

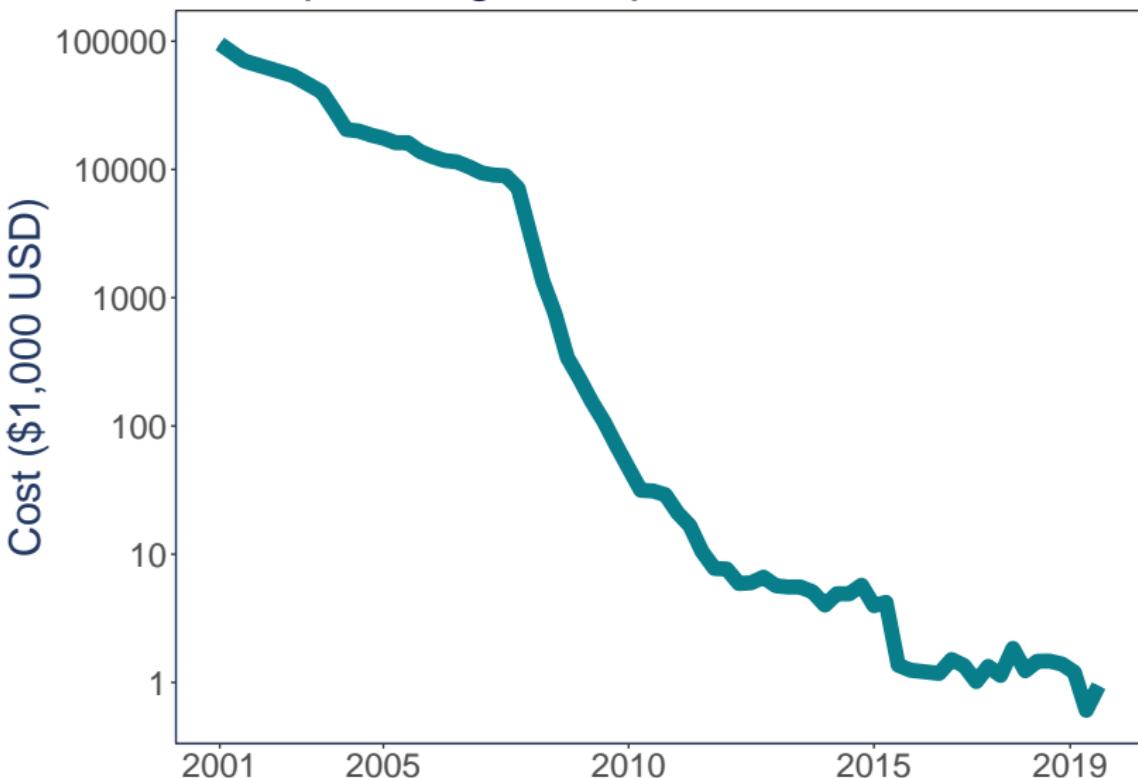
Bioinformatics: the hot interdisciplinary field

Daniella Lato
(She/Her)
PhD Candidate
Biology Department
Golding Lab, McMaster

✉ latodf@mcmaster.ca
GitHub: <https://github.com/dlato>

cagccagatggggggaggggtgagcgctctcccgctcaaa
acctccagcacttgcgatgcgttcgctcactgccgct
tcctaatctaaaaataaaaactgaattataaggtttat
ttcttagagtgcgtggtatcaaggattaaaatcaatcct
catcctatccagtcgcgtcaatctccggcaaagaggcgg
gagagattccggccgaattgcgcggtaagcgggaaaccc
ggtaaaacactgcagagtcaagcccttgccggcgtcgg
agtttgcggttcctgtggatgaggagtcgtatctgcgtgac
aatttgcgtggccatgagtcttgtccacatgcccggca
agagattccggtaggtgtcagatgtgaacaagtcgccc
ctttccactgcctgaagccaggcgccggccgttagctgt
tttgcgccttatcaacagaccgcggagaattgcgtg
gagaaccgcaaaaactacttccatctgcattgtatctccg
attcgaccggcgaaaactctgatgcgcggccgagcggc
tgctgcgaattccagtcctccatgcgtggaaacacgtc
tatccqctqatccqcaaccqqaagcqagctqatgcaggtca

Sequencing Cost per Human Genome

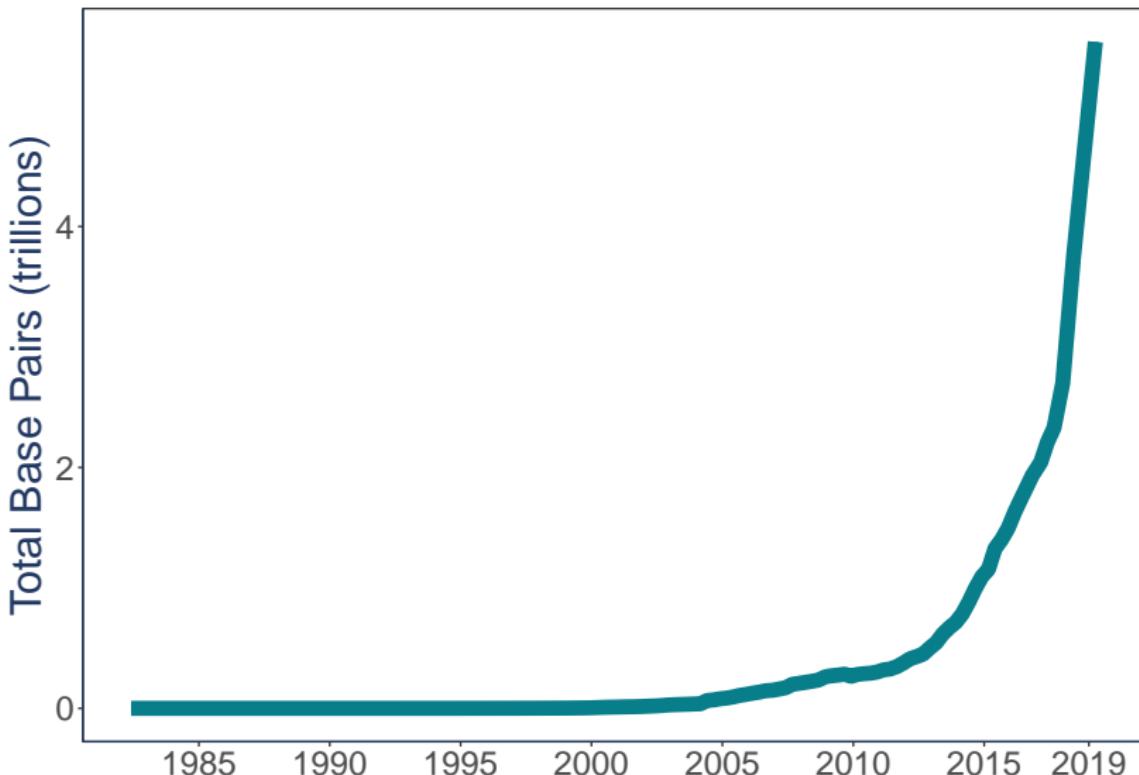


adapted from: Kris Wetterstrand, National Human Genome Research Institute

Cost to Sequence a Human Genome in 2019



EMBL Sequence Data



Data from: European Molecular Biology Laboratory (EMBL), publicly available

How many GB is that?

All that sequence data would equal about

11,066,077,343,692 GB

How many GB is that?



\approx **512GB**

The background consists of numerous laptops, all showing a blue screen with a white wavy pattern. The laptops are oriented in various directions, creating a dense, overlapping texture.

22 BILLION LAPTOPS

Bioinformatics to the rescue!



What is Bioinformatics?

Bioinformatics is taking **biological data, processes and theories** and **applying** “informatics” techniques (derived from disciplines like **math, computer science, and statistics**) to understand, organize, and predict biological processes.

Broad Types of Bioinformatics

1. Data Analysis
2. Software Development
3. Modeling

Broad Types of Bioinformatics

1. Data Analysis
2. Software Development
3. Modeling

1. Data Analysis

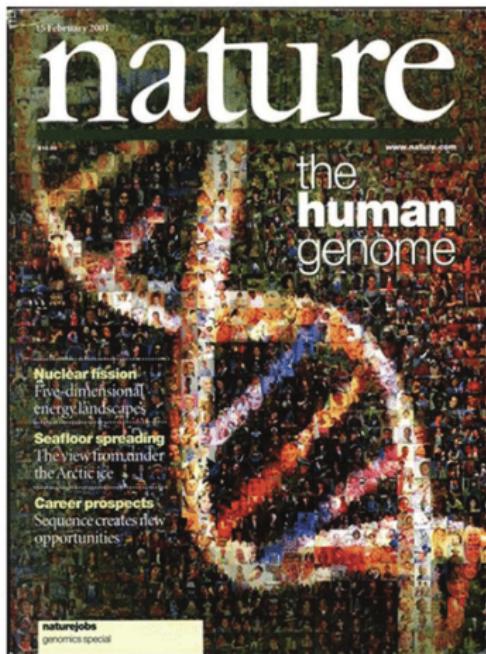
**Generating, interpreting,
and explaining any
biological data**

1. Data Analysis

The Human Genome is Sequenced!

1. Data Analysis

The Human Genome is Sequenced!



Broad Types of Bioinformatics

1. Data Analysis
2. Software Development
3. Modeling

2. Software Development

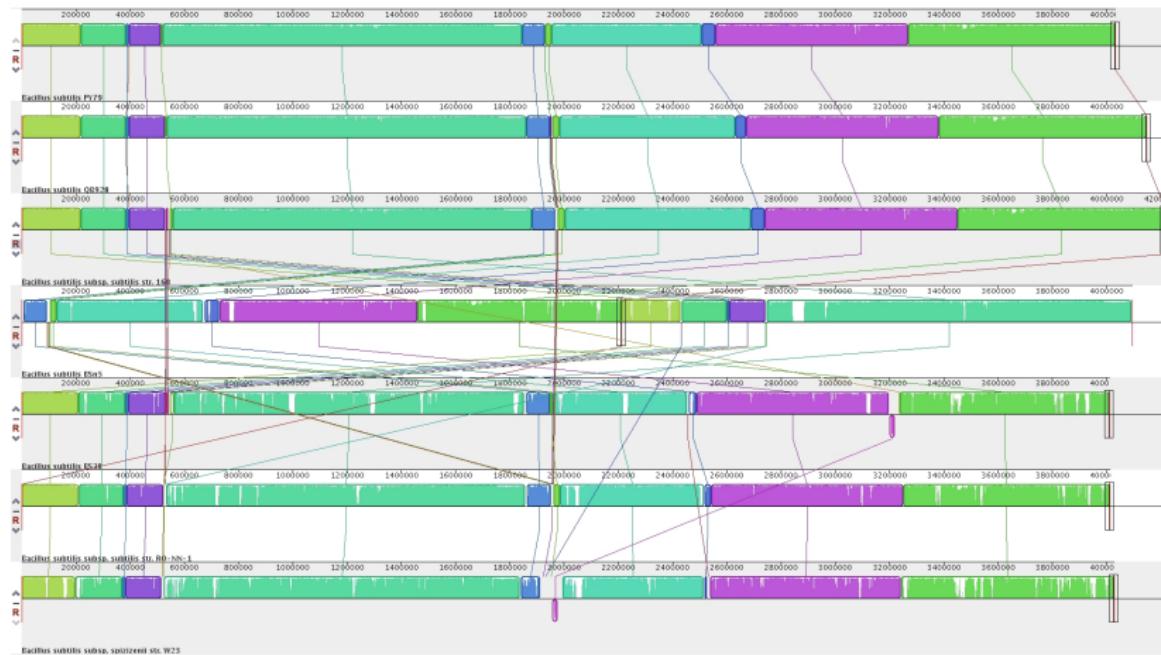
**Creating bioinformatics
tools for other people to
use**

2. Software Development

`progressiveMauve`: Whole genome alignment program

2. Software Development

progressiveMauve: Whole genome alignment program



2. Software Development

progressiveMauve: Whole genome alignment program

```
info15 Mauve% /home/dlato/Mauve_snapshot/mauve_snapshot_2012-06-07/Linux-x64/progressiveMauve --help
/home/dlato/Mauve_snapshot/mauve_snapshot_2012-06-07/linux-x64/progressiveMauve: unrecognized option '--help'
progressiveMauve Usage:

When each genome resides in a separate file:
/home/dlato/Mauve_snapshot/mauve_snapshot_2012-06-07/Linux-x64/progressiveMauve [options] <seq1 filename> ... <seqN filename>

When all genomes are in a single file:
/home/dlato/Mauve_snapshot/mauve_snapshot_2012-06-07/linux-x64/progressiveMauve [options] <seq filename>

Options:
--island-gap-size=<number> Alignment gaps above this size in nucleotides are considered to be islands [20]
--profile=files[...] (Not yet implemented) Read an existing sequence alignment in XMFA format and align it to other sequences or alignments
--apply-backbone[...] Read an existing sequence alignment in XMFA format and apply backbone statistics to it
--disable-backbone Disable backbone detection
--numS Find MUMs only, do not attempt to determine locally collinear blocks (LCBs)
--seed-weight=<number> Use the specified seed weight for calculating initial anchors
--output=<file> Output file name. Prints to screen by default
--backbone-output=<file> Backbone output file name (optional)
--match-input=<file> Use specified match file instead of searching for matches
--input-id-matrix=<file> An identity matrix describing similarity among all pairs of input sequences/alignments
--max-pairs-align=<number> Maximum number of base pairs to attempt aligning with the gapped aligner
--input-guide-tree=<file> A phylogenetic guide tree in NEWICK format that describes the order in which sequences will be aligned
--output-guide-tree=<file> Write out the guide tree used for alignment to a file
--version Display software version information
--debug Run in debug mode (perform internal consistency checks--very slow)
--scratch-path=1=<path> Designate a path that can be used for temporary data storage. Two or more paths should be specified.
--scratch-path=2=<path> Designate a path that can be used for temporary data storage. Two or more paths should be specified.
--extant=sp|sp+|sp-|spc Selects the anchoring score function. Default is extant sum-of-pairs (sp).
--no-weight-scaling Don't scale LCB weights by conservation distance and breakpoint distance
--max-breakpoint-distance-scale=<number> [0,1]> Set the maximum weight scaling by breakpoint distance. Defaults to 0.5
--conservation-distance-scale=<number> [0,1]> Scale conservation distances by this amount. Defaults to 0.5
--muscle-args=<arguments in quotes> Additional command-line options for MUSCLE. Any quotes should be escaped with a backslash
--skip-refinement Do not perform iterative refinement
--skip-anchoring Do not perform gap opening
--hp-dist-estimate=min|scm|numbers Minimum LCD score for estimating pairwise breakpoint distance
--mem-clean Set this to true when debugging memory allocations
--gap-open=<number> Gap open penalty
--repeat-penalty=<negative|zero> Sets whether the repeat scores go negative or go to zero for highly repetitive sequences. Default is negative.
--gap-exclude=<number> Gap extend penalty
--substitution-matrix=<file> Nucleotide substitution matrix in NCBI format
--weight-penalty=<number> Minimize pairwise LCB score
--hmm-scaled-penalty=<number> Minimize breakpoint penalty after scaling the penalty by expected divergence
--hmm-p-go-homologous=<number> Probability of transitioning from the unrelated to the homologous state [0.00001]
--hmm-p-go-unrelated=<number> Probability of transitioning from the homologous to the unrelated state [0.000000001]
--hmm-identity=<number> Expected level of sequence identity among pairs of sequences, ranging between 0 and 1 [0.7]
--seed-family=Use a family of spaced seeds to improve sensitivity
--solid-seeds=Use solid seeds. Do not permit substitutions in anchor matches.
--coding-seeds=Use coding pattern seeds. Useful to generate matches coding regions with 3rd codon position degeneracy.
--disable-cache=Disable recursive anchor search caching to workaround a crash bug
--no-recursion=Disable recursive anchor search

Examples:
/home/dlato/Mauve_snapshot/mauve_snapshot_2012-06-07/Linux-x64/progressiveMauve --output=my_seqs.xmfa my_genome1.gbk my_genome2.gbk my_genome3.fasta

If genomes are in a single file and have no rearrangement:
/home/dlato/Mauve_snapshot/mauve_snapshot_2012-06-07/linux-x64/progressiveMauve --collinear --output=my_seqs.xmfa my_genomes.fasta
info15 Mauve%
```

Broad Types of Bioinformatics

1. Data Analysis
2. Software Development
3. Modeling

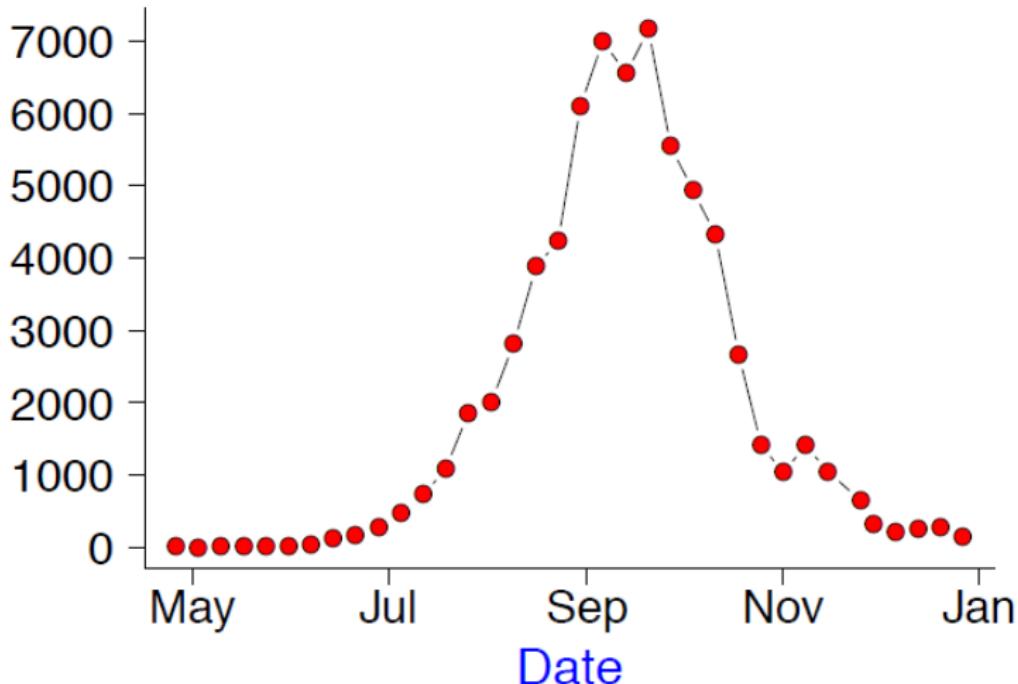
3. Modelling

Using mathematical and statistical principals to represent and predict biological systems or data

3. Modelling

SIR model and the Great Plague of London

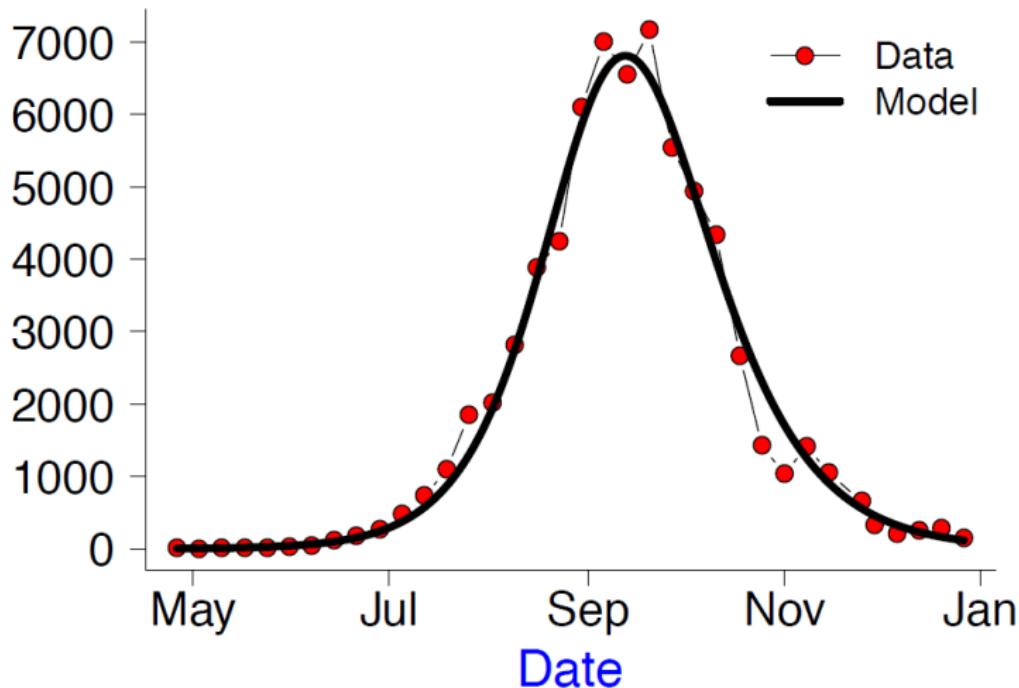
Weekly Deaths from Plague



3. Modelling

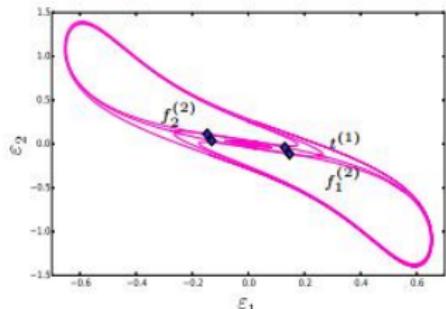
SIR model and the Great Plague of London

Weekly Deaths from Plague

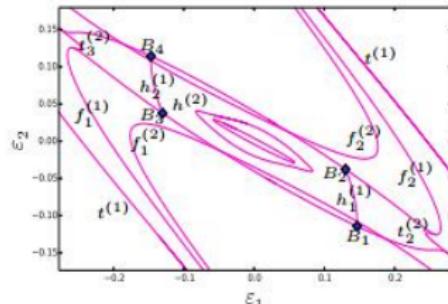


3. Modelling

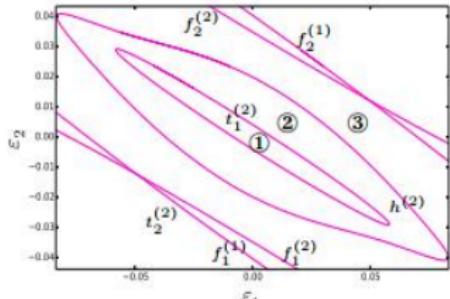
Bifurcation theory and predator-prey relationships



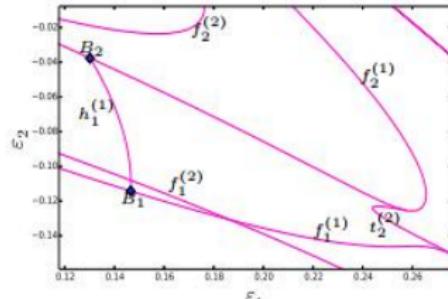
(A)



(B)



(C)



(D)

Broad Types of Bioinformatics

1. Data Analysis
2. Software Development
3. Modeling

Broad Types of Bioinformatics

1. Data Analysis
2. Software Development
3. Modeling
4. Combination!

4. Combination

- Modelling + Software Development
- Data Analysis + Modelling
- Data Analysis + Software Development
- Data Analysis + Modelling + Software Development

Spatial Patterns of Molecular Trends in Bacterial Genomes

(How do molecular trends change with position in the genome?)

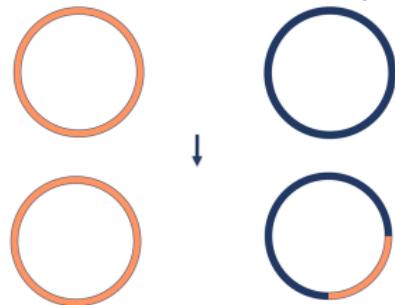
Bacteria are bizarre!

should I replace this slide with just one crazy pic of HGT?? should I remove this all together? or just shorten it?

Bacteria are bizarre!

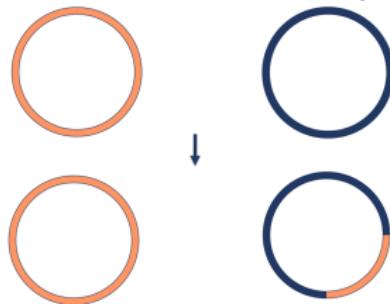
should I replace this slide with just one crazy pic of HGT?? should I remove this all together? or just shorten it?

Horizontal Gene Transfer (HGT)

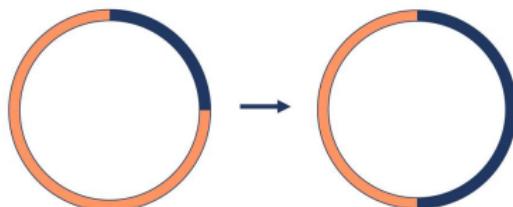


Bacteria are bizarre!

Horizontal Gene Transfer (HGT)

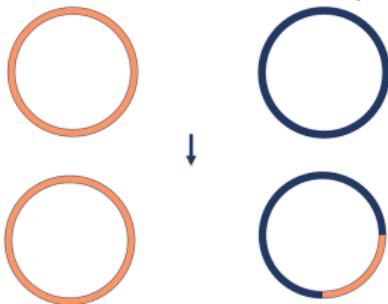


Duplication

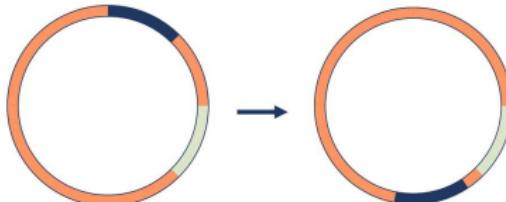


Bacteria are bizarre!

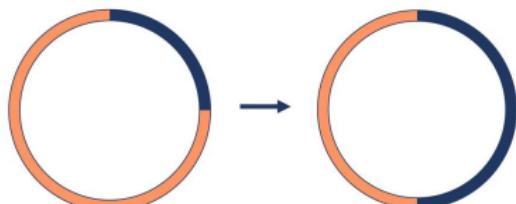
Horizontal Gene Transfer (HGT)



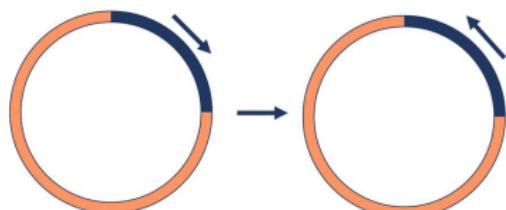
Rearrangement and Translocation



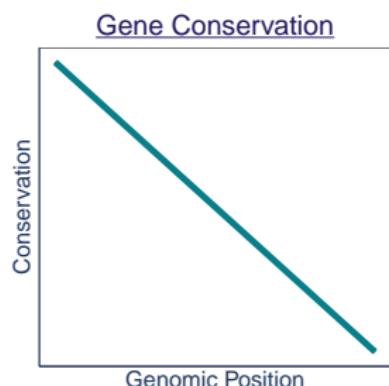
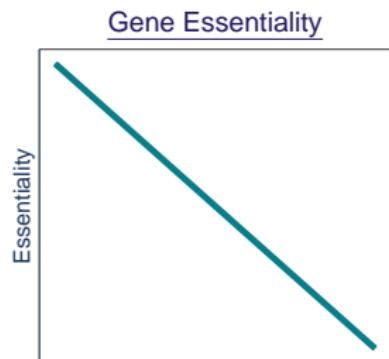
Duplication



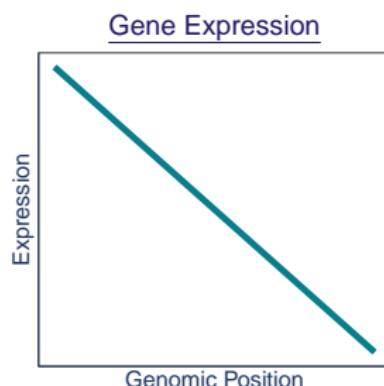
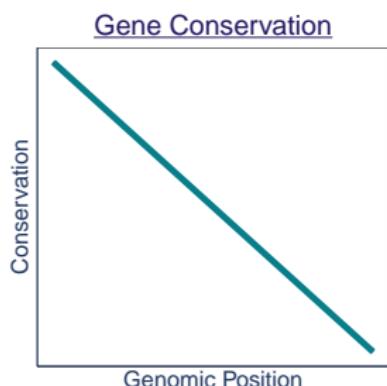
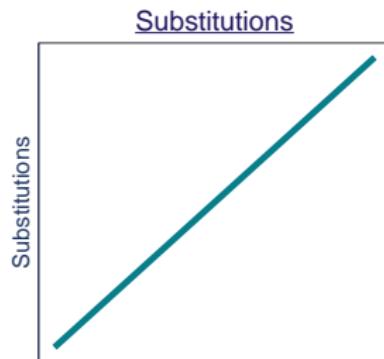
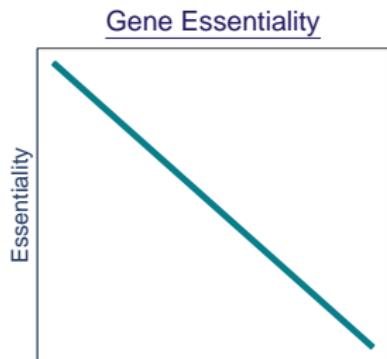
Inversion



My Research: Spatial molecular trends

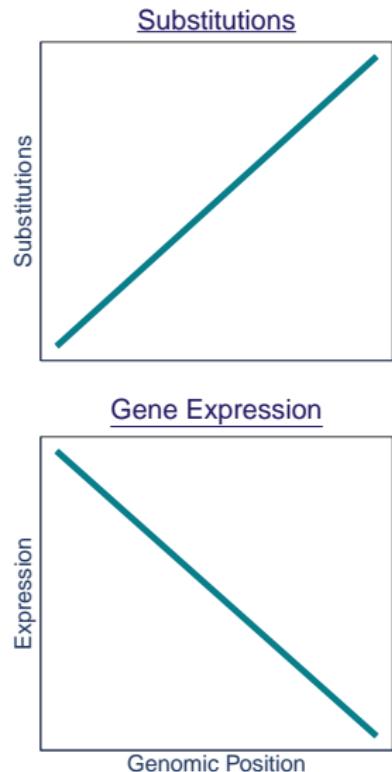


My Research: Spatial molecular trends



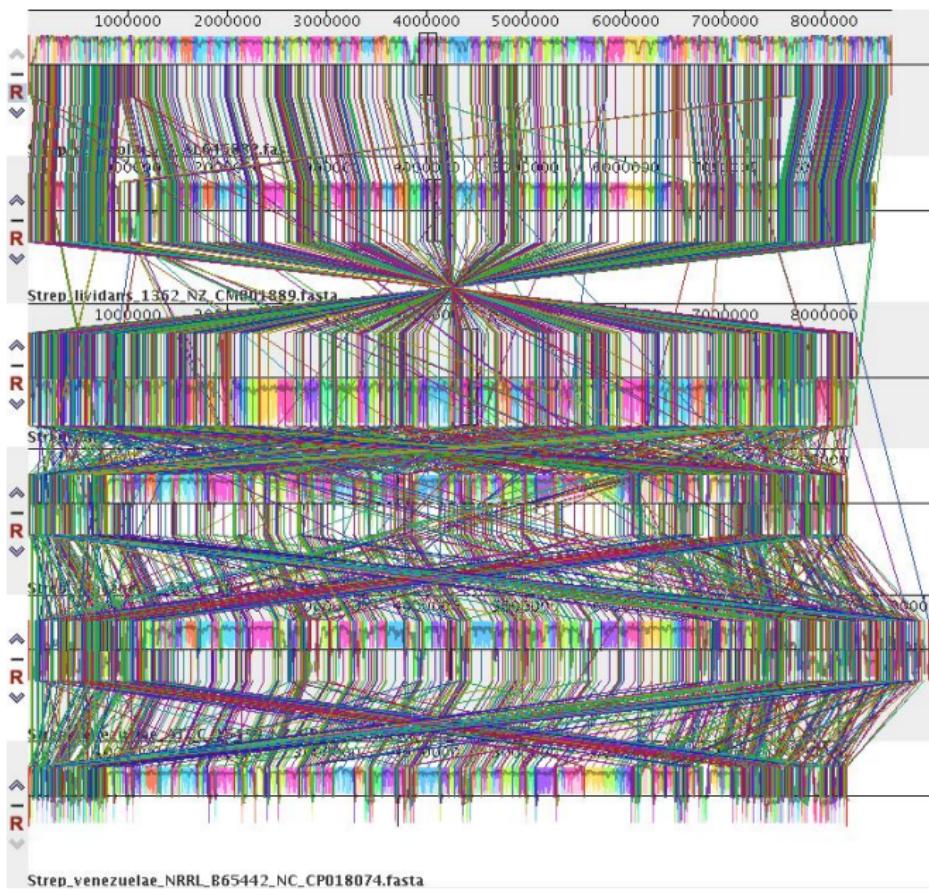
Couturier et al. 2006, Cooper et al. 2010, Sharp et al. 2005, Morrow et al. 2012, Cooper and Rocha 2006

My Research: Spatial molecular trends



Couturier et al. 2006, Cooper et al. 2010, Sharp et al. 2005, Morrow et al. 2012, Cooper and Rocha 2006

My Research: Incorporating Bacteria Genome Shuffling!



My Research: The Organisms

Bacteria:

- *Escherichia coli*
- *Bacillus subtilis*
- *Streptomyces*
- *Sinorhizobium meliloti*

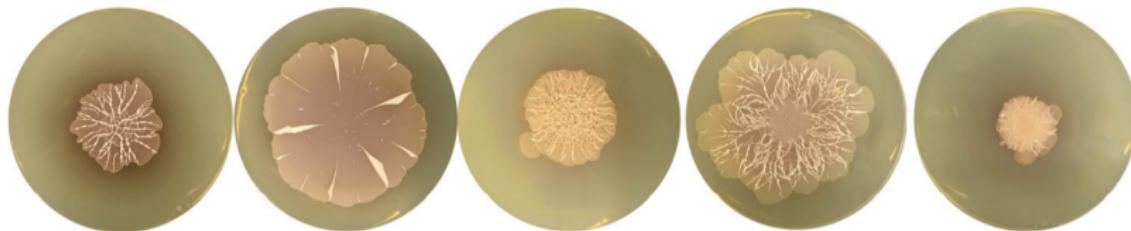
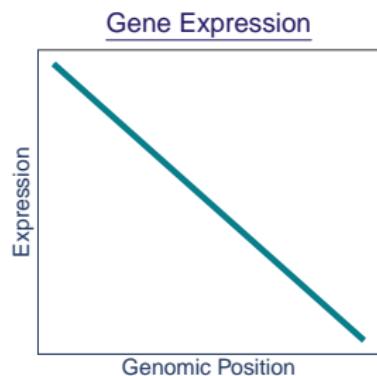
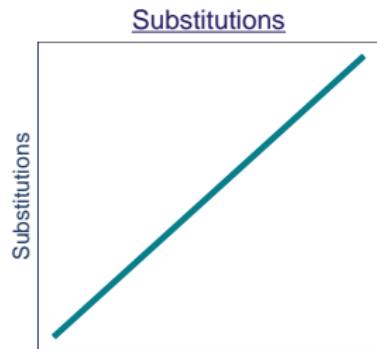


Photo: *Streptomyces* by Stephanie Jones, Marie Elliot's Lab at McMaster University

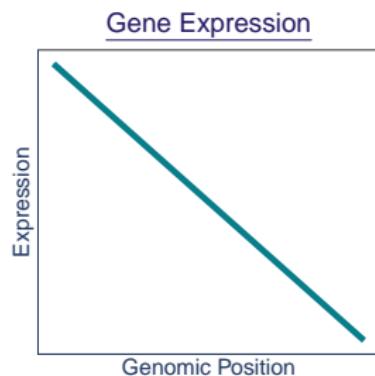
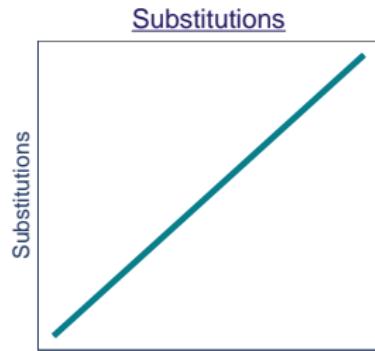
My Research: Conclusions

Previous Studies:

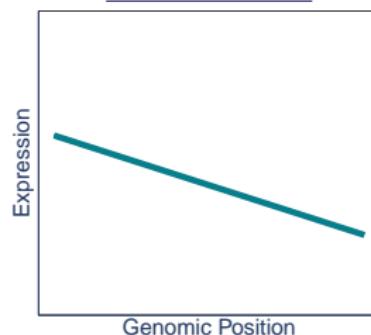
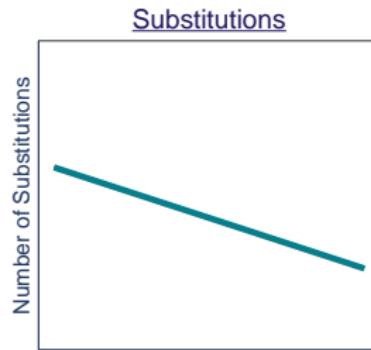


My Research: Conclusions

Previous Studies:



My Research:

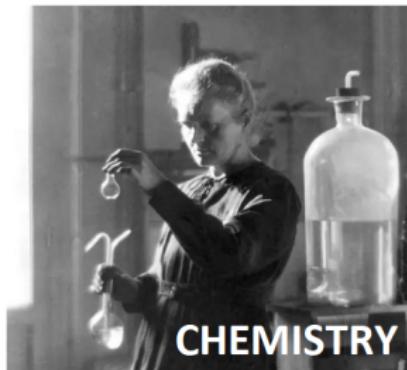


Why become a Comp Bio Geek?



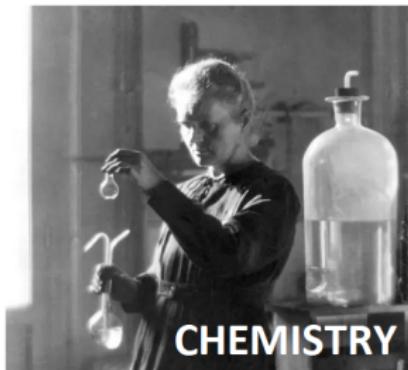
Not Just Biology!

Not Just Biology!



Marie Curie, Nobel Prize in Chemistry

Not Just Biology!

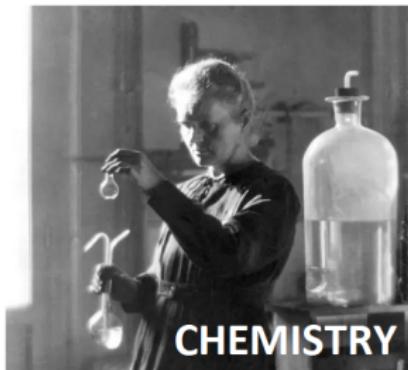


Marie Curie, Nobel Prize in Chemistry



Katherine Johnson, NASA Physicist

Not Just Biology!



Marie Curie, Nobel Prize in Chemistry



Katherine Johnson, NASA Physicist



Courses you should take or audit:

- **Online Resources!**
 - DataCamp, Coursera, Codeacademy
- **Bio 3S03: Intro to Bioinformatics**
- **Bio 3SS3: Population Ecology**
- **Bio 3SA3: Applied Statistics for Biology**
- **Math 4MB3: Mathematical Biology**
- **Math 3MB3: Introduction to Modelling**

Questions?

latodf@mcmaster.ca

