# Bioinformatics and Evolutionary Genomics

## Evolution of Genomes, Proteomes, Networks and Complexes

Berend Snel
Theoretical Biology and Bioinformatics
Department of Biology
Science Faculty
Utrecht University

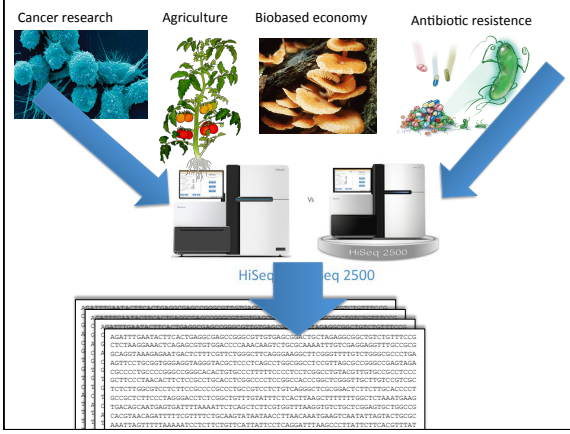


## Today

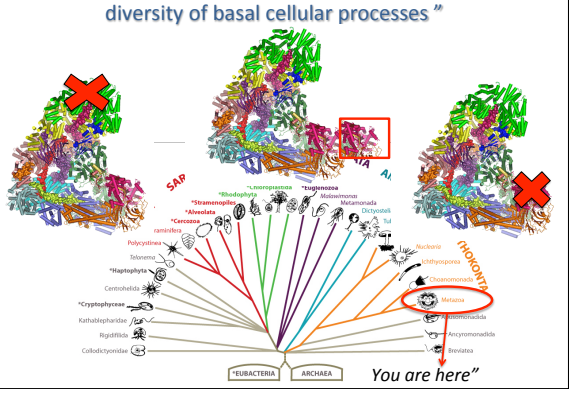
- Introduction on general aims of the course and on procedural stuff
- Lecture on homology and domains (see how far we get …)

## Requests

- **very** heterogeneous with respect to previous knowledge (IBMB, GB, research projects, PhD students)
- PLEASE: interrupt / ask questions when I am going to fast, when I use jargon, when I make jumps/conclusions that to me seem obvious 100% logical, but to your are erratic; please point out my implicit assumptions regarding what everybody knows
- -> Master course …
- Computer exercises: more experienced people help




"Mapping and understanding fate of ancestral eukaryotic diversity of basal cellular processes"

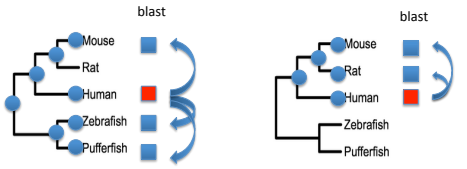



*You are here"*

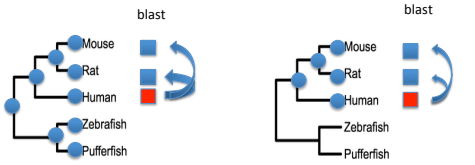
## What kind of questions?

## What kind of questions?

- What is the history of my gene? (what happened)
  - When was my gene invented? (what do you mean your gene? Duplication vs *de novo* gene invention)
  - How conserved is my gene? In the meaning of: in which distant species does it also occur?
  - When did this motif arise?
- "± which other genomes contain my gene"
- "What is the same thing in different organisms and how did it evolve … "



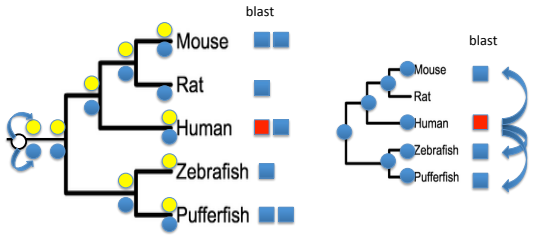

Gene originates in common ancestor (or later) … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.




Gene originates in common ancestor… but evolves rapidly (coiled coil, disordered, very short globular domain)

Gene originates later… evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

Problem: divergent homology = orthology

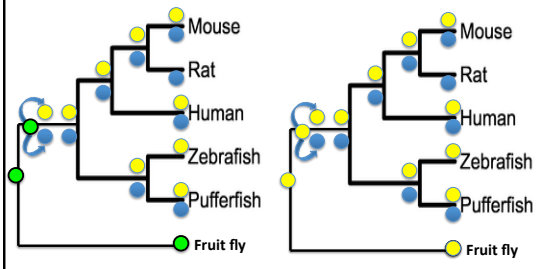



Gene originates before common ancestor. Duplicates. … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

Gene originates in common ancestor … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

Problem: distinguishing paralogs from orthologs

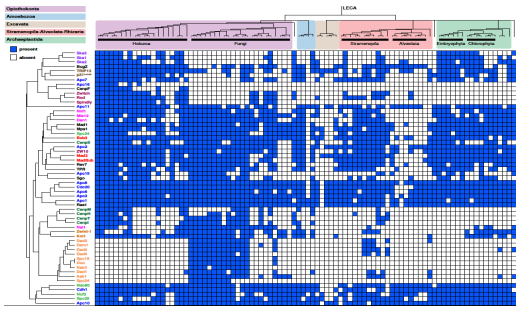
### Same scenario?




In blast they are likely the same …

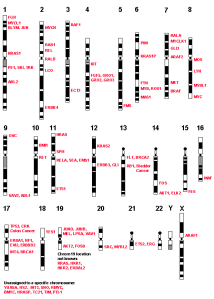
Problem: same gene evolution vs same gene function

### Phyletic patterns of basal cellular processes across eukaryotes: what do we see?

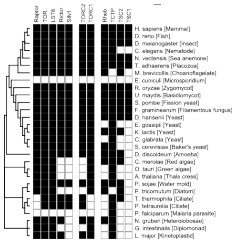


Expanded from Dev Cell. 2012 Aug 14;23(2):239-50. doi: 10.1016/j.devcel.2012.06.013. Evolution and function of the mitotic checkpoint. Vleugel M, Hoogendoorn E, Snel B, Kops GJ.

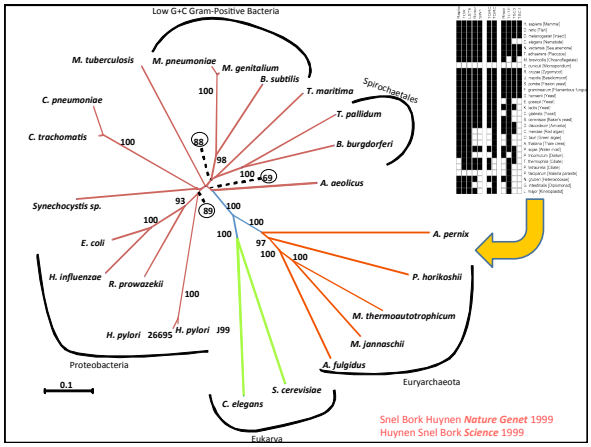
## (protein coding) Gene Content



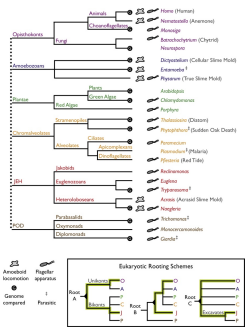
bag of genes

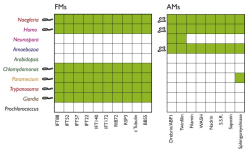


Why does the genome contain the genes that it does



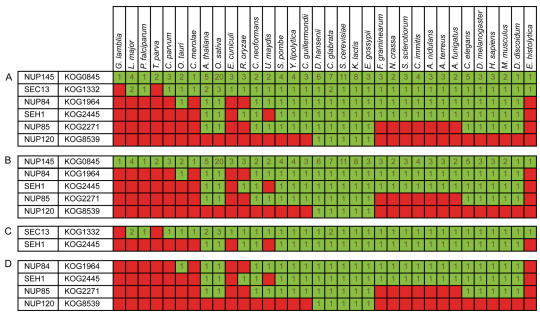

Snel Bork Huynen *Nature Genet* 1999
Huynen Snel Bork *Science* 1999





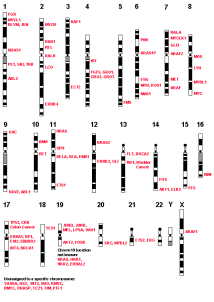
Fritz-Laylin et al. cell 2010

## … however

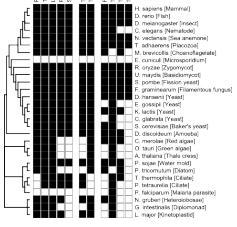



Fokkens and Snel *PLoS Comp Biol* 2009

## Presence of genes across genomes: ORTHOLOGY

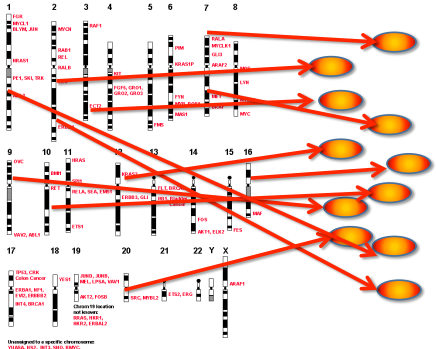


bags of genes

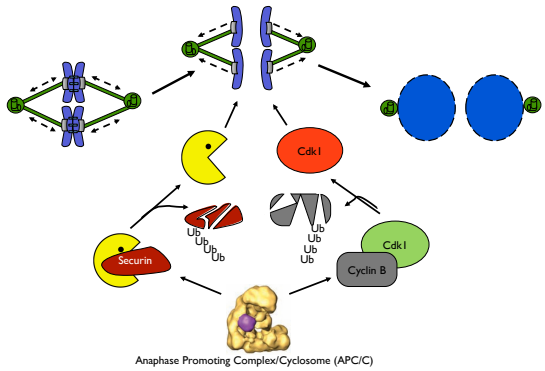


Why does the genome contain the genes that it does

## Bag of genes, what about the homologs within a genome? (paralogs, duplicates)

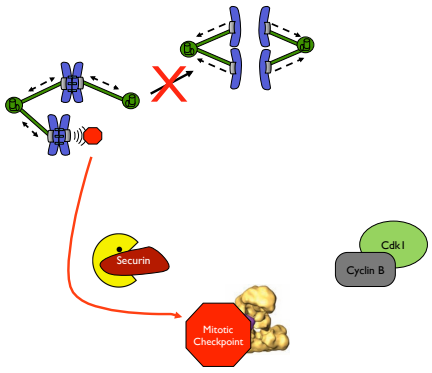


## Initiating Anaphase

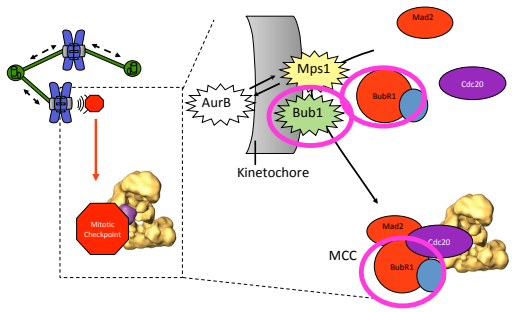



Anaphase Promoting Complex/Cyclosome (APC/C)

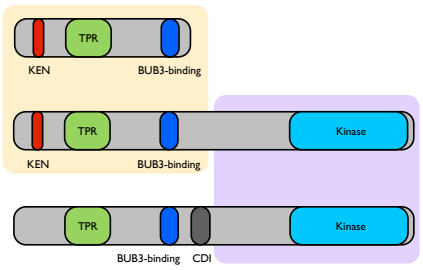
## The Mitotic Checkpoint

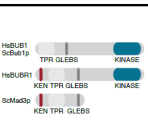



## The Mitotic Checkpoint Complex (MCC)

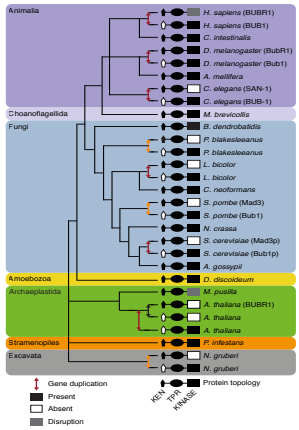



## scMad3p (fungi)



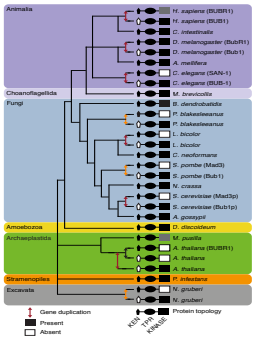



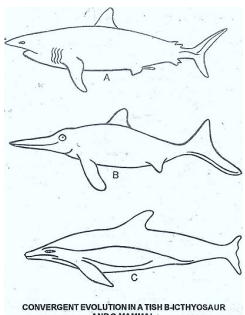

9 independent duplications.
7 cases where a mad3-like and a bub1-like protein arose out of a bubmad-like ancestor.



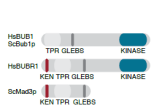

Suijkerbuijk *et al.*, Snel and Kops *Dev Cell* 2012

## Recurrent (convergent/parallel) evolution in molecular systems!

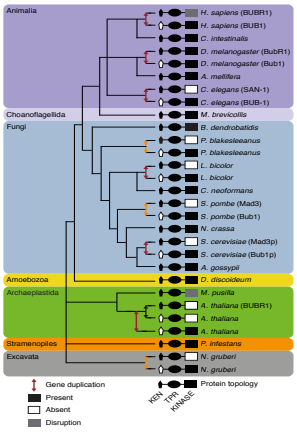





CONVERGENT EVOLUTION IN A TISH B-ICTHYOSAUR AND C-MAMMAL.




What about the kinase domain in human (and fly)




## What about the kinase domain in human bubr1?

a

| | Gly-rich loop I | II | III | Catalytic loop VI-B | VII |
|---|---|---|---|---|---|
| Consensus | GxGxxGxV | VAIK | E | HRDxKxxN | DFG |
| PKA | GTGSFGRV 52 57 | YAMK 72 | E 91 | YRDLKPEN 166 168 171 | DFG 184 |
| HsBUB1 | GEGAFAQV 796 801 | FVLK 821 | E 830 | HGDIKPDN 917 919 922 | DLG 946 |
| HsBUBR1 | CEDYKLF- 777 781 | TVIK 795 | D 804 | HGDLSPRC 862 864 867 | DFS 911 |

degeneration of motifs essential for catalysis

Further experiments showed vertebrates are not exception. The kinase domain of BubR1 lacks enzymatic activity.

*"This explained the field's inability to identify substrates of BubR1, and dispelled a leading theory of SAC silencing based on inactivation of BubR1 after kinetochore-microtubule attachment."*

## This course

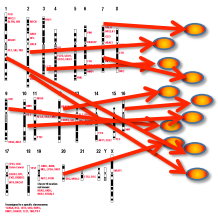
- I want study evolution of genomes pathways and networks, so that is why I study gene/protein evolution
- At the end, able to analyse your own protein
- Understanding that many bioinformatic challenges are a mix of conceptual and technical problems (e.g. why orthology is such an incredibly persistent problem)
- "what you should ~know" in order to this kind of research
- Topics are interrelated
  - e.g. orthology already in homology lecture but proper explanation a day after
  - e.g. that trees can be used to time a duplication to eukaryogenesis but proper discussion of eukaryogenis has its own lecture

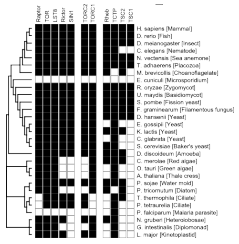
## Small scale & Large Scale

- How did my protein, complex, pathway evolve? (collaborations)(COO, mini project)
- Large scale, how do genome, networks and complexes evolve (context/expectation, bioinformatics senior authorships)(paper discussions)
- What can we infer about eukaryogenesis?

## Homology (& domains)

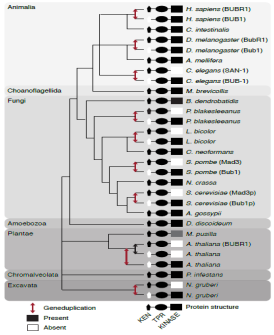
- Absolute basis of any comparative analysis, affects MSA and trees, detection still being improved,



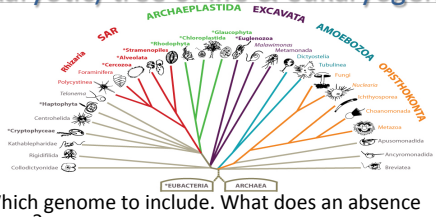


## Gene Phylogeny & Orthology

- How do we get such trees **and how do we interpret them**
- Trees reveal some of the most important genome evolution processes: LGT, duplication, loss,



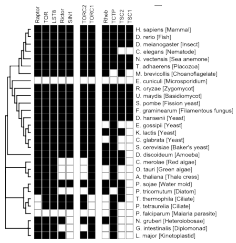

## (Eukaryotic) tree of life & eukaryogenesis



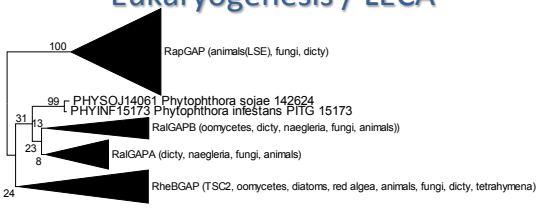

- Which genome to include. What does an absence mean?
- Essential for interpreting gene trees:
  - Knowing (at least the outline) by heart >>> having to look it up
- With regards to evolutionary signaling cell biology ( kinases, smallgtpases etc. )the diversity in present day genomes is staggering and dwarfs e.g. human-fruit fly difference

## Large scale orthology

- Needed to move beyond anecdotes, but difficult to get

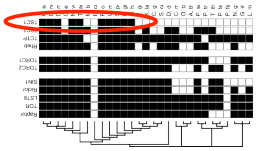


## Eukaryogenesis / LECA



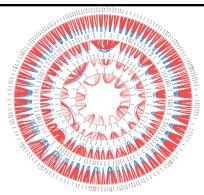

- Biological topic, eukaryogensis / LECA for which these types of analyses are telling us a lot. But it also impacts a lot of things we do: we see it back in gene trees and it impacts getting orthologous groups across eukaryotes.

## Gene content evolution



- Fundamental level of genome evolution
- Gene invention -> inability to detect homologs vs real lack of homologs does not simply mean novel gene
- Evolutionary modules?
- Trying to move large scale but remember the pitfalls

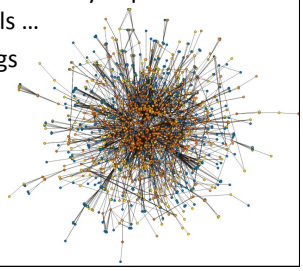
## Whole Genome Duplication (WGD)



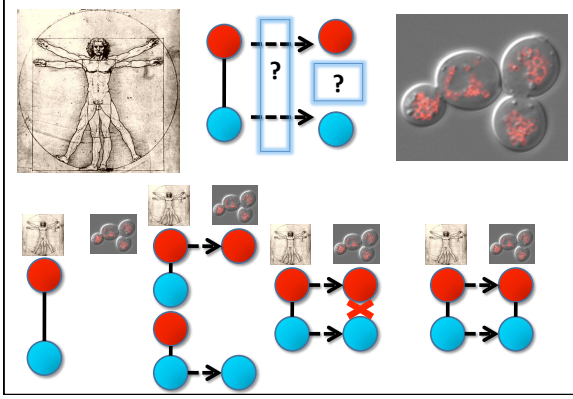
- Like LECA, WGD is important biology for which bioinf needed to research but which also impacts our data
- And which is welcome source of information for our analyses (Lidija, bubmad): independent and reliable reconstruction of the history of part of the history of genes

## Using HTP data to study evolution of networks / complexes

- Is the number of conserved interactions between e.g. yeast and human 10% or 95%???
- On top of all the genome analysis pitfalls also all the HTP data pitfalls …
- Duplicates vs orthologs



## What can happen to an interaction in evolution




## Techniques AND biology

- Detective/forensics vs concepts; Large scale biology vs small-scale biology; Bioinformatics biology vs data/technique problems;
- A lot like police investigation … and less like Nobel prize winning physics …
- Anything goes in genome evolution; many processes often entangled (i.e. google subneofunctionalization)

## Practical stuff

- Schedule
- BYOD
  - Tuesday March 7
  - Tuesday March 14
- Literature discussion (have you already read Zmasek et al?)
  - You should have read the papers in depth before the discussion
  - I will shortly introduce and then invite people to discuss figures / pieces of the results
  - This + participation in the discussion is 20% of grade
- Mini project, let me first explain some bioinformatics … than tomorrow afternoon let's discuss it & pick proteins

## Computer Exercises

- Mostly use of web resources.
- Computer exercises for some topics many others more difficult (i.e. evolution of interaction networks based on HTP analysis).
- Ask help from fellow students.
- Should tie strongly into mini-projects
- (I am slightly afraid the data bases are getting unwieldy w.r.t. number of genomes … searches very slow … you need to already know the ToL to pick relevant species)

## Requests

- **very** heterogeneous with respect to previous knowledge (IBMB, GB, research projects, PhD students)
- PLEASE: interrupt / ask questions when I am going to fast, when I use jargon, when I make jumps/conclusions that to me seem obvious 100% logical, but to your are erratic; please point out my implicit assumptions regarding what everybody knows
- Computer exercises: more experienced people help
- And also apologies for some redundancy

## Mini project

- **The protein.**
- **What does my protein look like (protein topology)**
- Optional: Size of the (super)family in the genome you're sequence is from
- **Homologs across tree of life**
- **Tree of relevant sequences in diverse genomes**
- **Orthologs in genomes from your tree. (or from homology searches)**
- Optional: Does your potein or any of its orthologs in other species have Whole Genome Duplicates (WGD)/ Ohnologs?
- Optional: Point of invention of the eukaryotic orthologous group your protein belongs to.
- Optional: interactions of proteins in your tree according to biogrid
- Optional: orthology of interactors of your proteins according to biogrid and an automatic ortology database such as. E.g. panther.

## Mini project

- Species tree, you really get to know the outline if you are using the ToL to describe the evolution of a protein. Similarly for e.g. smart/pfam etc.
- Students are often finished too long after the course … for your own benefit try to prevent that
- Some students get stuck on is what they find novel. It does **not** have to be novel! Just describe what you find!

## Mini project / Molecular evolution is recursive / iterative 3: generalized

- To study the a evolution of a gene you need a model / framework of the evolution of the gene, but to get an idea of a proper framework / model of the evolution of a gene, you need the need to study the evolution of a gene
- Thus: heuristics & build on previous results. Start from stuff you trust (alignment of highly identical sequences), and/or only the use the general but flawed overview (e.g. guide tree). Then iterate
- Not yet so automatically solved for evolutionary history of a gene and its homologs as it is for other case …