



Bioinformatics and its applications

Alla L Lapidus, Ph.D.

SPbAU, SPbSU,

St. Petersburg





Term Bioinformatics

Term **Bioinformatics** was invented by Paulien Hogeweg (Полина Хогевег) and Ben Hesper in 1970 as "the study of informatic processes in biotic systems".

Paulien Hogeweg is a Dutch theoretical biologist and complex systems researcher studying biological systems as dynamic information processing systems at many interconnected levels.

Definitions of what is Bioinformatics:

Bioinformatics is the use of IT in biotechnology for the data storage, data warehousing and analyzing the DNA sequences. In Bioinformatics knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and of course sound knowledge of IT to analyze biotech data.

Bioinformatics is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful

The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.

Bioinformatics develops computer to analyze biology for example ingredients and metabolism.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

<http://www.bisti.nih.gov/CompuBioDef.pdf>

My additions:

- 1. Bioinformatics is a SCIENCE**
- 2. Not only to develop algorithms, store, retrieve, organize and analyze biological data but to CURATE data**

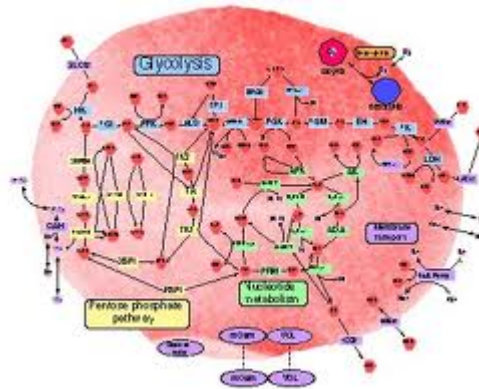
Bioinformatics is being used in following fields:

- Microbial genome applications
- Molecular medicine
- Personalised medicine
- Preventative medicine
- Gene therapy
- Drug development
- Antibiotic resistance
- Evolutionary studies
- Waste cleanup
- Biotechnology
- Climate change Studies
- Alternative energy sources
- Crop improvement
- Forensic analysis
- Bio-weapon creation
- Insect resistance
- Improve nutritional quality
- Development of Drought resistant varieties
- Veterinary Science

Sequencing projects



Data analysis



Results Interpretation



Data applications

LIMS - Lab Information Management Software

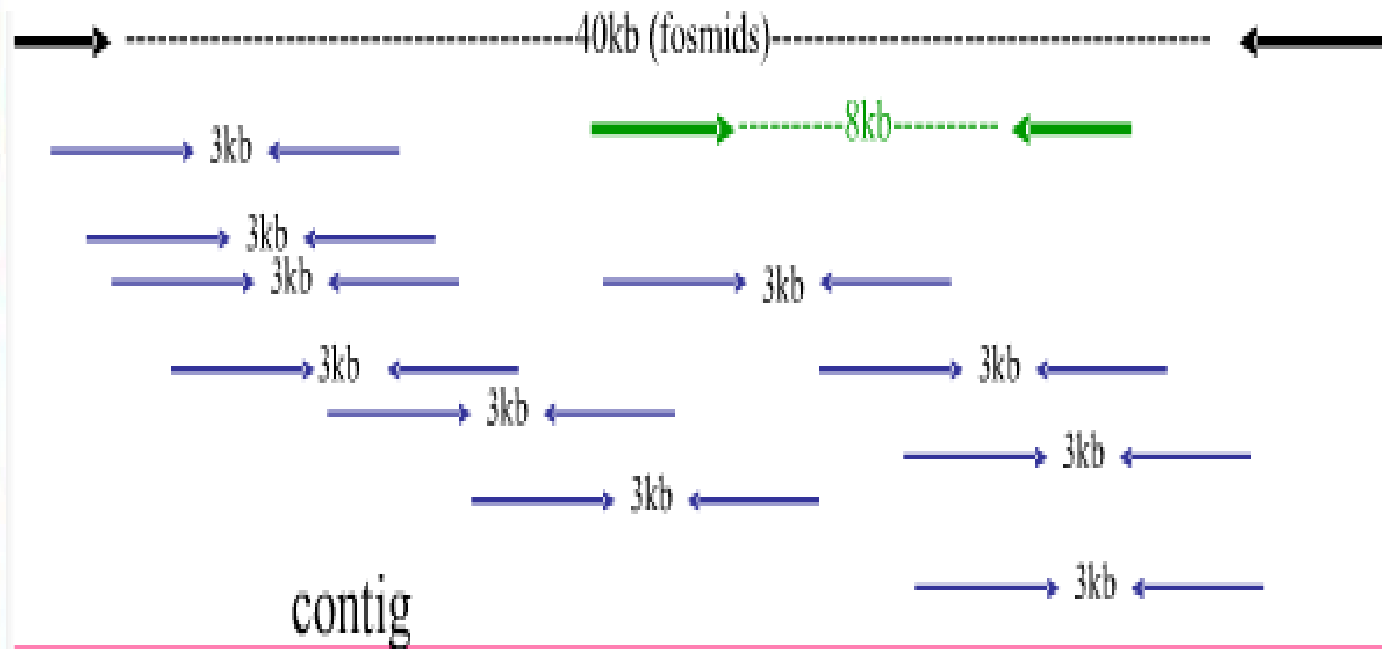
Microbial genome applications

- Genome assembly
- Re-sequencing
- Comparative analysis
- Evolutionary studies
- Antibiotic resistance
- Waste cleanup
- Biotechnology

Genome Assembly

- Genome assembly is a very complex computational problem due to enormous amount of data to put together and some other reasons reasons.
- Ideally an assembly program should produce one contig for every chromosome of the genome being sequenced. But because of the complex nature of the genomes, the ideal conditions just never possible, thus leading to gaps in the genome.

De Novo assembly - puzzle without the picture



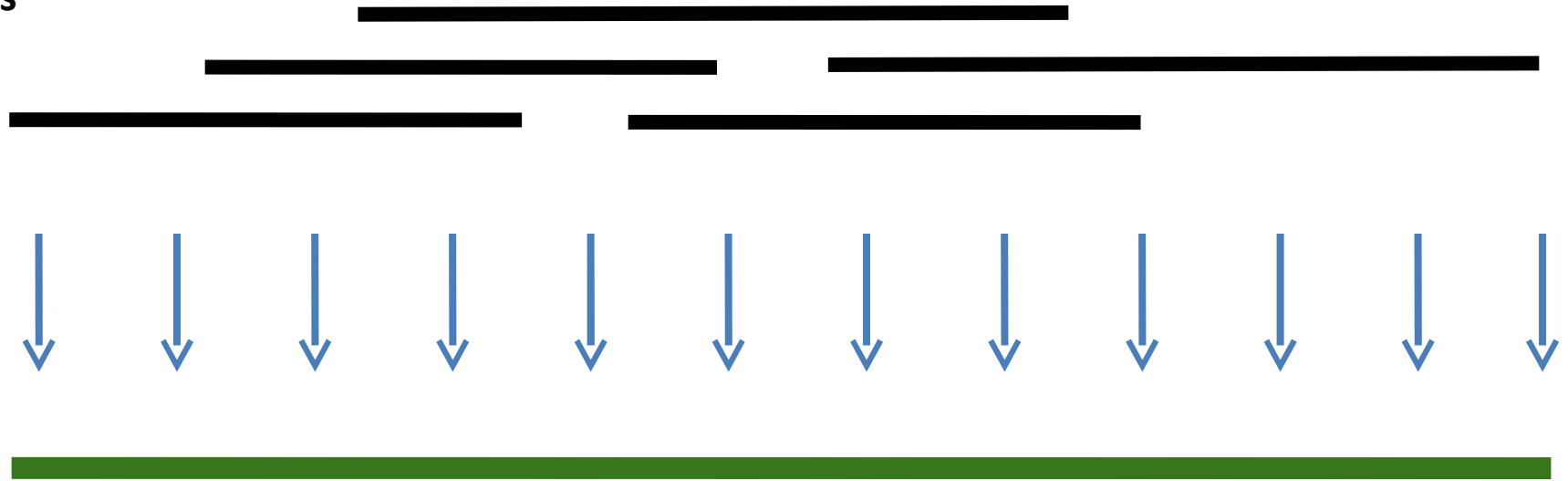
Assembly Challenges

- Presence of **repeats**. Repeats are identical sequences that occur in the genome in different locations and are often seen in varying lengths and in the multiple copies. There are several types of repeats: tandem repeats or interspersed repeats. The read's originating from different copies of the repeat appear identical to the assembler, causing errors in the assembly.
- **Contaminants** in samples (eg. from Bacteria or Human).
- **PCR artefacts** (eg. **Chimeras** and **Mutations**)
- **Sequencing errors**, such as “**Homopolymer**” errors – when eg. 2+ run of same base.
- **MID**'s (multiplex indexes), **primers/adapters** still in the raw reads.
- **polyploid** genomes

Assembly algorithms

Overlap-Layout-Consensus - Find overlaps between all reads

reads



Consensus

Problems caused by new sequencing technologies:

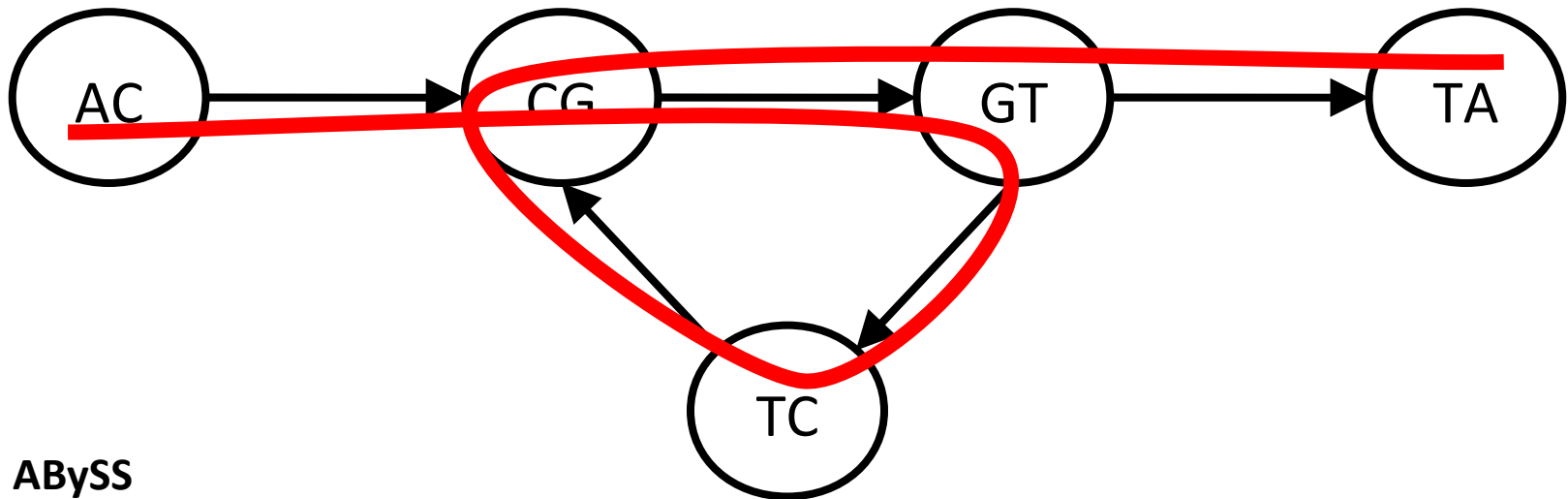
- ❖ Hard to find overlaps between short reads
- ❖ Impossible to scale up



De Bruijn graph

ACGTCGTA

k=2

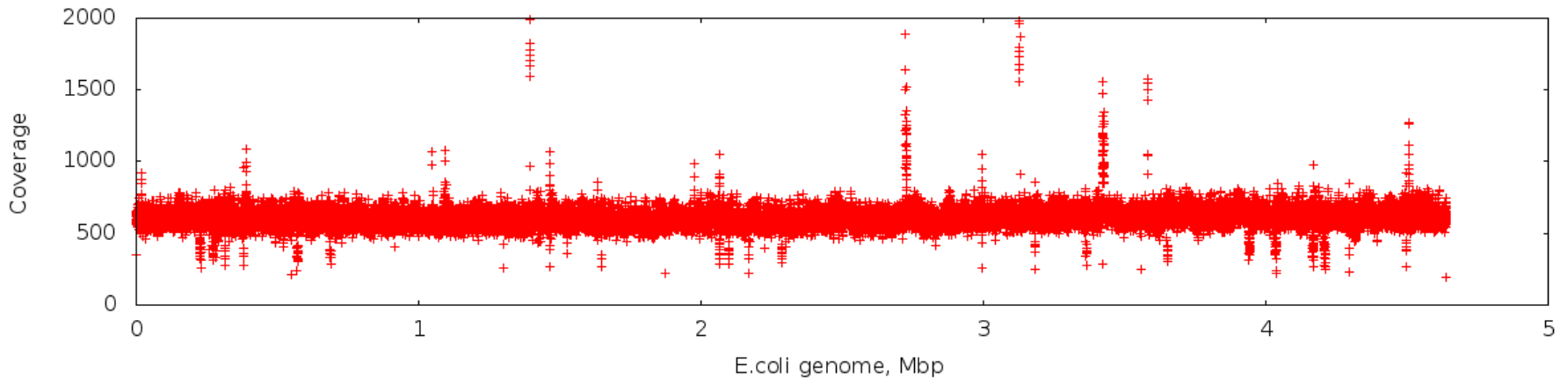


- ▲ ABySS
- ▲ ALLPATHS-LG
- ▲ EULER
- ▲ IDBA
- ▲ Velvet

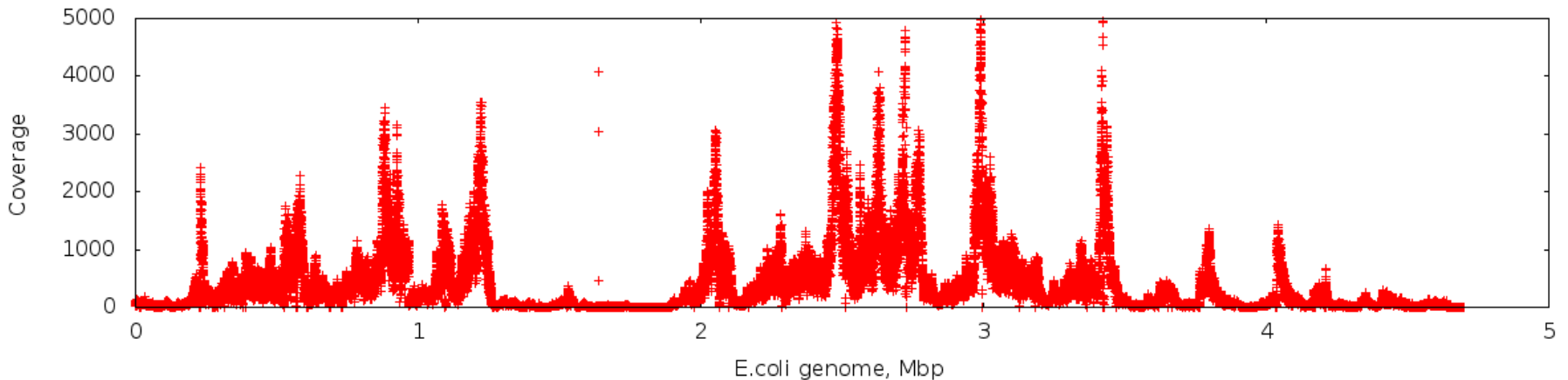
Single-cell dataset

- *E. coli* isolate dataset

- IDBA-UD
- SPAdes
- Velvet

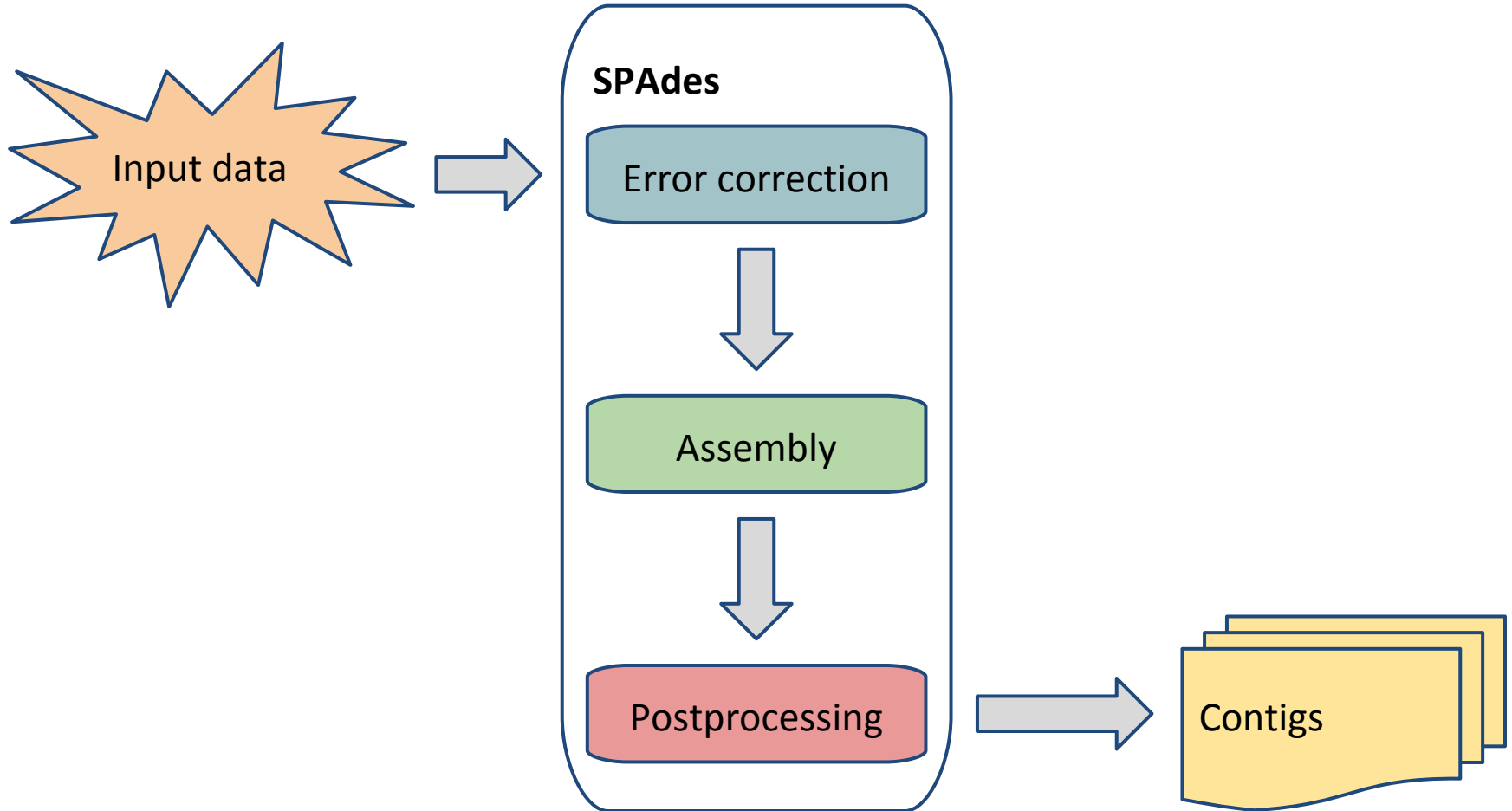


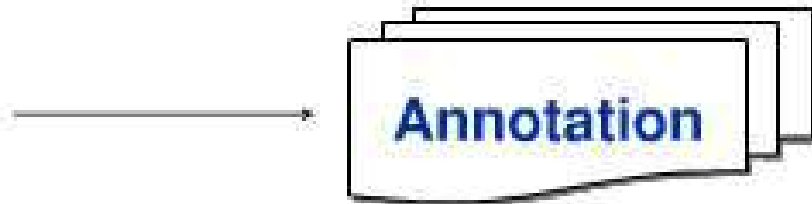
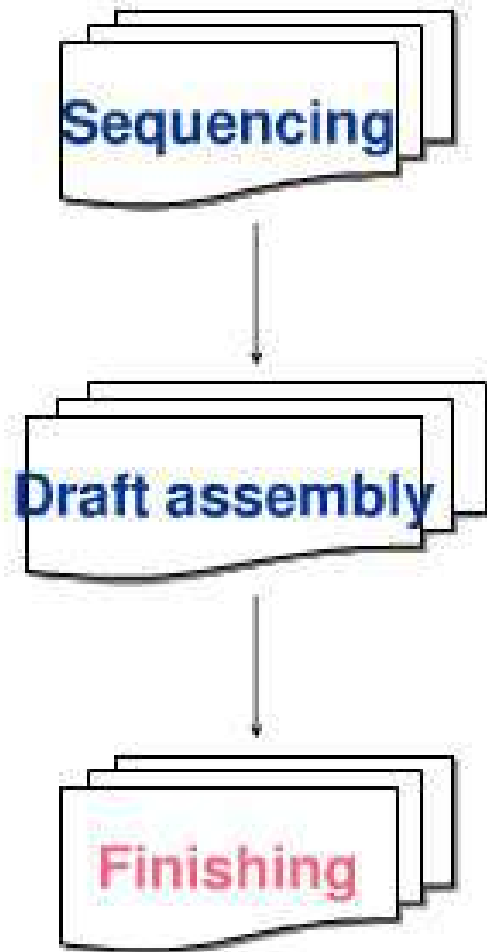
- *E. coli* single-cell dataset





SPAdes pipeline





Gene Prediction and Genome Annotation

```
1 aacaggggtgt atctcgcaca ttctcatcca ctagtataac tgctgctgac agtaatcgaa
61 ctagatagac tgttctggat gctatcattc gatatnttga caacacggga gccatcctgt
121 tcgttgatcc gagattcgac gagtcatgca acaagatcca gaccgttgcc tgcaaacgcc
181 taggctgtga atgaacgact cgatcacgat cgctagtcgc acgtctgac tcaccgattg
241 aagccgtatt ccacagagtg cgagaaccgg tcatttactg agtggttcgg ctctgtttaa
301 atacggaaag cccactcggg agagatatct ctcccttaatg ggctatgaaa ggtatgaatg
361 gtggcggcga accgcgttcc ccagaggctc ggcgactcca gtactccccg gaacgctggg
421 gggcttatct tccgtgttcg ggatgggtac gggaggcaac cccaccgctg tggccgcta
481 acgtcagatc acggaatcga accgcgatag taccagtctc gattaactct tccaccggt
541 gattacgtgc gatccagttt ggcctgggac tcgttcagcg acgagttaa tcgatggtga
601 atgagtcaca gtgcgtatga atgatggctt tggctctgta gtgctcgtgg gcttaaccctc
661 tegttaacct gacgcgcaca ccccgagtct atcgaccgcg tcttgtagcg gggacctcgg
721 cgggtgtctct tttccaagtg ggtttcagc ttagatgcgt tcagctotta ccccggtgg
781 cgtggctacc cggcacgtgc tctctcgaac aaccggtaca ccagtggcca ccaaccgtag
841 ttccctctgt actatacggg cgttctgtgc agacaccatt acacaccag tagatagcag
901 cggacctgtc tcacgacggg ctaaaccag ctacgacat cctttaatag gogaacaacc
961 tcacccttgc cgccttctgc acgggcagga tggagggaac cgacatcgag gttagcaagcc
1021 actcggctga tatgtgctct tgcgagtgc gctctgta tccctagggt agctttctg
1081 tcatcaattg cccgcatoaa gcaggcta atggttcgtg gaccacgct tcgctgtagc
1141 gttcctcgtt gggaagaaca ctgtcaagct taatcttctt cttgcactct tcgcccggtc
1201 tctgtcccgg ctgagatagc catagggcgc gctcagatc ttttcgagcg cgtaccgccc
1261 cagtcaaact gcccggtat cgggtgcctc ctcccgaggt gagagtcgca gtcaccgacg
1321 ggtagtattt cactggtgac tcgggtggccc gctagcggcg gtacctgtgt agtgtctcct
1381 atgtatgctg cacatcggcg accacgtctc agcgcagacc tgcagtaaag ctccataggg
1441 tcttcgcttc cccctgggtg tctccagact ccgcactgga atgtacagt caccgggccc
1501 aacgttggga cagtgaagct ctggttaatc cactcatgca agccgctact gatcgggaa
1561 ggtactacgc taccttaaga gggctatagt tacccccgcc gttgacaggt ccttcgtcct
1621 cttgtacgag tggttcagat acctgcactg ggcaggatc agtgaccgta cgagtccttg
1681 cggatttgcg gtcacctatg ttgttactag acagtccgag ctcccgagtc actgcgacct
1741 gctccgcttc ggagcaggca tccctcttc cgaaggtacg ggactaact gccgaattcc
1801 ctaacgttgg ttgctcccga caggccttgg ctttcgcgc catggacacc tgtgtcgtt
```

Based on similarity to known genes – blastX (NCBI)

Gene finding programs

- **Glimmer** – for most prokaryotic genomes
- **GenMark** – for both prokaryotic genomes and eukaryotic genomes

IMG Content

Datasets

Bacteria	8120
Archaea	248
Eukarya	183
Plasmids	1193
Viruses	2809
Genome Fragments	579
Total Datasets	11132
GEBA	245

Last updated: [2013-07-05](#)

[IMG 4.0 is dedicated to the memory of our colleague, Iain Anderson](#)

- [Genome by Metadata](#)
- [Project Map](#)
- [Content History](#)
- [System Requirements](#)
- [About IMG](#)
- [FAQ](#)



Hands on training available at the

[Microbial Genomics & Metagenomics Workshop](#)

The Integrated Microbial Genomes (IMG) system ([Nucleic Acids Research, Vol 40, 2012](#)) serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in a uniquely integrated context. Plasmids that are not part of a specific microbial genome sequencing project and phage genomes are also included into IMG in order to increase its genomic context for comparative analysis.

For details, see [IMG Release Notes](#) (Dec. 12, 2012), in particular the workspace and background computation capabilities available to IMG registered users.



Count	Total
DNA, number of bases	60,476,662,654
Total Genes	25,395,838
Total Genomes	11,132

[IMG Statistics](#)

[IMG ER Account Request](#)

All Genomes

Genome Count

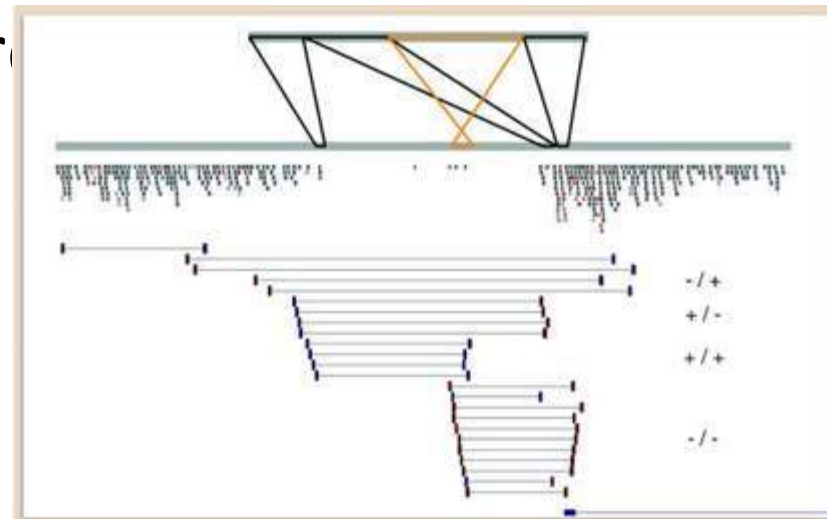
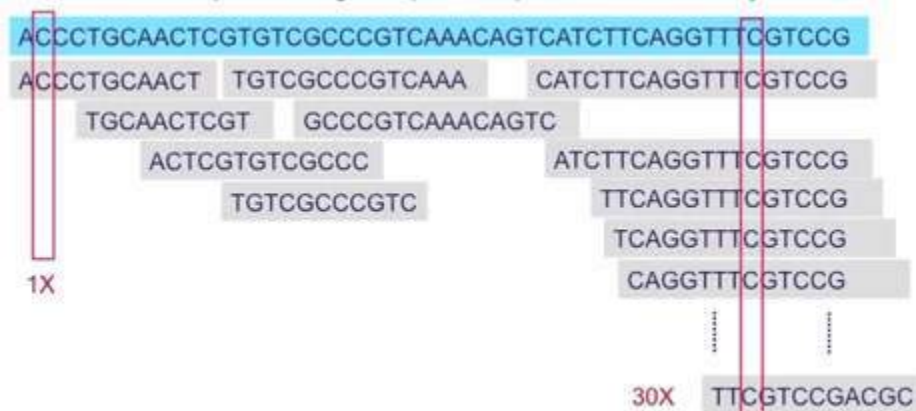
Status	Bacteria	Archaea	Eukaryota	Plasmids	Viruses	Genome Fragments	Total
Finished	2131	154	37	1190	2809	579	6900
Draft	2407	28	146	3	0	0	2584
Permanent Draft	1582	66	0	0	0	0	1648
Total	6120	248	183	1193	2809	579	11132

Re-sequencing

Projects aimed at characterizing the genetic variations of species or populations

Resequencing of bacterial and archaeal isolates etc is possible if reference genomes are available

This approach can help to better understand bacterial community structure, gene function in bacteria under selective pressure



Climate change Studies

Increasing levels of carbon dioxide emission are thought to contribute to global climate change.

One way to decrease atmospheric carbon dioxide is to study the genomes of microbes that use carbone dioxidet as their sole carbon source


Human microbiome

MetaHIT - Europe

Human Microbiome Project –US

The human microbiome includes viruses, fungi and bacteria, their genes and their environmental interactions, and is known to influence human physiology.

There's very broad variation in these bacteria in different people and that severely limits our ability to create a “normal” microflora profile for comparison among healthy people and those with any kind of health issues.



The diagram shows a human silhouette with red dots indicating the locations of the microbiome. The background is dark with large, glowing blue and green circular shapes representing microbes. Labels on the left side of the silhouette are connected to red dots on the body by thin white lines.

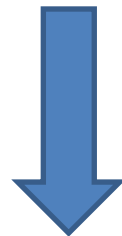
Nasal
Oral
Skin
Gastrointestinal
Urogenital

Children with **autism** harbor significantly fewer types of gut bacteria than those who are not affected by the disorder, researchers have found.

Prevotella species were most dramatically reduced among samples from autistic children—especially *P. copri*. (helps the breakdown of protein and carbohydrate foods)

Bioinformatics combining biology with computer science

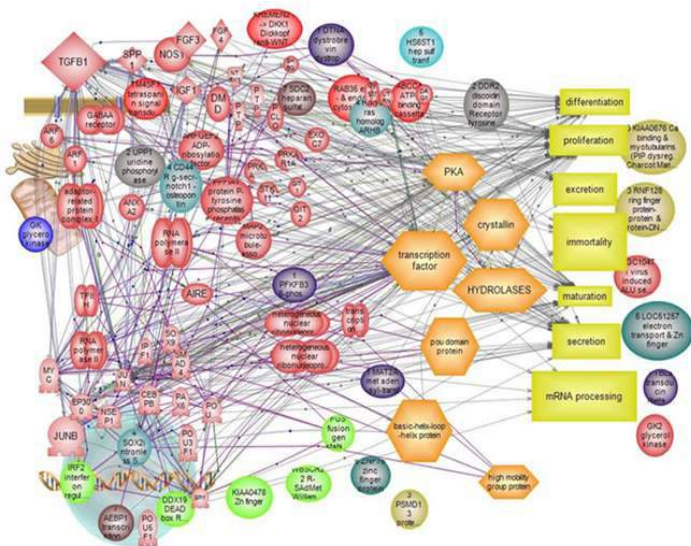
- it can explore the causes of diseases at the molecular level
- explain the phenomena of the diseases on the gene/pathway level
- make use of computer techniques (data mining, machine learning etc), to analyze and interpret data faster
- to enhance the accuracy of the results



Reduce the cost and time of drug discovery

To improve drug discovery we need to discover
(read "develop") efficient bioinformatics
algorithms and approaches for

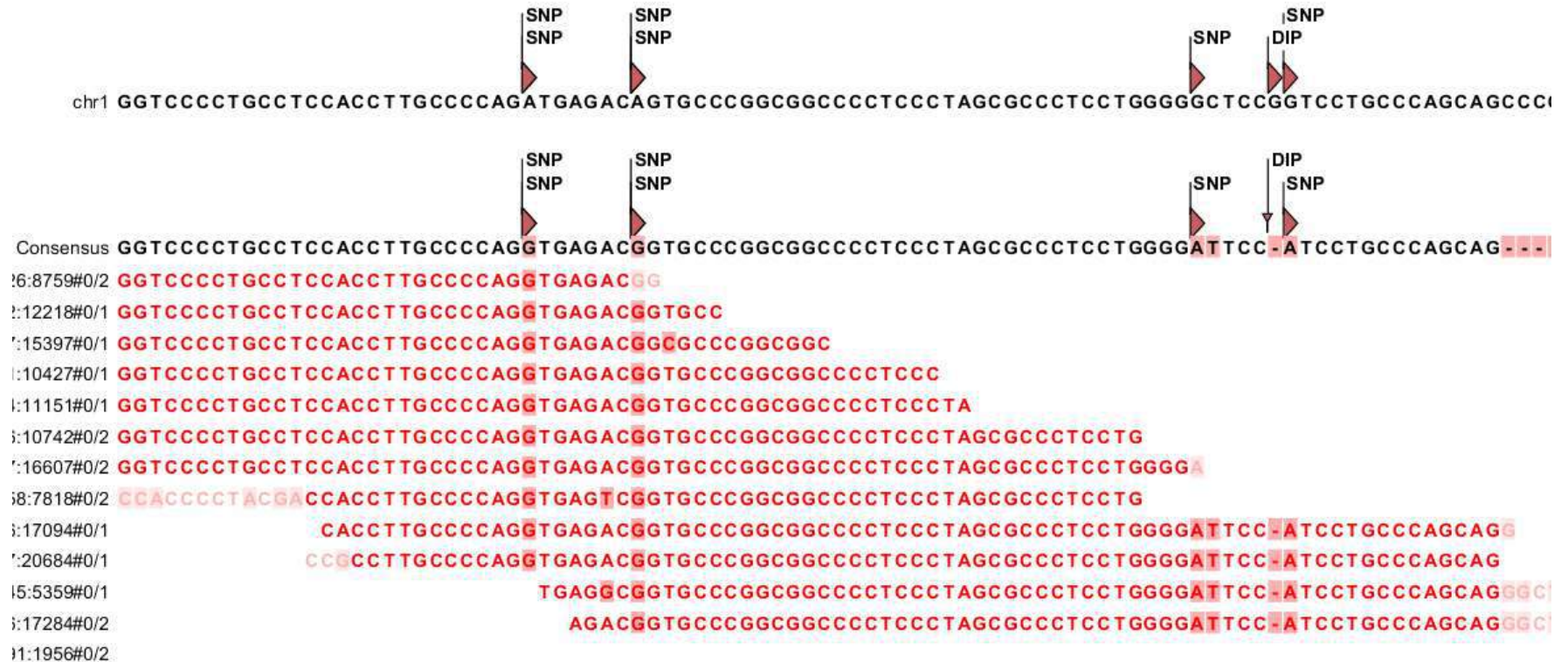
target identification
target validation
lead identification
lead optimization



Advantages of detecting mutations with next-generation sequencing

- High throughput
 - Test many genes at once
- Systematic, unbiased mutation detection
 - All mutation types
 - Single nucleotide variants (SNV), copy number variation (CNV)-insertions, deletions and translocations
- Digital readout of mutation frequency
 - Easier to detect and quantify mutations in a heterogeneous sample
- Cost effective **precision** medicine
 - “Right drug at right dose to the right patient at the right time”

Homozygous SNPs and indel



Missed SNP?

chrGCAGAGGCCAAGCCAGAGGTTCCAGGCTTAAACCCAGCCCTGCCCTGCCAGTCCA

ConsensusGCAGAGGCCAAGCCAGAGGTTCCAGGCTTAAACCCAGCCCTGCCCTGCCAGTCCA
:002.4009#0/

3866:4795#0/GCA

1407:2153#0/

1308:3912#0/GCAGAGGCCAAGCCAGAGGTTCCAGGCTTAAA

3914:7870#0/GCCAAGCCAGAGGTTCCAGGCTTAAACCCAGCCCTGCCCTGCCAGTCCA

5555:2114#0/CAAGCCAGAGGTTCCAGGCTTAAACCCAGCCCTGCCCTGCCAGTCCA

579:16341#0/CCAGGCTTAAACCCAGCCCTGCCCTGCCAGTCCA

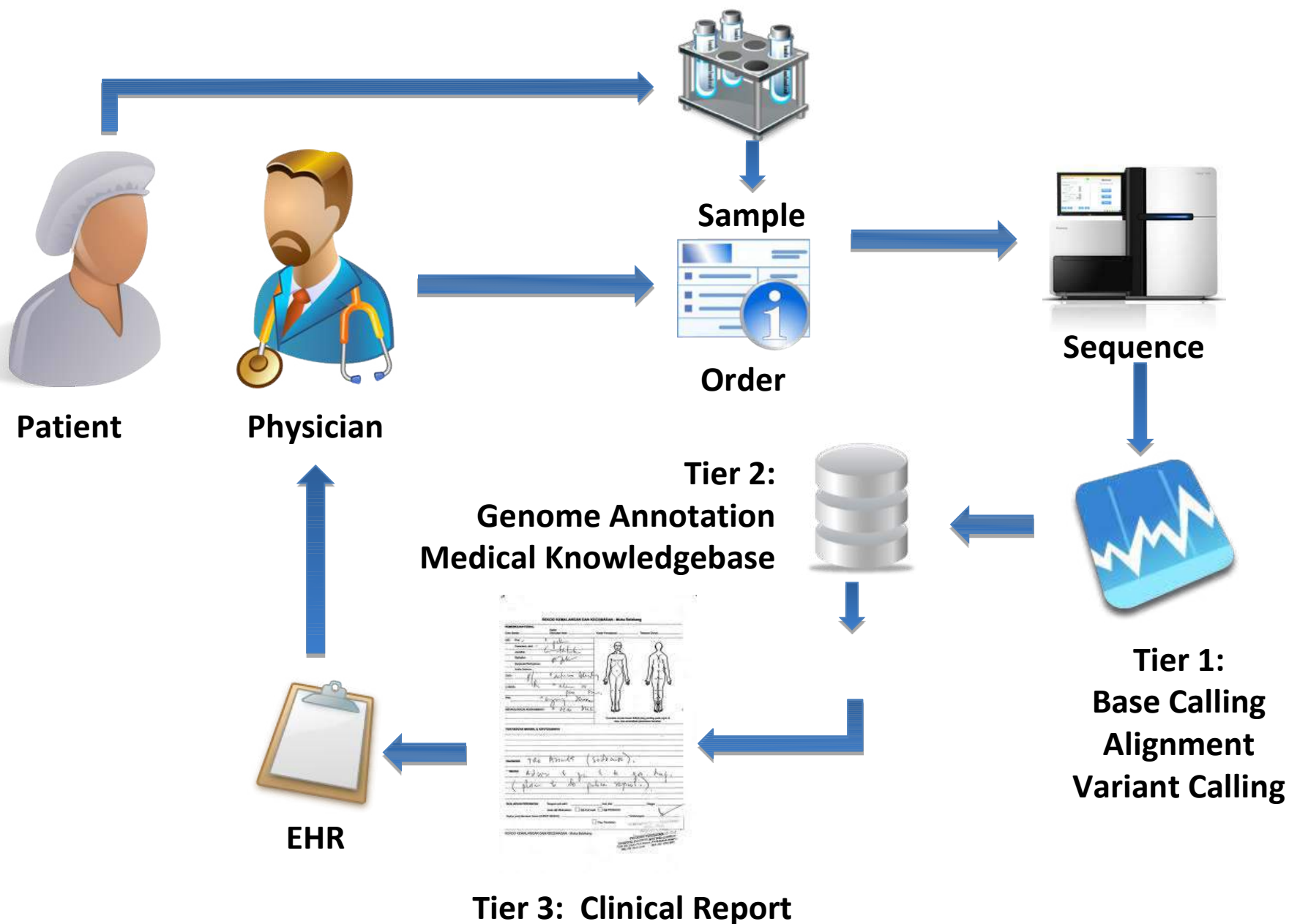
3944:9734#0/GGCTTAAACCCAGCCCTGCCCTGCCAGTCCA

108:13945#0/

Bioinformatics and Health Informatics

If bioinformatics is the study of the flow of information in biological sciences, Health Informatics is the study of the information in patient care

Medicine: Informatics pipeline workflow

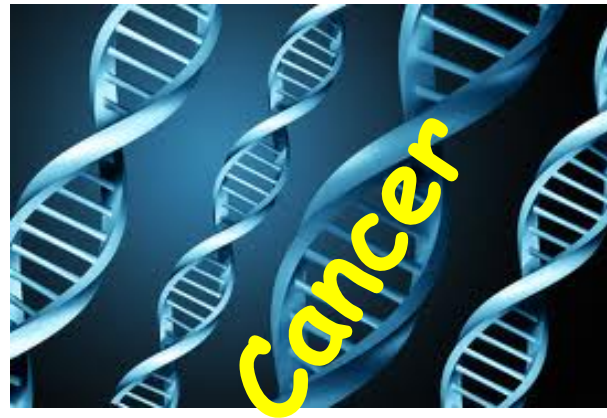


Huge need in bioinformatics tools

Simple pipelines/protocols and easy to read reports



Sample sequencing

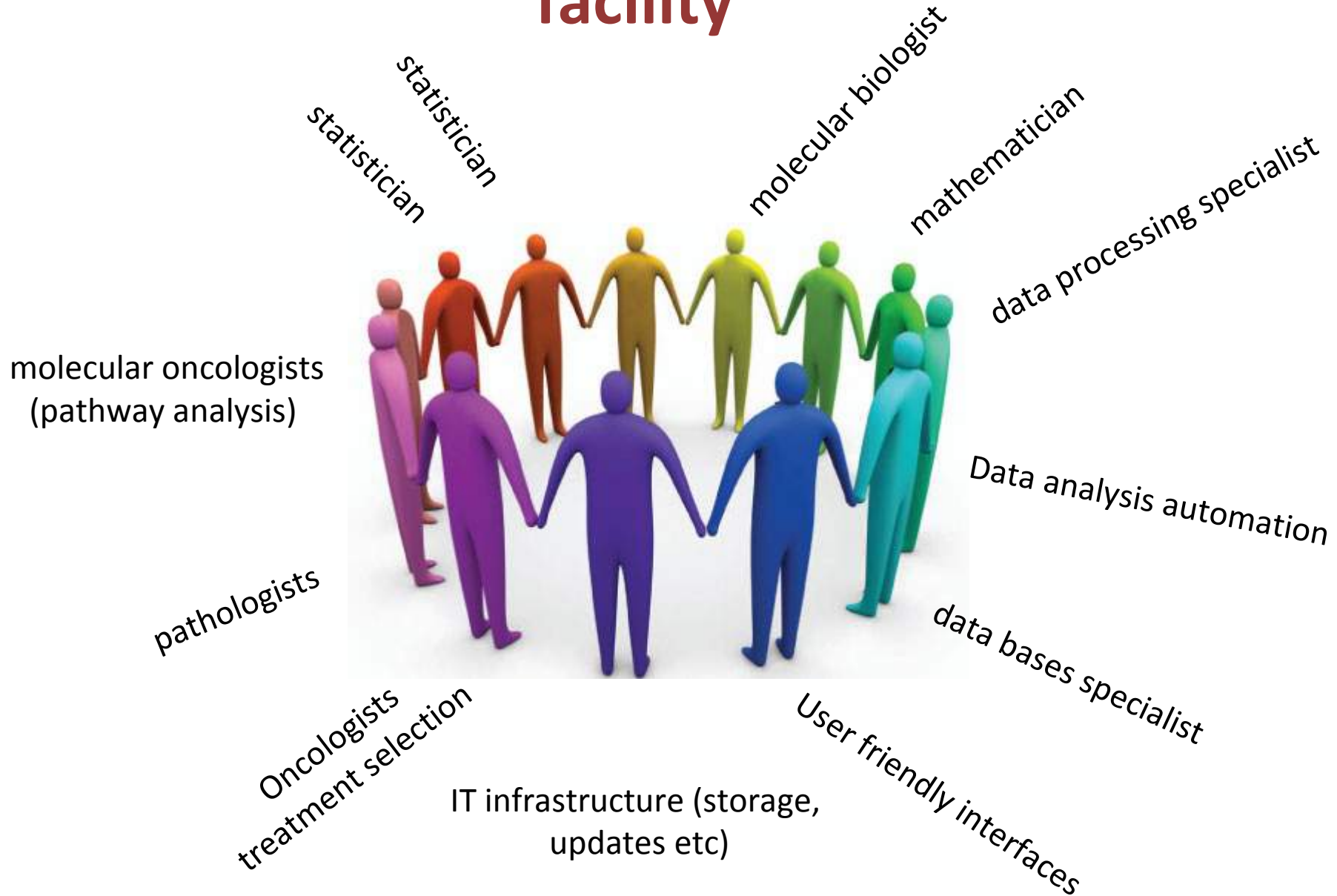


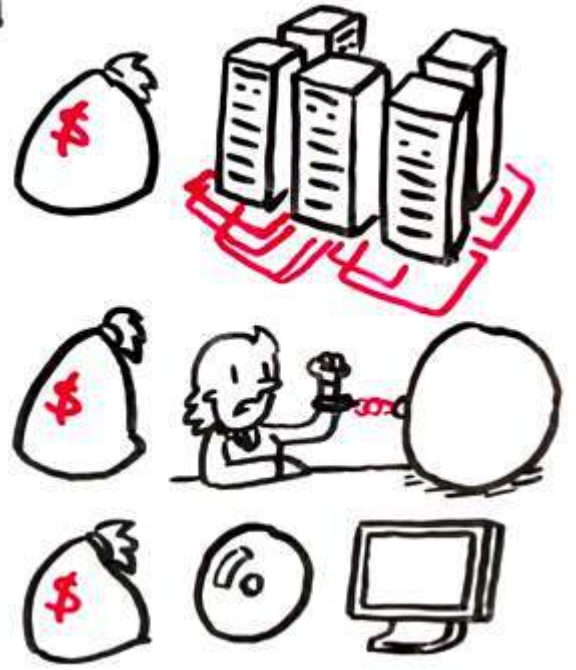
Data Analysis



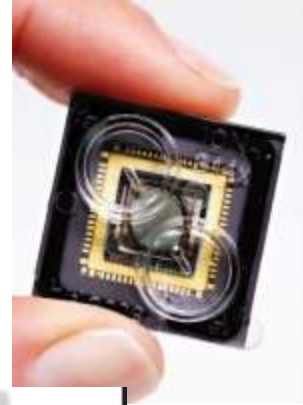
Patients treatment

Team work to set up cancer sequencing facility





Ion Torrent: Torrent Suite Software



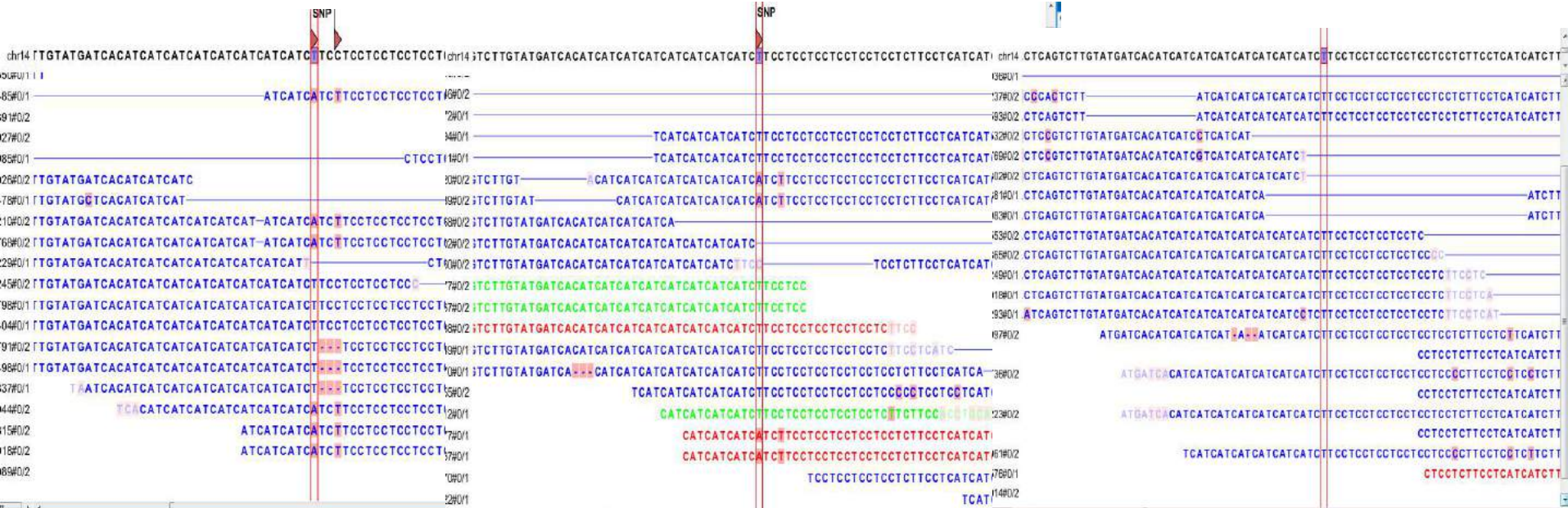
HOMEZ_22814666



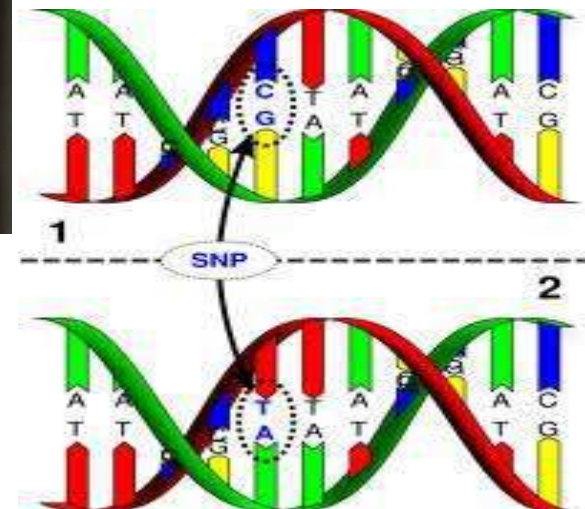
S2
S8

s4
s5

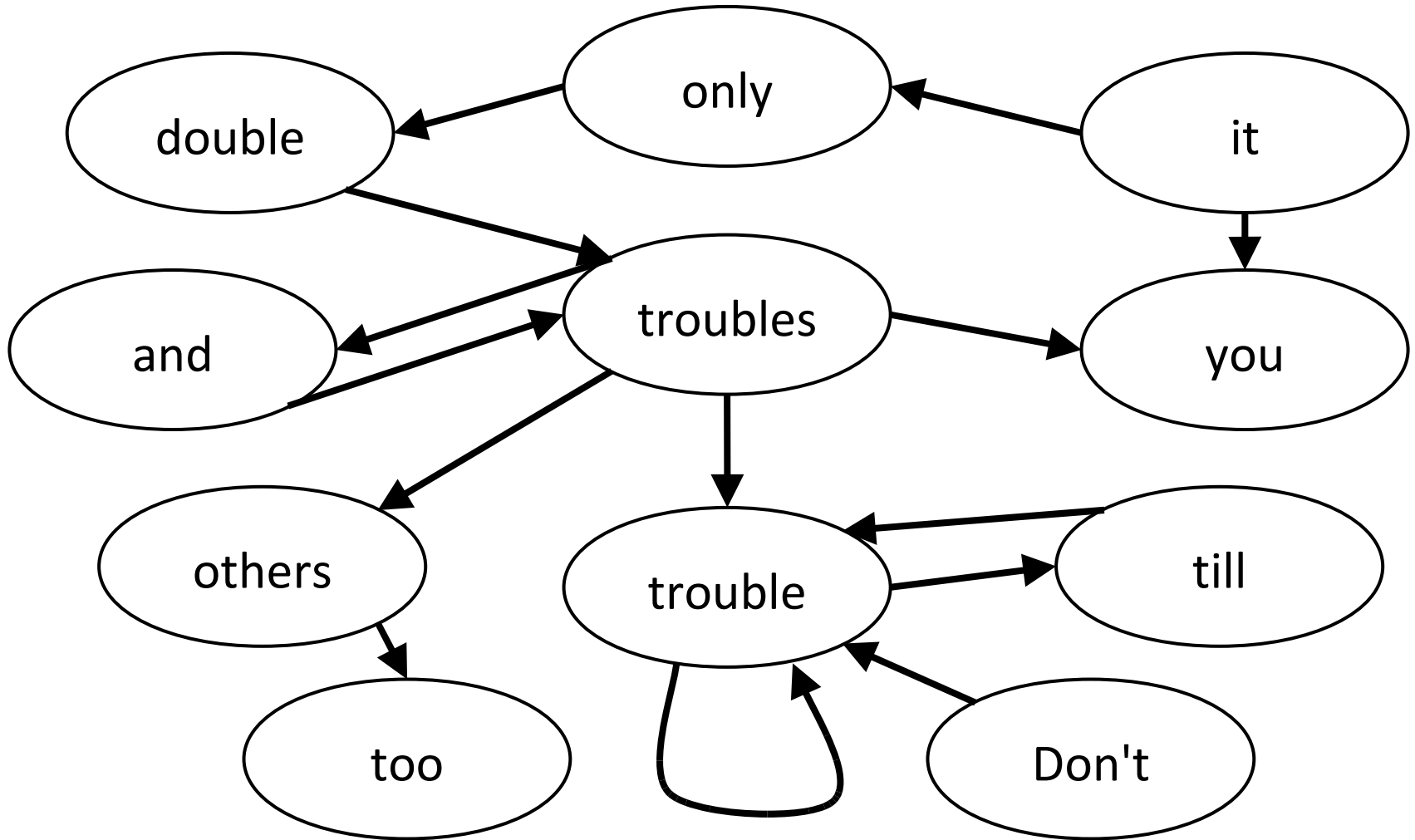
s3
s7



Each baby to be sequenced at birth: personal reference



Funny De Bruijn graph





THANK YOU!