CENTER FOR BIOLOGICAL COMPUTATION initiative

ARISTOTLE UNIVERSITY OF THESSALONIKI

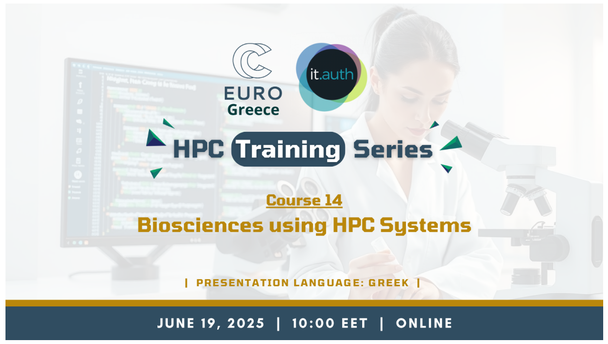
ARISTOTLE UNIVERSITY OF THESSALONIKI
KNOWLEDGE I NNOVATION CULTURE

BCCB

Bespoke comparative genomics software architectures

a case-study for future HPC on Aristotelis

Biosciences using High Performance Computing (HPC) Systems



BERKELEY LAB
Lawrence Berkeley National Laboratory

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

Christos A. Ouzounis cao@csd.auth.gr • 19 jun 2025

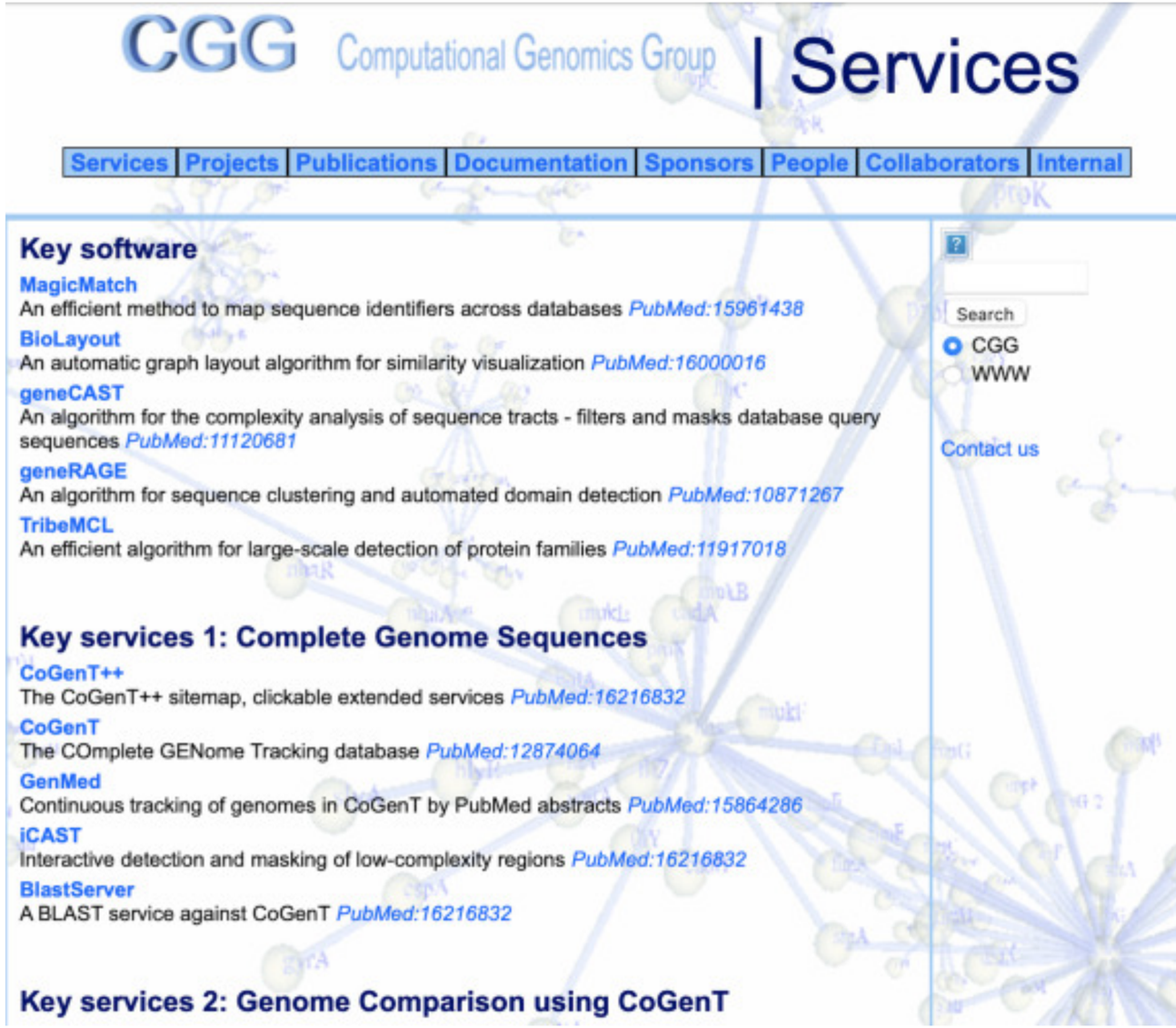
elixir GREECE

HB H.bioinfo

HPC for Bioinformatics

cgg_toolkit Hands-On Demo

need a short intro to concepts?
and some elements of HPC



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011498

life is complex matter

an information universe

- biosciences: PBytes of information with large-scale installations
- analysis of these data $\Rightarrow$ top-five players of Big Data
- along with financial, military, comms and physics/astronomy
- noisy, multi-scale: ideal playground for complex systems

central dogma
from DNA to protein

- DNA contains 'genes'
- .. that encode
- .. (via an intermediate, mRNA)
- .. proteins

- both have distinct alphabets
  - DNA: 4 characters {ACGT}
  - proteins: 20 characters {aa's}

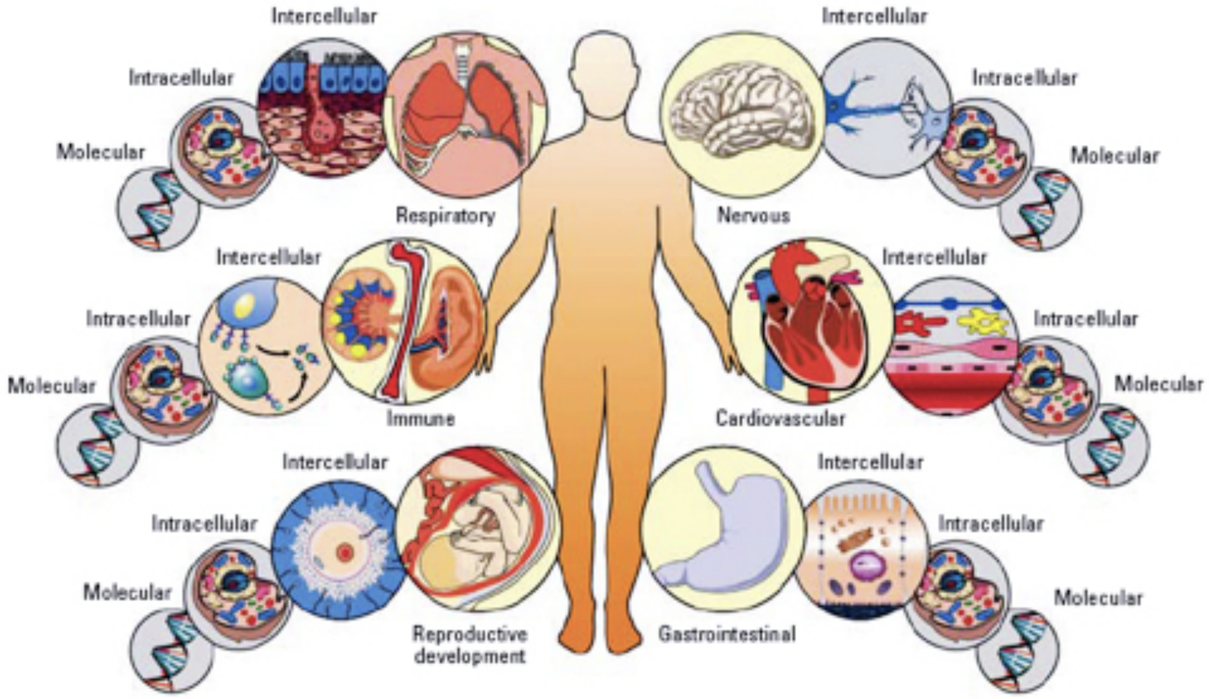
```
ATGCCCGTTGCCCACGTTGCCTTGCCCGTTCCGCTTCCTCGTACCTTTGACTATCTGCTGCCAGAAGGCATGACGGTTAAAGCTGGGTGTCGCGTGCGCGT
GCCGTTTGGCAAACAGCAGGAGCGCATCGGGATTGTGGTATCAGTTAGCGATGCCAGCGAACTGCCGCTCAATGAGCTAAAAGCGGTAGTCGAAGTGCTGG
ATAGTGAGCCGGTGTTTACTCACTCCGTCTGGCGATTGCTGCTATGGGCGGCAGATTACTATCATCATCCGATTGGCGATGTGCTGTTTCATGCCTTGCCG
ATTTTACTACGCCAGGGGCGGCCTGCGGCGAACGCGCCGATGTGGTACTGGTTTGCCACTGAACAAGGCCAGGCGGTGGATCTGAACAGCCTGAAACGCTC
CCCCAAGCAACAACAGGCGCTGGCGGCGTTACGGCAAGGCAAAATCTGGCGCGACCAGGTCGCCACGCTCGAATTTAATGATGCCGCGTTGCAGGCGCTAC
GCAAAAAAGGTCTGTGTGATTTAGCAAGTGAAACACCAGAGTTTAGCGACTGGCGAACGAACTATGCCGTTTCTGGTGAGCGGTTGCGATTGAATACCGAA
CAGGCCACCGCCGTTGGCGCAATTCATAGCGCGGCAGATACTTTTTCTGCCTGGCTGCTGGCGGGCGTTACCGGTTCCGGTAAAACGGAGGTTTATCTCAG
CGTACTGGAAAACGTGCTCGCTCAGGGCAAACAGGCGCTGGTGATGGTGCCGGAAATCGGCCTGACACCGCAAACTATCGCCCGTTTTCGTGAACGTTTTA
ATGCCCCCGTGGAAGTTCTGCATTCCGGCCTGAACGACAGCGAGCGTCTTTCGGCGTGGCTGAAAGCGAAAAATGGTGAGGCGGCGATTGTGATCGGCACC
CGCTCCGCGCTGTTTACGCCGTTTAAAAATCTCGGCGTGATTGTCATTGATGAAGAGCACGACAGCTCCTACAAGCAGCAGGAAGGCTGGCGCTATCATGC
CCGCGACCTGGCGGTGTATCGTGCGCACAGCGAGCAAATCCCGATTATTCTTGGCTCCGCAACGCCCGCGCTGGAAACGTTATGCAACGTCCAGCAGAAAA
AATACCGCCTGCTGCGCCTGACCCGTCGGGCAGGGAATGCGCGTCCGGCAATTCAACATGTGCTGGATTTAAAAGGTCAGAAGGTGCAGGCAGGTCTGGCT
CCGGCGTTAATCACTCGTATGCGCCAGCATTTACAGGCTGATAACCAGGTCATTCTCTTTCTTAACCGCCGTGGCTTTGCGCCTGCACTGCTGTGCCACGA
CTGTGGCTGGATTGCCGAATGCCCACGTTGCGATCACTACTACACGCTGCATCAGGCGCAGCACCATCTGCGCTGCCACCACTGTGACAGTCAGCGTCCGG
TGCCGCGCCAGTGCCCTTCCTGCGGTTCCACGCACCTGGTCCCCGTGGGGCTGGGCACCGAACAGCTTGAACAGACGCTCGCGCCGTTGTTCCCCGGCGTG
CCCATTTCTCGTATCGACCGCGATACCACCAGCCGCAAAGGGGCGCTGGAACAGCAACTGGCAGAAGTACATCGCGGCGGCGCGCGGATTTTGATTGGTAC
ACAAATGCTGGCGAAAGGTCACCATTTCCCGGATGTGACGCTGGTTGCATTACTGGACGTGGACGGCGCGCTGTTTTCTGCCGATTTTCGCTCGGCAGAGC
GTTTCGCTCAGCTTTACACCCAGGTCGCCGGTCGTGCCGGGCGTGCGGGTAAACAGGGCGAAGTGGTGCTGCAAACGCACCATCCGGAACATCCTCTGTTG
CAAACGTTGCTCTATAAAGGCTACGACGCCTTTGCCGAACAGGCGCTGGCTGAGCGGCGAATGATGCAGCTACCGCCGTGGACCAGCCATGTGATTGTGCG
TGCGGAAGATCATAACAATCAGCACGCGCCATTGTTCCTGCAACAACTGCGTAATCTGATCCTCTCCAGCCCACTGGCAGACGAGAAACTGTGGGTTCTCG
GTCCGGTTCCGGCTCTGGCACCTAAACGTGGCGGTCGCTGGCGCTGGCAGATATTGTTGCAGCACCCTTCCCGCGTGCGCTTGCAACACATCATTAACGGT
ACGCTGGCGCTCATCAATACAATACCGGATTCCCGTAAGGTGAAATGGGTGCTGGATGTTGATCCGATTGAGGGTTAA
```

```
MPVAHVALPVPLPRTFDYLLPEGMTVKAGCRVRVPFGKQQERIG
IVVSVSDASELPLNELKAVVEVLDSEPVFTHSVWRLLLWAADYY
HHPIGDVLFHALPILLRQGRPAANAPMWYWFATEQGQAVDLNSL
KRSPKQQQALAALRQGKIWRDQVATLEFNDAALQALRKKGLCDL
ASETPEFSDWRTNYAVSGERLRLNTEQATAVGAIHSAADTFSAW
LLAGVTGSGKTEVYLSVLENVLAQGKQALVMVPEIGLTPQTIAR
FRERFNAPVEVLHSGLNDSERLSAWLKAKNGEAAIVIGTRSALF
TPFKNLGVIVIDEEHDSSYKQQEGWRYHARDLAVYRAHSEQIPI
ILGSATPALETLCNVQQKKYRLLRLTRRAGNARPAIQHVLDLKG
QKVQAGLAPALITRMRQHLQADNQVILFLNRRGFAPALLCHDCG
WIAECPRCDHYYTLHQAQHHLRCHHCDSQRPVPRQCPSCGSTHL
VPVGLGTEQLEQTLAPLFPGVPISRIDRDTTSRKGALEQQLAEV
HRGGARILIGTQMLAKGHHFPDVTLVALLDVDGALFSADFRSAE
RFAQLYTQVAGRAGRAGKQGEVVLQTHHPEHPLLQTLLYKGYDA
FAEQALAERRMMQLPPWTSHVIVRAEDHNNQHAPLFLQQLRNLI
LSSPLADEKLWVLGPVPALAPKRGGRWRWQILLQHPSRVRLQHI
INGTLALINTIPDSRKVKWVLDVDPIEG
```

DNA changes at very slow rates for Myrs ⟩ EVOLUTION

```
>gi|82583714|emb|CT009680.1| Human chromosome X complete sequence
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTCTGAAAGTGGACCTATCAGCAGGATGTGGGTG
GGAGCAGATTAGAGAATAAAAGCAGACTGCCTGAGCCAGCAGTGGCAACCCAATGGGGTCCCTTTCCATA
CTGTGGAAGCTTCGTTCTTTCACTCTTTGCAATAAATCTTGCTATTGCTCACTCTTTGGGTCCACACTGC
CTTTATGAGCTGTGACACTCACCGCAAAGGTCTGCAGCTTCACTCCTGAGCCAGTGAGACCACAACCCCA
CCAGAAAGAAGAAACTCAGAACACATCTGAACATCAGAAGAAACAAACTCCGGACGCGCCACCTTTAAGA
ACTGTAACACTCACCGCGAGGTTCCGCGTCTTCATTCTTGAAGTCAGTGAGACCAAGAACCCACCAATTC
CAGACACACTAGGACCCTGAGACAACCCCTAGAAGAGCACCTGGTTGATAACCCAGTTCCCATCTGGGAT
TTAGGGGACCTGGACAGCCCGGAAAATGAGCTCCTCATCTCTAACCCAGTTCCCCTGTGGGGATTTAGGG
GACCAGGGACAGCCCGTTGCATGAGCCCCTGGACTCTAACCCAGTTCCCTTCTGGAATTTAGGGGCCCTG
GGACAGCCCTGTACATGAGCTCCTGGTCTGTAACACAGTTCCCCTGTGGGGATTTAGGGACTTGGGCCTT
CTGTCTTTGGGATCTACTCTCTATGGGCCACACAGATATGTCTTCCAACTTCCCTACACAGGGGGGACTT
CAAAGAGTGCCTTGAGCTGATCTGGTGATTGCTTTTTTGTACTGTTATTTATCTTATTCTTTTCATTGTG
AGGTACTGATGCAAACACTTTGTACGAAAAGGTCTTTCTCATCTCGGGAGTCCCCGTCTATTTGTCCCGG
TCCCTGTTAACCCAGTCCCCGACAGGAGCCCCTTCTGCACCTTGAGCTCTCACCACTCACCGTCCATCCA
GCCCCAGCTCTGCCTGCAACCCACCCATCCCTGGGACTCGGGCCTCCCCTCTCTAGTGGTCTGGTCATCA
GGCCAGGGGCACGTGGAAGAAGCTATCGTGGCAAAGGGAGCAGTCATATCCCCAAAATCTGTGGTTGGTT
TACCACCACCATGGAAACCCCAGGGTGGGACTCTAGTTTCAGGTTGGAGCTGAGCCCTGTCGGGAATGAG
CTTTCCCCAGCTATGGCTTCTTGGGGCCCCTGTGCCCTGAGCTGTGTCTCCCAGCATCGGGTCCCCACCA
TGCATATGGCCCACTCAGGCACAGTGCCGCGATGGCTGCATGCGTGAGGGGGGCCTGGGCCCAGGGCTGG
GAGTCCTTTGTGTCTCATGGCCATGATTGTCCTTCCGAGTATGATATGGTGGCCAATTTCTTTTATTCTG
TCGTTCAGAGTGAGTAAATGATGTAGAGTTCATGCAGAAAAAAATACAACAAAAACCAAGGGAACATAGA
ATTGGAAAACGCGTCACAGCAATGAGTTAAATAGGTAACAAATTTCATCATTTGAAGAAAGACTTAGAGT
GCCAAAAGTGCCTCTTAAGTCTCCTTTAAAAAGTAGCAAAATTCATCCCTGAAGAAGCATCTTGGCCTTT
TTCATGTACTCAGAGTGCTGGTGAAGAACAAAGATTGCTGAAACATTATGTACCTAACAGCGTTACAGGG
TGTAGATAACACACTGGAAAACCTGGTCGTTACAGTGGACATATTCCAGGAAGTCCTTGCCTGAGGTTTT
CCAAGTTATGGAATTGCTTGAGATTGGAAGAGGCGATGGAGGGTACAACTGTAATGCCCAACCTCATTTT
TGCTAACCCTGTTTTTAGACTCTCCCTTTCCTTCAATCACCTAGCCTTCTTTCCACCTGAAAGGACTCTC
CCTTAACTGAGAGAACCGGACAGACTCCATCTTGGCTCTTTCACTGGCAGCCCCTTCCTCAAAGACTTAA
CTCGTGCAAGCTGACTCCCAGGACATCCGAGAATGCAATTAACTGACAACCTACTGTGGCGAGCTACATC
CGCAGTCCCCAGGAATTCGTCCGATTGATAACGCCCAATTACCCGCGTCTATCACCTTGTAATAGTCTTA
AAGCACCTGCACCTGGAACTGTTTACTTTCCTGTAACCATTTATCCTTTTAACATTTTGCCTGATTTACT
TATGTAAAATTCTTTTAACTAGACCGCCACTCCCCTTTCTAAACAAAAGTATAAAAGAAAATCTAGCCCC
TTCTTTGGGACTGAGACAATTTTGAGGTTAACGCAGGGTGCCTGTAATCCTAAGGGAGGAGACCGCCACT
TCTGCTGCCCTTCCCTTCCCCACACCCCCTTCTCTAGTTTATGAAACAGGGAAAAAGGGAGAAAGCAAAA
AGATAAAAAAAACAGAAGTAAGATAAATAGCTAGACGACCTTGGCAGCACCACCCGGCACTGGTGGTTAA
AATAATAATAATAATAATATTAACCCCTGACCTAAACTACTTGTGTTATCTGTAAATTCCAGACACTGTA
TGAGGAAGCCCTGCAAAACTTTCTGTTCTGTTATCTGATGCGTGTAGCCCCCAGTCACGTTCCGATGCTT
GCTCGATCTATCACGACCCTTTCAAGTGAACCCCTTAGAGTCGTAAACCCTTAAAAGGGCCAGGAATTTC
GTTTTCGGGGAGCTCGGCTCTTCAGGCCCAAGTAAACCTGCCGTATCTCACCTGAGACCAACCCCCAACT
ACAAAACTCAACCTGGAATTTTCCCAGGACCAAACCCATCTATATTCTGTAACCCGAAACCTCAAAGCCT
AACCCTAACCCTAACCCCTACAGTTGAGGTCCCCCCGCCCCTGTGGTTCCAGCTCAAGACAACCTGCCCC
TCCGTGGGTTTGCAGGCCCTCTGGTGGGGGTGGGAGCTGGGGGCCACATACAGCTCTCTGAGCTTAAGCC
```

species share many common characteristics ⟩ HOMOLOGY

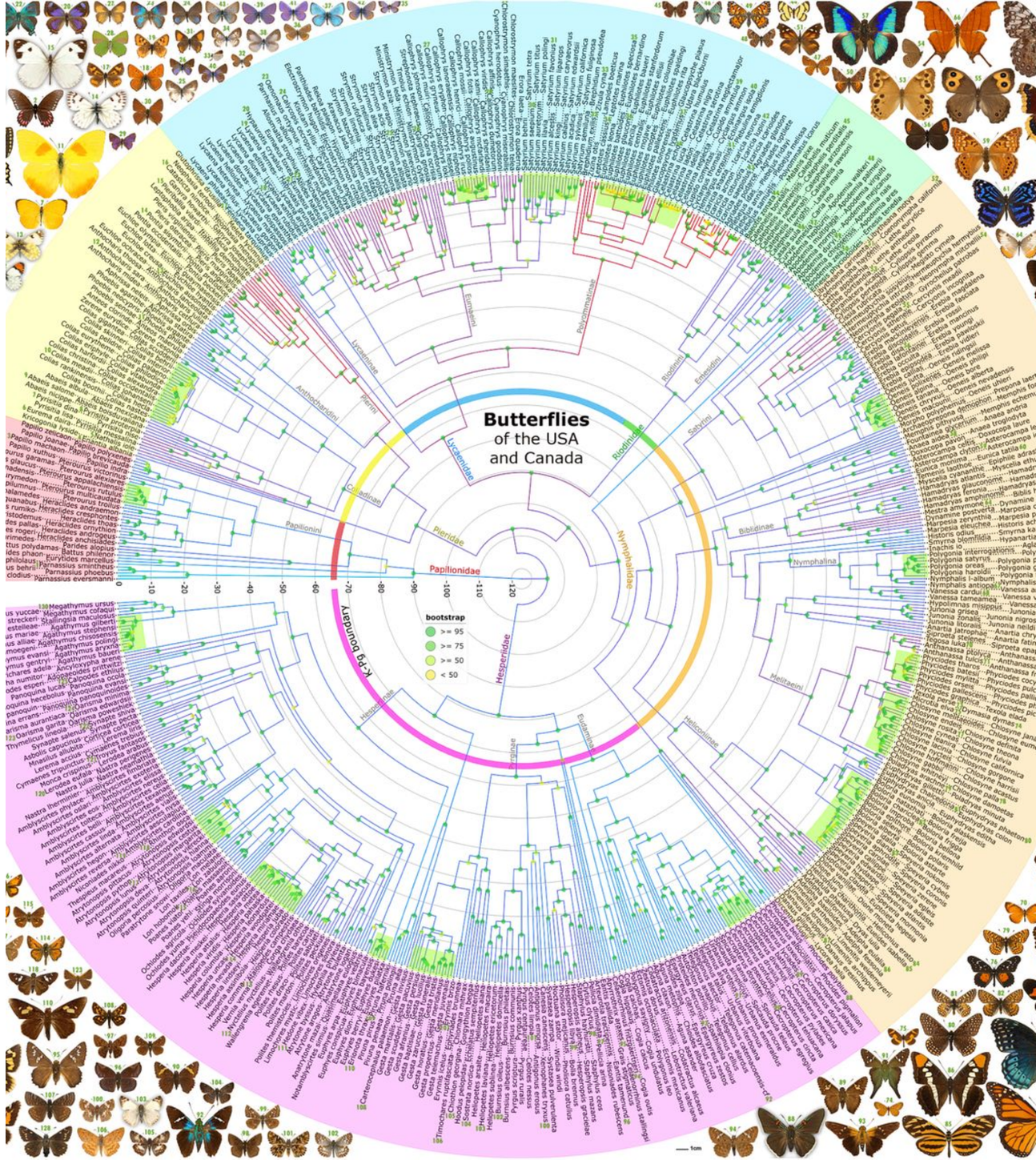


HPC

dozens, 100s, 1000s of cores

- *Runs complex, parallel workloads*
  - *used for physics, climate, genomics*

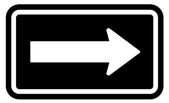
- *HPC in the biosciences*
  - *primarily for genome science, also ecosystems*
  - *many comparisons, combinatorial explosion*
  - *time-critical (e.g. COVID-19)*



today

on HYPATIA @elixir-greece

- CPU pool VM
  - 96 CPUs
  - 1TB RAM
  - 2.4 Ghz
- Intel Xeon Cascadelake
  - >50 TB space
  - shared resource
  - project-based access etc.



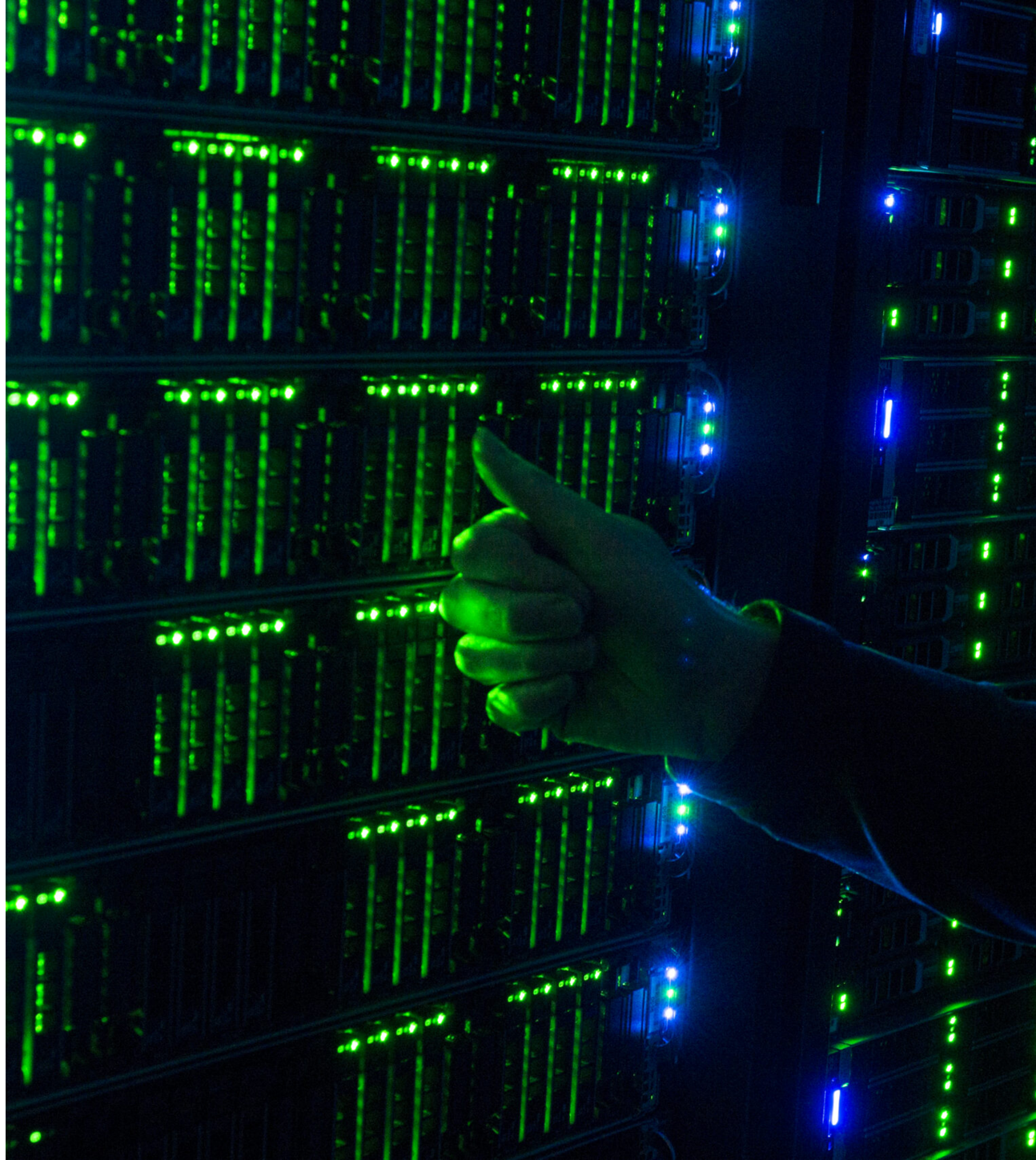
```
Architecture:                    x86_64
CPU op-mode(s):                  32-bit, 64-bit
Byte Order:                      Little Endian
Address sizes:                   40 bits physical, 48 bits virtual
CPU(s):                          96
On-line CPU(s) list:             0-95
Thread(s) per core:              1
Core(s) per socket:              1
Socket(s):                       96
NUMA node(s):                    1
Vendor ID:                       GenuineIntel
CPU family:                      6
Model:                           85
Model name:                      Intel Xeon Processor (Cascadelake)
Stepping:                        6
CPU MHz:                         2400.004
BogoMIPS:                        4800.00
Hypervisor vendor:               KVM
Virtualization type:             full
L1d cache:                       3 MiB
L1i cache:                       3 MiB
L2 cache:                        384 MiB
L3 cache:                        1.5 GiB
NUMA node0 CPU(s):               0-95
```

cgg toolkit

collection of tools, implicit

- Runs on HPC CPU pool
  - SLURM
  - parallel: MPI, multithreading

- A comparative genomics suite
  - input: genomes, annotations, metadata, data fusion
  - processing: mapping, masking, matching, clustering, visualization
  - output: matrices, clusters, etc.



https://github.com/bcpl-certh/cgg-toolkit

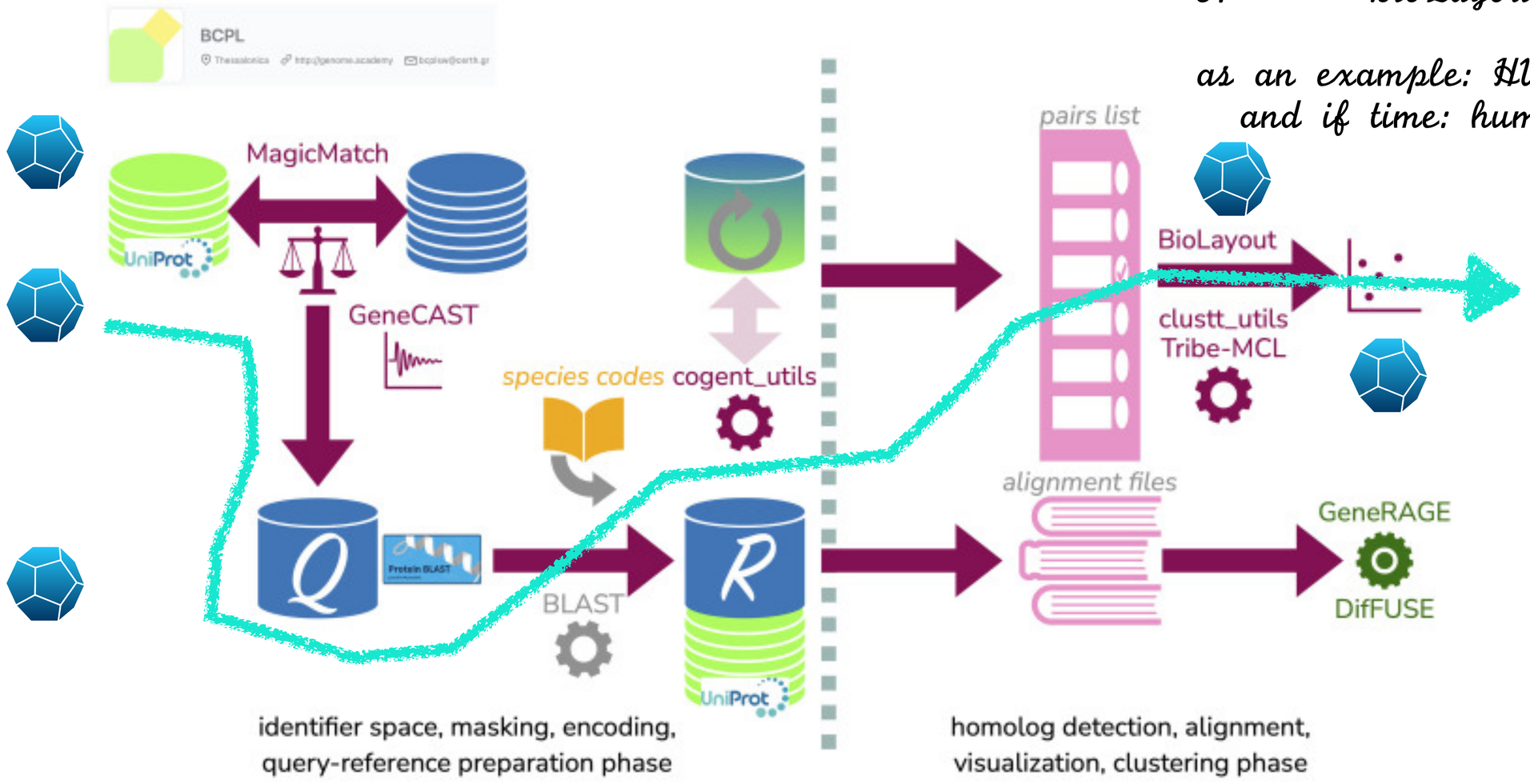
cgg toolkit

multiple decision points for analysis

steps today:

1. magicmatch (IDs)
2. CAST masking
3. BLAST search
4. MCL clustering
5. BioLayout visuals

as an example: HUMAN GENOME
and if time: human vs plant



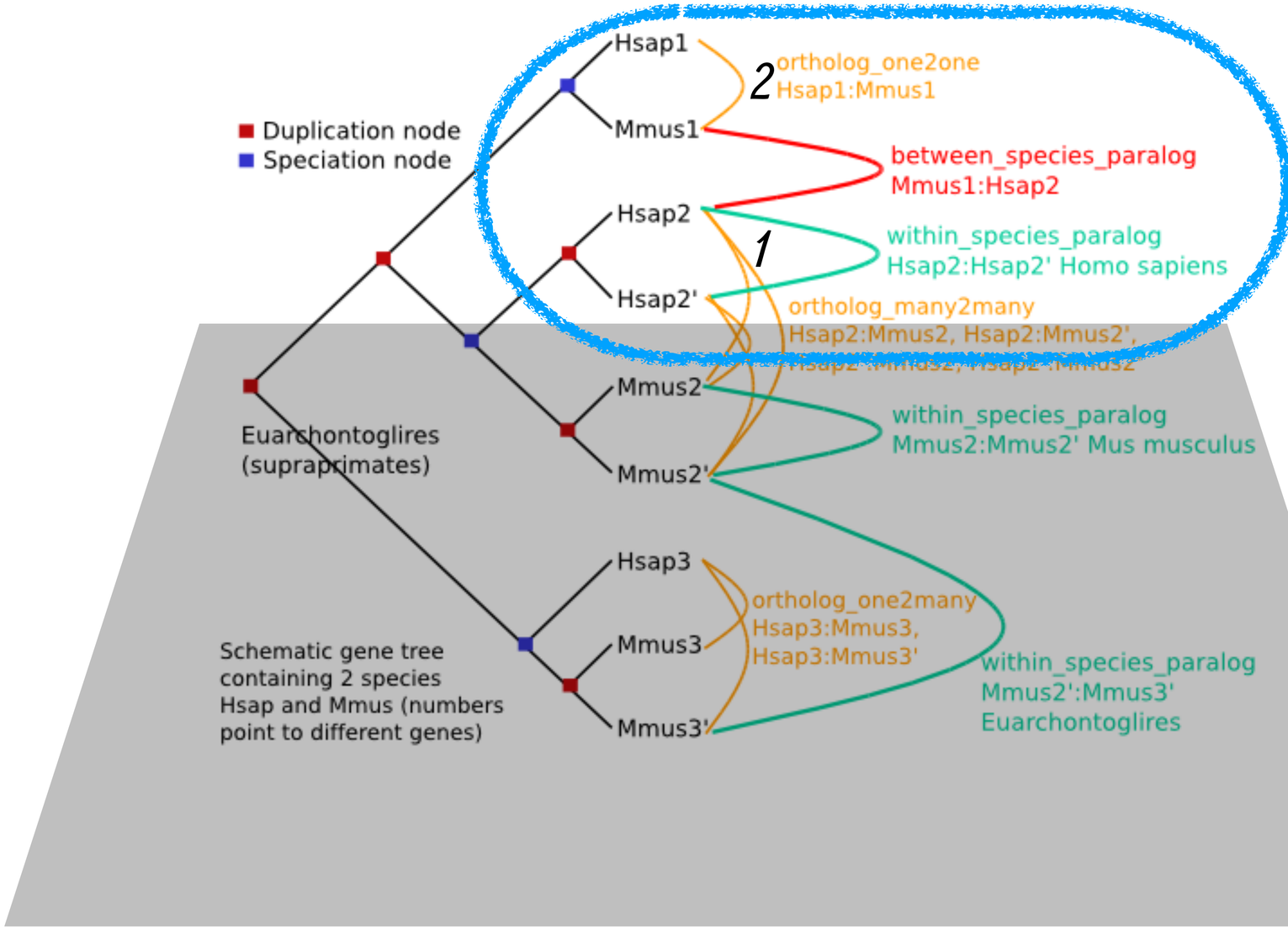
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011498

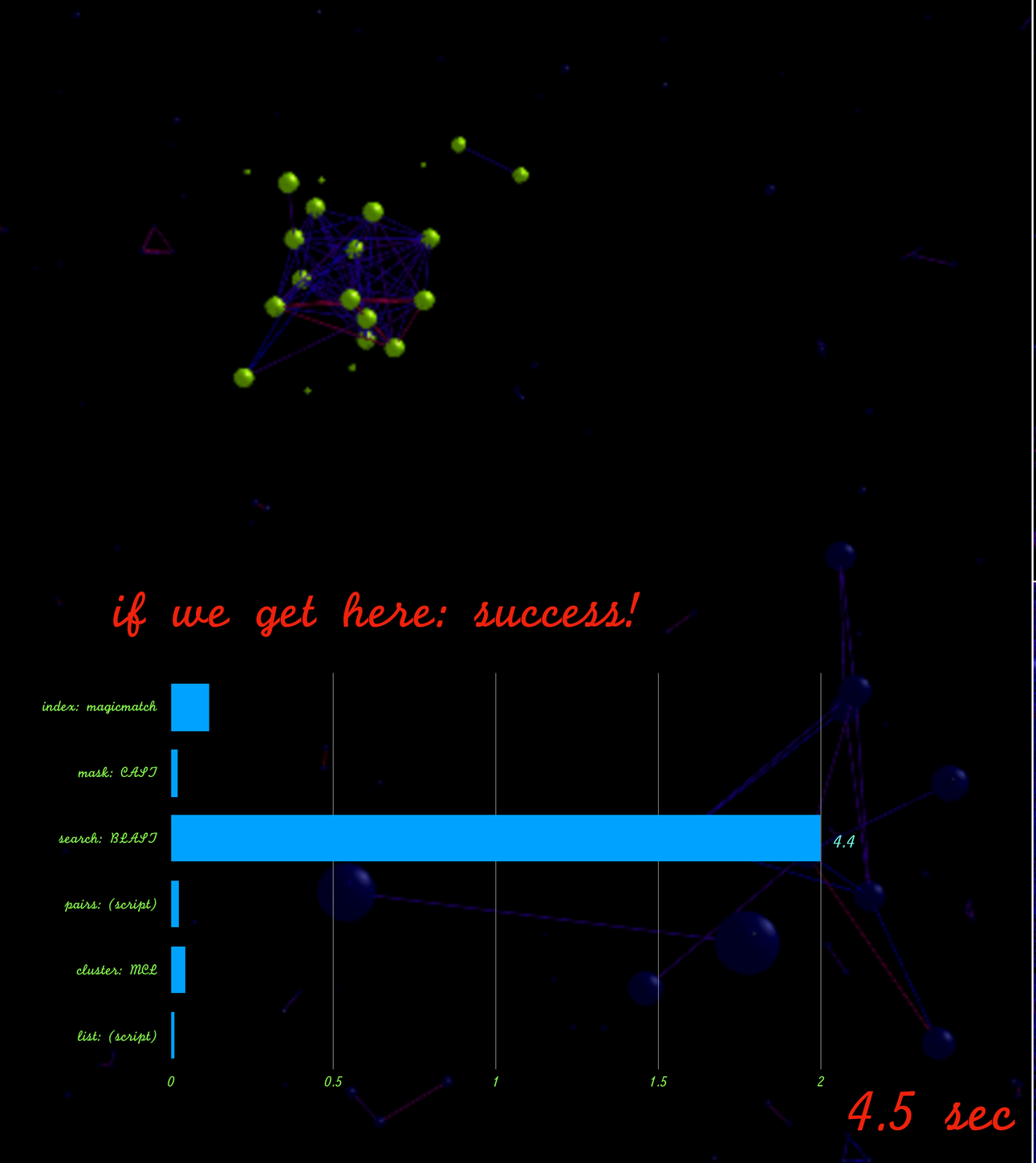
live demo

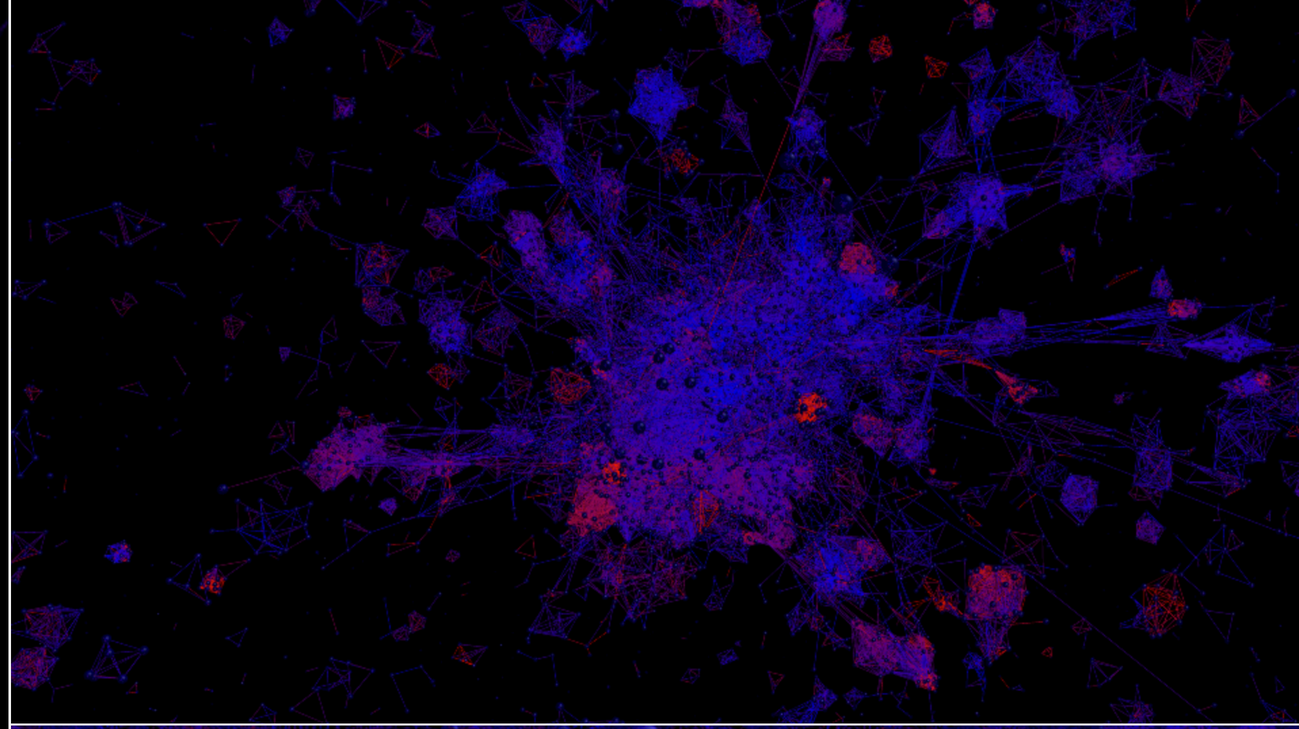
step-by-step

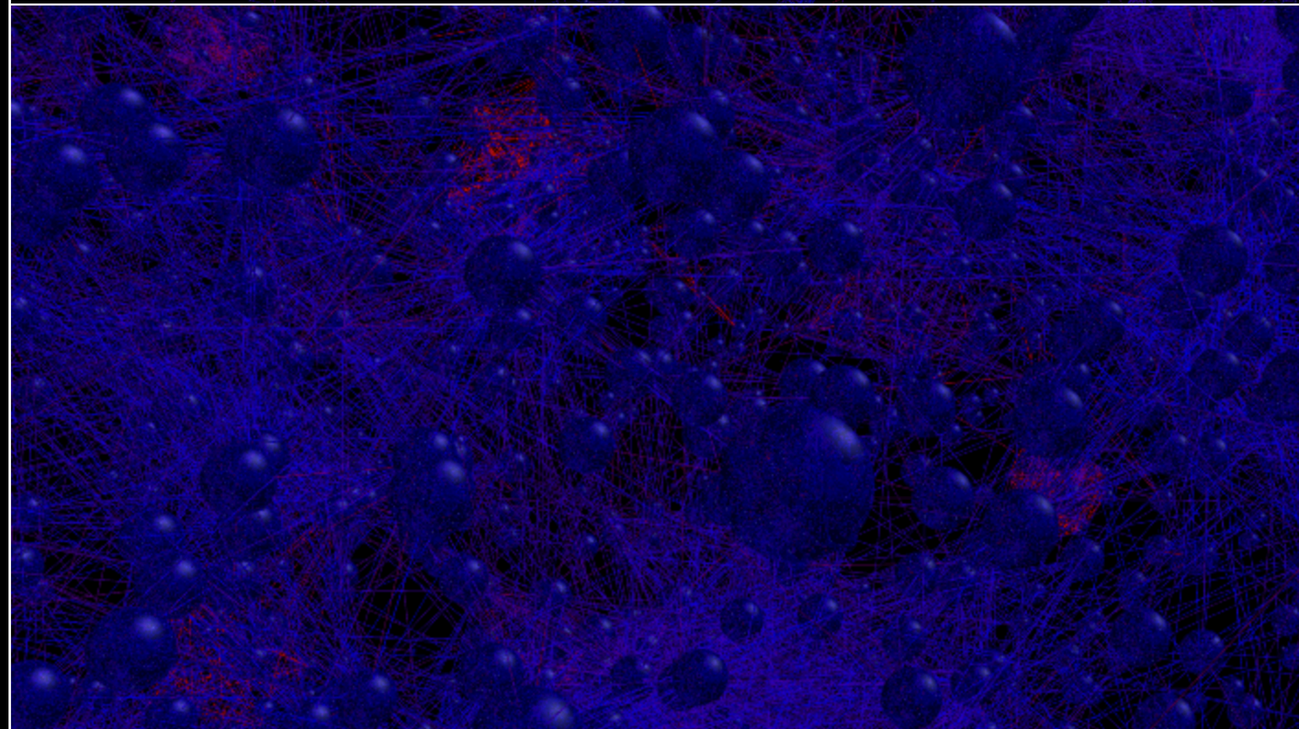
- typically automated
  - will explain config
  - will submit jobs one-by-one

- will view results
  - explain
  - discuss
  - and, if time, visualize




if we get here: success!
index: magicmatch
mask: CAPT
search: BLAST
4.4
pairs: (script)
cluster: MCL
list: (script)
0
0.5
1
1.5
2
4.5 sec





conclusion(s)

and a couple of recommendations

- HPC is essential
  - for bioinformatics, much more
  - reproducibility & scalability
    - all genomes: 1bn sec = 30 yrs
- resources
  - get as much storage as you possibly can !!
    - CPU power OK, not crucial
  - prepare for REALLY big data !!

Q&A

http://genome.academy/how/links