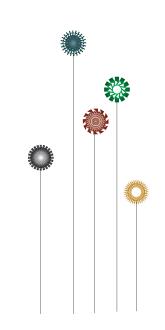


# Shared bioinformatics database within Unipro UGENE

BOSC July 11–12, 2014

#### by Ivan Protsyuk

UniPro LLC Novosibirsk State University





- About UGENE
- Shared databases in UGENE
  - generally
  - technically
  - numerically
- Prospects

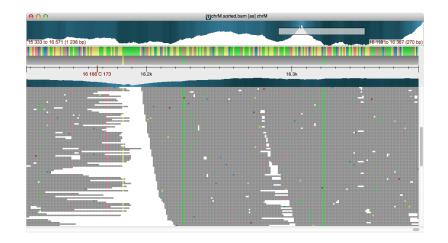




# About UGENE

Dozens of algorithms:

- BLAST
- Multiple and pairwise alignment
- Short reads alignment
- •And other...





- Visualization capabilities
- > 20 data formats
- Remote databases
- NGS framework

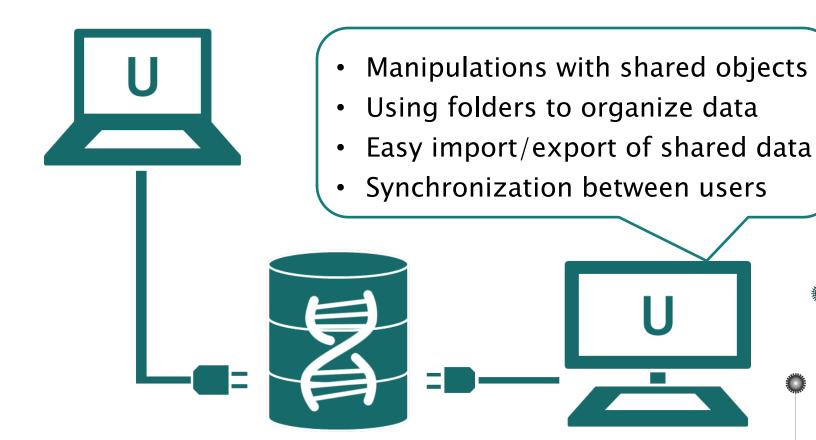




#### Data Storage

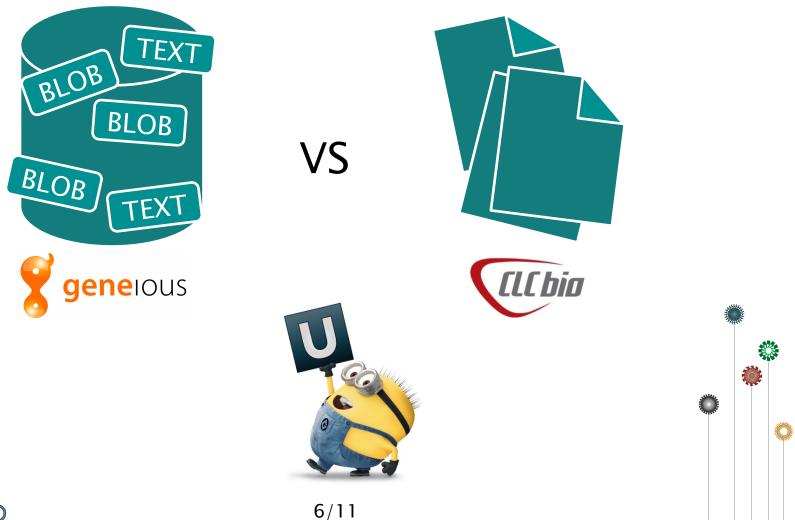
- Data can be big
- Simultaneous access

# Shared Bioinformatics Database



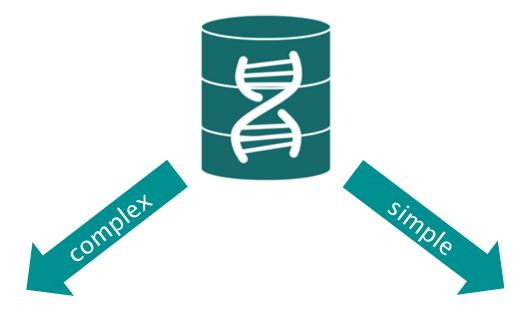


#### **Existing Approaches**





#### **UGENE Data Model**



- Sequences
- Annotations
- Multiple alignments
- Genome assemblies

- Phylogenetic trees
- Chromatograms
- Protein tertiary structure
- Weight matrices
- Text



# Storage Capabilities

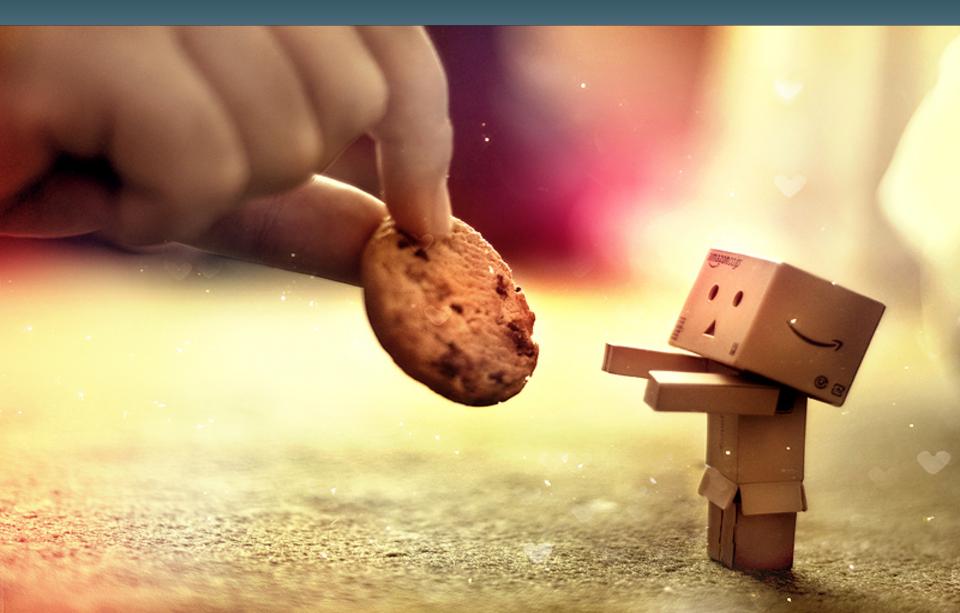
- Millions of biological objects
- NGS data supported
- Import 100 MBp ~ 1 min, but instant access later on





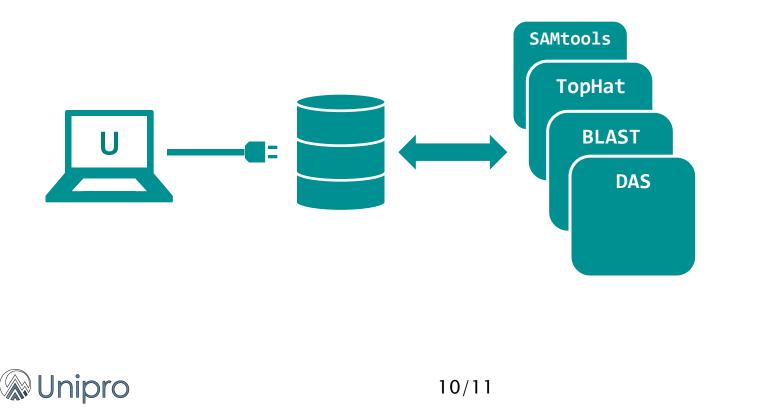


#### Public Storage for Everyone



#### Future Plans

- Performance improvements
- Distributed workflow system
- User requests



#### Contacts

#### UGENE:

http://ugene.unipro.ru/shared-db
ugene@unipro.ru

Ivan Protsyuk: iprotsyuk@unipro.ru

#### **Graphical View**

0	0 0	-* UGENE		
Co 🗁 🖶   A 🐚 🐚 🐚 🐚 🐚 🐚 🐨 Col 🖓 🗸 14 🗸 🤫 🍕 💐 💐 🗮 🖾 🛠 👯 🤗				
H	Project	00	℅ [s] NC_016856	
1: Project	Name filter	🚱 NC_016856 [dna]	O ⊠ - "> - C C= T≣ T - D € ⊂ 0, 0, =	
Pr	Objects	200k 400k 600k 800k 1m 1.2m 1.4m 1.6m		
	▼ 🗐 Lab 42	misc leader (E)		
- <b>1</b> 2	<ul> <li>Arabidopsis thaliana (TAIR 10)</li> </ul>	ncRNA (9)	ACE	
	C. elegans (ce6)	rRNA (22)		
	🕨 🗁 Drosophila melanogaster (dm3)	RNA (86) 500k 1m 1.5m	2m 2.5m 3m 3.5m 4m 4.6m 4.870.265	
	Escherichia coli str. K-12 substr. MG1	1 000 1m 1.5m	2m 2.3m 5m 3.3m 4m 4.3m 4.0m 4.0m	
	▶ 🗁 Human (hg19)			
	Human Immunodeficiency Virus (HIV-2)	R P G F C * T H R I L	S F V D I L F L L I N K L M L I S S F V Y	
	<ul> <li>Mouse (mm9)</li> <li>Mycobacterium tuberculosis (NC_0009</li> </ul>	APRLLLNASYP	LIC Y F V F T D K Q I D V N Q Q F C I	
	Recycle bin	CGCCCAGGCTTCTGTTAAACGCATCGTATCCT	CTCATTTGTTGATATTTTGTTTTTACTGATAAACAAATTGATGTTAATCAGCAGTTTTGTATAC	
	<ul> <li>Salmonella Enterica(NC_016856.1)</li> </ul>	14 797 14805 14810 14815 14820 14825 GCGGGTCCGAAGACAATTTGCGTAGCATAGGA	14830 14835 14840 14845 14850 14855 14860 14865 14867 14875 14880 14885 14 892 GAGTAAACAACTATAAAACAAAATGACTATTTGTTTAACTACAATTAGTCGTCAAAACATATG	
		R G P K Q * V C R I R	ENTSIKNKSIGIGIAINAACAAAAGAIAIG	
	🚱 [a] NC_016856 features	G L S R N F A D Y G	R M Q Q Y K T K V S L C I S T L * C N Q I R	
	[t] INFO	A W A E T L R M T D E	* K N I N Q K * Q Y V F Q H * D A T K Y V	
	Vibrio cholerae (NC_002505)		Value	
	Yeast (Saccharomyces cerevisiae) (sac	Name Auto-annotations [Lab 42   NC_0168		
	Cebrafish (Danio rerio) (danRer7)	<ul> <li>MC_016856 features [Lab 42]</li> </ul>		
	▶ 👔 NC_014267.1.gb ▼ 👔 human_T1.fa	► Ø CDS (0, 5315)		
	[s] human_T1 (UCSC April 2002 chr7:1	🔻 🧭 gene (0, 5516)		
	► R HIV-1.aln	gene	190255	
	chrM.sam.bam.ugenedb	gene	3372799	
	chrM.fa	🕨 🗖 gene	28013730	
		🕨 🗖 gene	37345020	
		▶ □ gene	complement(51145887)	
		<ul> <li>gene</li> <li>gene</li> </ul>	complement(59667396) 76658618	
		<ul> <li>gene</li> </ul>	87299319	
		▶ □ gene	complement(93769942)	
		gene	complement(1009210805)	
ç	💆 <u>2</u> : Tasks 🛛 💷 <u>3</u> : Log No active tasks 💡 🗎			