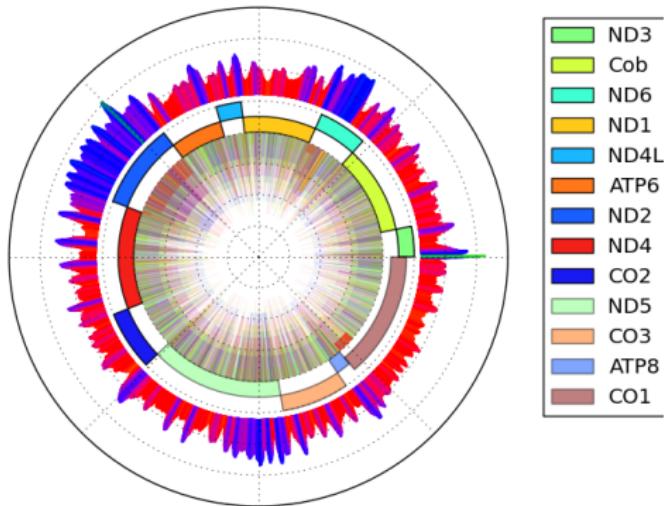
Amphipods: mitochondrial genomes



Gene rearrangements occur in different lineages independently!

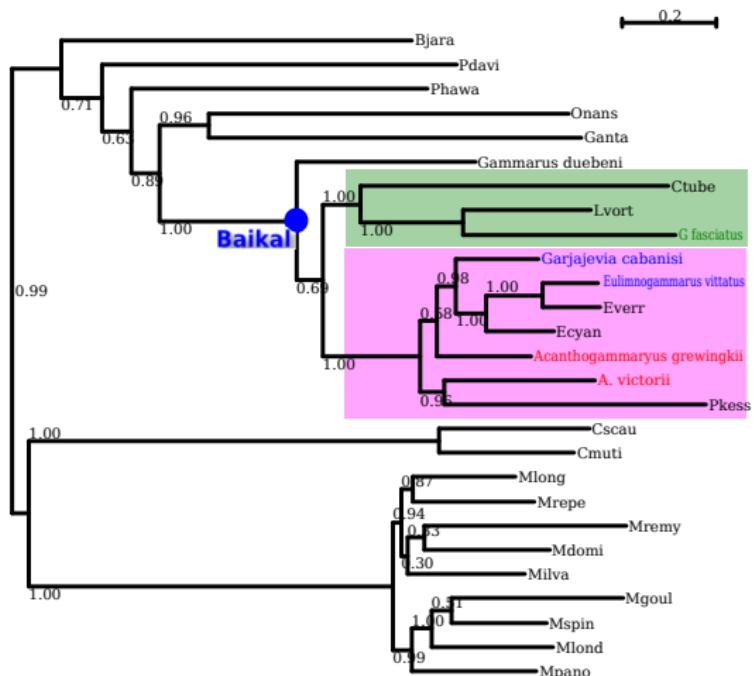
Relative variability of mitochondrial genes

Variability of the mitochondrial nucleotide sequences
of protein coding genes of 10 species of Baikalian endemic amphipodes



- Faster evolution: ND4, ND4L, ND6
- Slower evolution: COI
- baikalian pattern of diversity differs from the general one
- Evidences for positive selection concentrate in faster evolving parts of mitogenome

Coding mitochondrial genes-based phylogeny



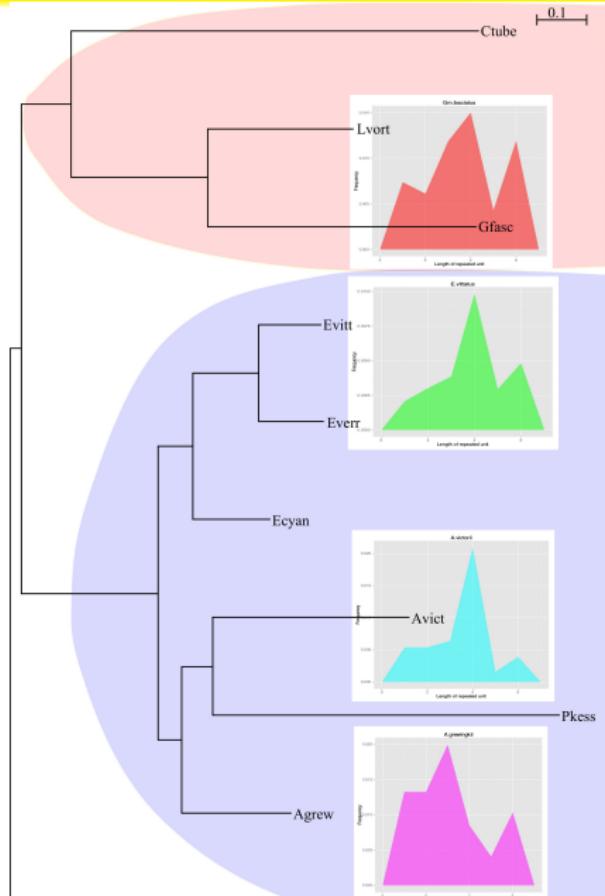
- There are 2 major lineages of amphipods in Baikal
- Some of taxonomic classification details remain unconfirmed
- gen. *Acanthogammarus* is polyphyletic, thus is not a genus
- gene order rearrangements are **not a good conservative diagnostic traits** as they were believed to be

Amphipods of Lake Baikal

Ratio of reads containing Simple Sequence Repeats from 1 to 6 bp long.

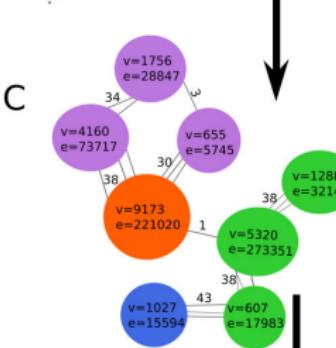
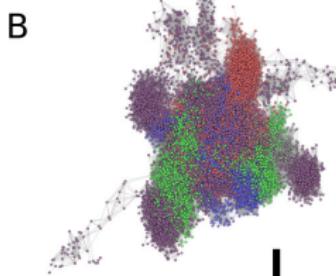
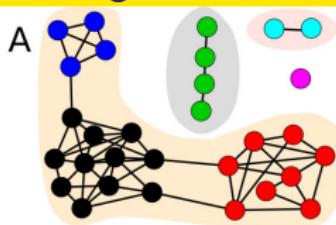
| Species | searched | reads containing microsats | percent |
|-----------------------------------|----------|----------------------------|---------|
| <i>Acanthogammarus grewingkii</i> | 1886693 | 79913 | 4.2% |
| <i>Acanthogammarus victorii</i> | 500407 | 19812 | 4% |
| <i>Eulimnogammarus vittatus</i> | 50666 | 1048 | 2% |
| <i>Garjajevia cabanisi</i> | 679544 | 39144 | 5.8% |
| <i>Gmelinoides fasciatus</i> | 61165 | 1343 | 2.2% |

Distribution of SSRs in amphipods



- there are dramatic differences in SSR content in the genomes studied;
- The differences may include total absence of certain SSR in one species and high abundance in the other
- so far no SSR shared between at least two species of amphipods have been found
- NGS fishing for SSR is still better than the traditional experimental approach and helps the studies of intra-specific variation
- One must beware the retroelements!**

Clustering method of searching repeats



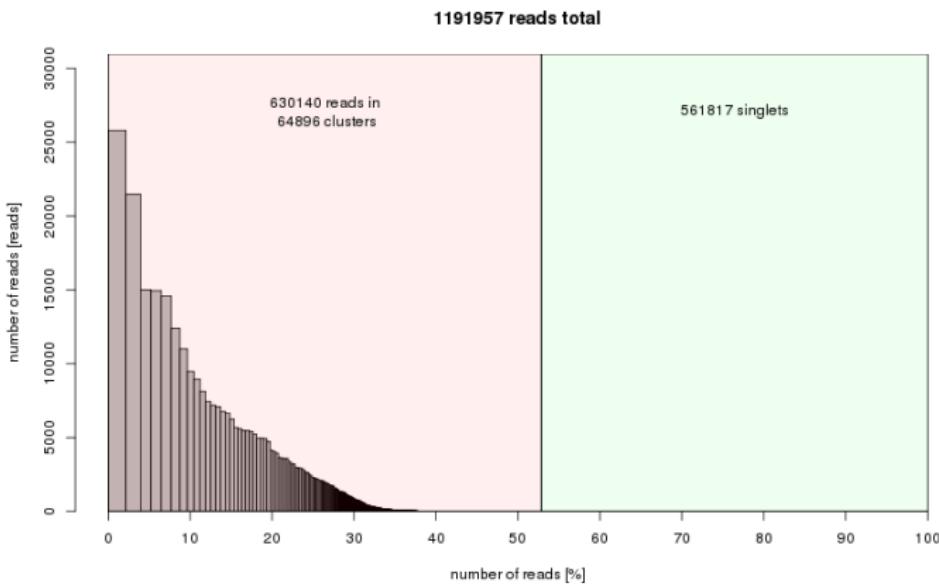
Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data
 P. Novák, P. Neumann and J. Macas* BMC Bioinformatics 2010, 11:378

General data on *E.vittatus*

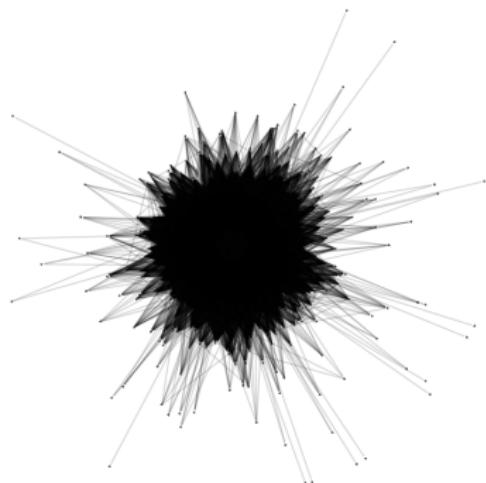
Novak, P., Neumann, P., Pech, J., Steinhäsl, J., Macas, J. (2013) - RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics 29:792-793.

or

Novak, P., Neumann, P., Macas, J. (2010) - Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378.



Most of the repeats belong to "simple repeats"



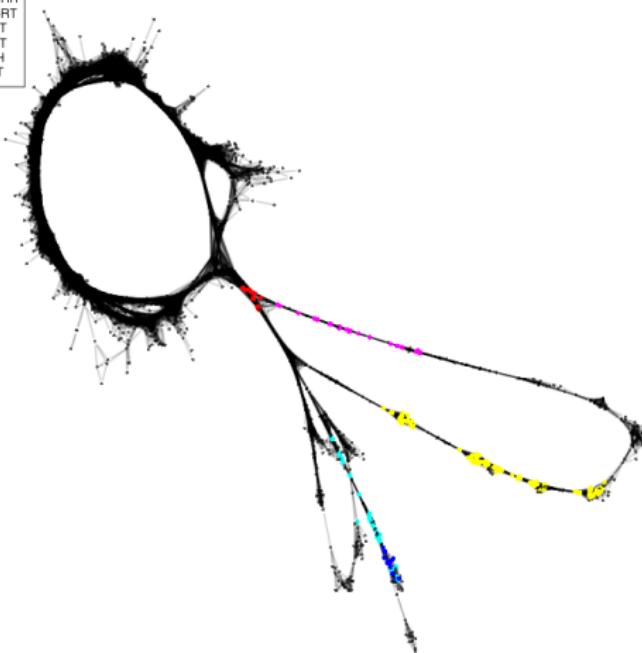
CL1 occupies 2.15% of the genome, no domain hits

Several interconnected clusters of simple repeats

| cluster | CL2 | CL3 | CL4 | CL5 |
|-----------------------------|-----------------|--------------------|-------------|-------------|
| % of genome main paralog | 1.8% LINE-RT | 1.25% LTR-gypsy | 1.25% No | 1.22% No |

More complex repeat

- PARA-RH
- PARA-RT
- Ty1-INT
- Ty3-INT
- Ty3-RH
- Ty3-RT



oo
Centro de cultura

LTR retro-element example

