ЛЕКЦИЯ:

Медицинская биоинформатика - настоящее и тенденции

Часть 1: Геномные данные - в поисках клинического смысла.

Павлов Александр

Руководитель проекта

Часть 2: Практические аспекты разработки ПО

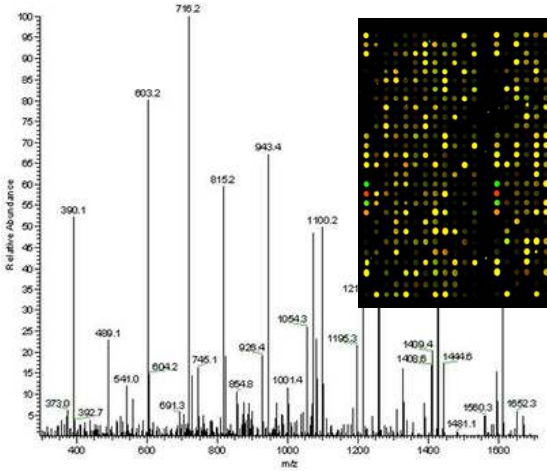
Брагин Антон

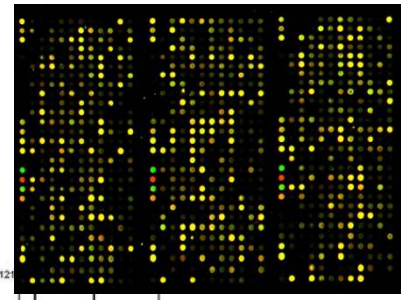
Руководитель группы биоинформатики

Москва

1 августа 2013 г.

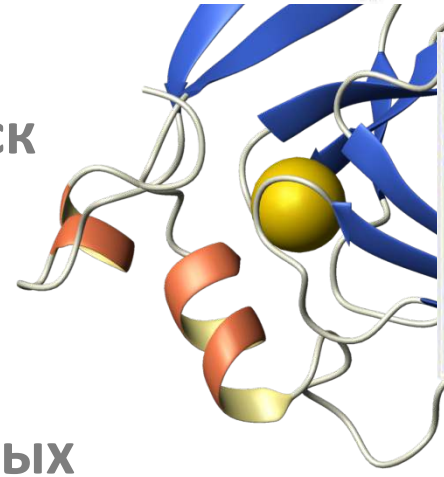
Биоинформатика в медицине





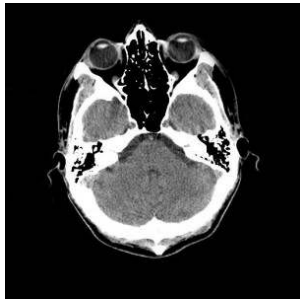
Области

- Масс-спектрометрия
- Анализ данных microarray
- Анализ данных секвенирования
- Биомоделирование
- Базы данных и эффективный поиск
- Анализ изображений



Предпосылки

- Рост объемов генерируемых данных
- Потребность в комплексной аналитике – data mining
- Виртуализация wet lab
- Data-driven research



Роль геномики в медицине

Области

- Наследственные заболевания
- Онкодиагностика
- Персонализированная медицина
- Пренатальная диагностика
- Medical research

Задачи

- Поиск генетических вариантов
- Установление ассоциаций (генотип - фенотип)
- Пациент, как «омный» профиль

Препятствия на пути

- Гетерогенность входных данных
- Критерии качества результатов
- Проблема стандартов (форматы, описания фенотипов, гайдлайны)
- Базы данных и их валидация

- Трудность в формулировании задач биологами
- Трансляция результатов
- Уровень достоверности
- Культура работа с данными (scientific data management)

Что можно сделать?

Уровень техники

- Введение стандартов для входных данных и метрики их качества
- Обеспечение прослеживаемости путей миграции данных
- Создание набора инструментов (инфраструктура + ПО) для разных задач и пользователей

Уровень концепции

- Стандарты для трансляционных исследований
- Введение принципов «информационной гигиены» при работе с клиническими данными
- Выработка требований к валидации процессов и результатов

Эволюция представлений

октябрь 2007

If you would like to purchase a private copy of your personal genome, please contact us.

↓

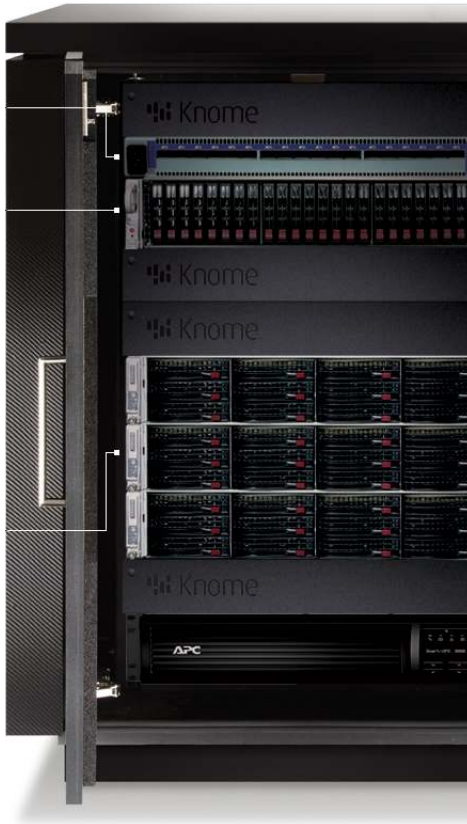
декабрь 2007

Knome Launches First Commercial Whole-Genome Sequencing and Analysis Service for Individuals

↓

13.07.2011

→



11.07.2013

Критическое мышление

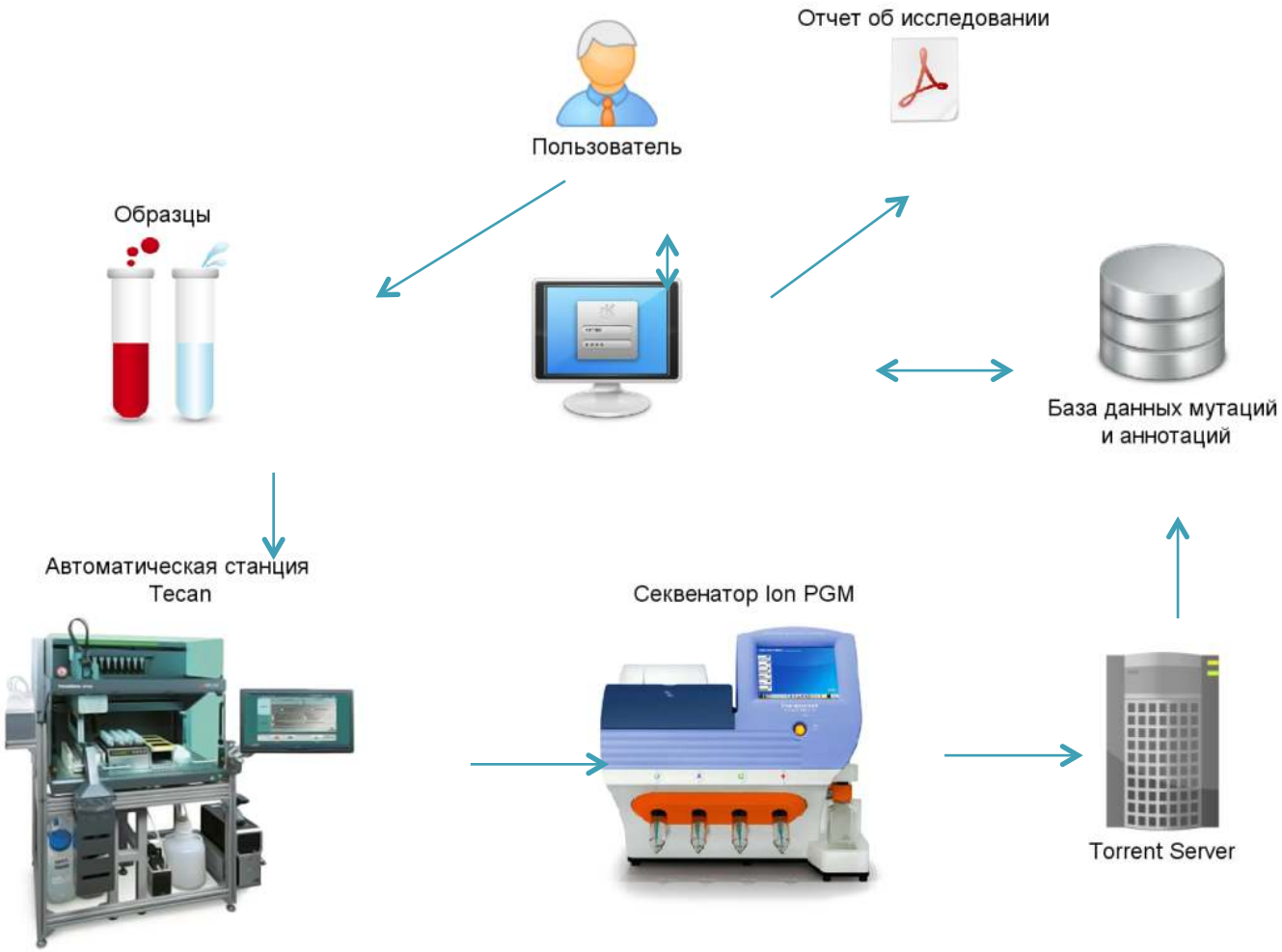
Новый технологический уровень исследований при старой парадигме интерпретации

Boston University:
GMS GE 705: Critical Thinking in Genetics and Genomics

Mike Snyder

Наш вклад

Неонатальная NGS-генодиагностика



Медицинская биоинформатика
- настоящее и тенденции

Практические аспекты разработки ПО

Антон Брагин
Sequoia Genetics
abragin@sequoiag.com

The Elements of Bioinformatics

Year of first release

Xx (00)

Tool name (XX)

KEY TO TOOL TYPE

- Aligners pairwise
- Aligners short read
- Multiple Sequence Aligners
- Assemblers Genomic Long read [L]
- Assemblers Genomic Short read [S]
- Assemblers mRNA
- Gene Prediction mRNA
- Gene Prediction ncRNA
- Database/Warehouse
- Workflows
- Genome Browsers
- ToolKit & APIs

OPEN SOURCE TOOLS

90 Bn BLAST NCBI	08 So SOAP													81 Py PHYLIP	00 Em EMBOSS	00 Bj BIOJAVA	01 Bc BIOCONDUCTOR	09 Qi QIIME
98 Hm HMMER	07 Ma MAQ												01 Bs BIOSQL	00 Eh eHIVE	00 En ENSEMBL API	00 Br BIORUBY	03 La LAGAN	09 Vc VCFTOOLS
99 Mu MUMmer	08 Bo BOWTIE	94 Cl CLUSTALW	06 Pe PECAN	97 Mi MIRA 97 [L]	07 Oz OLIGOZIP	07 Al ALLPATHS-LG [S]	09 Sd SOAP: DE NOVO [S]	03 Un UNIGENE	98 Gl GLIMMER	08 Mg MGENE	97 Tr tRNASCAN	03 Bm BIOMART	03 Ta TAVERNA	00 Eb ENSEMBL BROWSER	00 Bp BIOPERL	07 Pl PLINK	09 Ga GATK
03 Ss SSAHA	09 Bw BWA	00 Cf COMBINE, RCOFFEE T-COFFEE, MCOFFEE, EXPRESSO,	10 Pr PRANK	97 Ce CELERA [L]	09 Fo FORGE [L]	07 Ve VELVET [S]	09 Sg SGA [S]	09 Cu CUFFLINKS	99 Eu EUGENE	09 Au AUGUSTUS	07 Mp miPRED	07 In INTERMINE	07 Gx GALAXY	02 Ap APPOLO	04 Ig IGB	07 Ut UTGB	09 Sa SAMTOOLS
05 Ex EXONERATE	09 Bf BFAST	02 Ma MAFFT	04 Mv MAVID	02 Ar ARACHNE [L]	10 Me MERACULOUS [L]	08 Co CORTEX_CON_RP [S]	10 Pa PE-ASSEMBLER [S]	10 Oa OASES	00 Ep ENSEMBL PIPELINE	11 Sn SNAP	08 Mr miRDEEP	10 Ch CHADO	08 Kn KNIME	02 Gb GBROWSE	04 Nc NCBI-MAPVIEWER	07 Ag ARGO	10 Am AMPLICONNOISE
	10 Sm SMALT	04 Ml MUSCLE	05 Pb PROBCONS	03 Ph PHUSION(2) [L]	10 Qu QUAKE [L]	09 Ba BAMBUS 09 [S]			05 Ji JIGSAW		10 Rn RNAFOLD			02 At ARTEMIS	07 Jb JBROWSE	08 Iv IGV	11 Da DALLIANCE BROWSER
				04 Se SEQCLEAN	11 Pi PRICE [L]												

TOOLS FREE FOR ACADEMICS ONLY

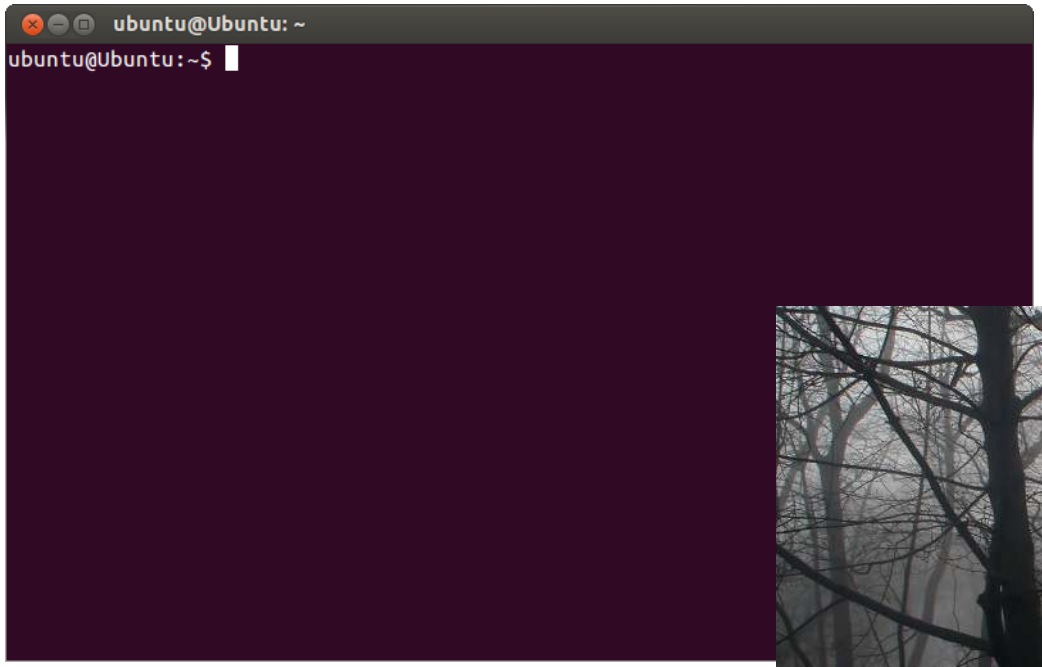
02 Bl BLAT	XX Vm VMATCH	01 Sh SSAHA2	XX Fa FASTA	XX Cm CROSSMATCH	XX No NOVOALIGN	09 Ab ABYSS [S]	99 Fg FGENESH	11 Mk MAKER	97 Gn GENSCAN	02 Uc UCSC

COMMERCIAL TOOLS

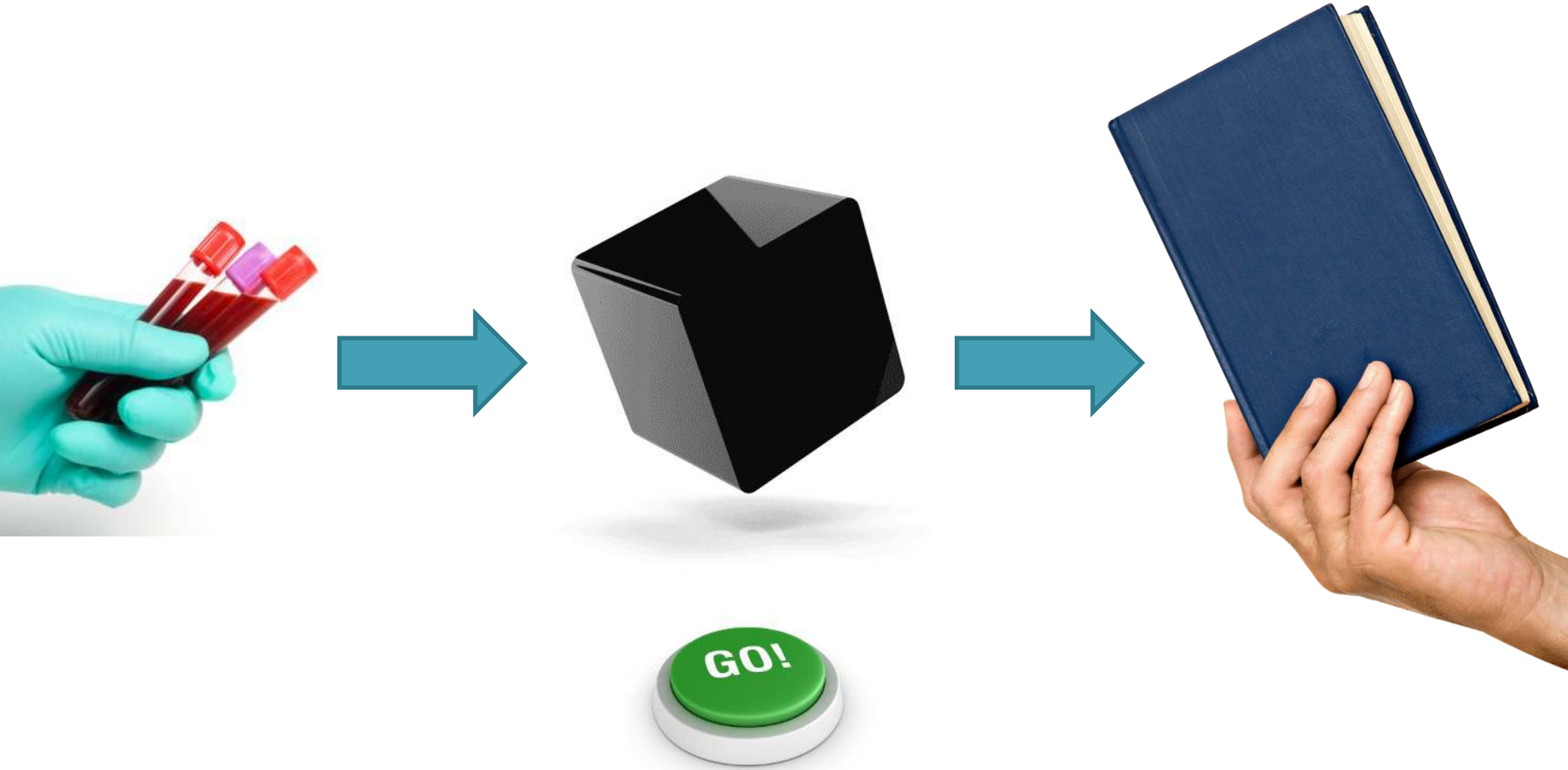
XX Zm ZOOM	01 Pc PCAP	11 Ne NEWBLER	XX Pp PIPELINE PILOT	XX Gs GENESPRING	05 Cb CLC BIO

www.eaglegenomics.com

Специалисты из разных областей по-разному воспринимают одни и те же вещи



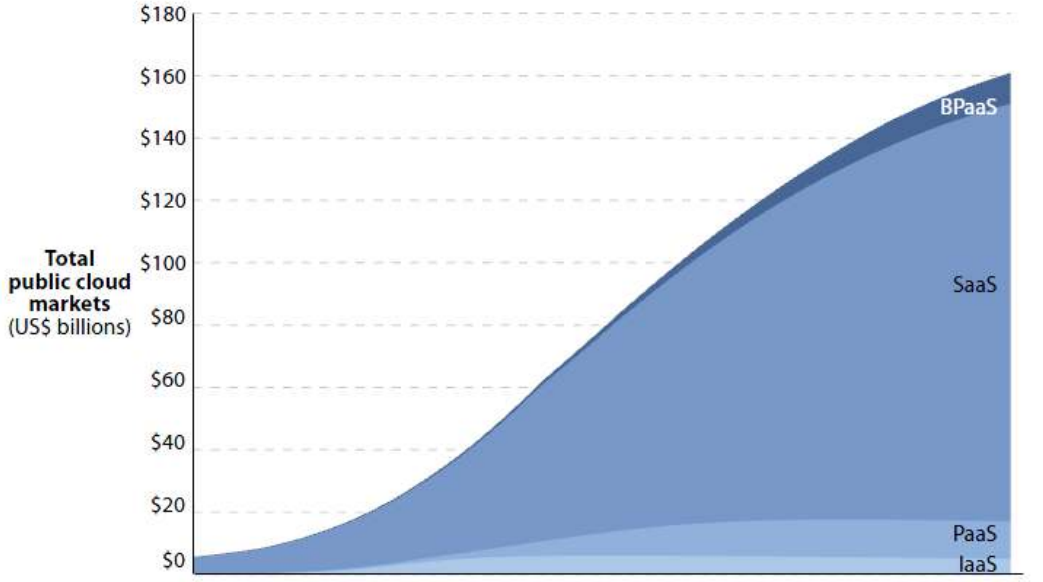
Текущие представления о работе биомедицинского ПО



Общая тенденция – от продукта к сервису

Figure 3 Forecast: Global Public Cloud Market Size, 2011 To 2020

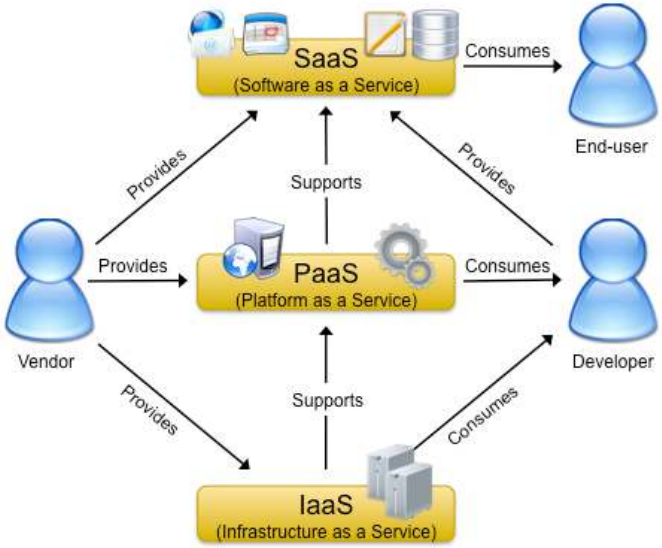
The spreadsheet detailing this forecast is available online.



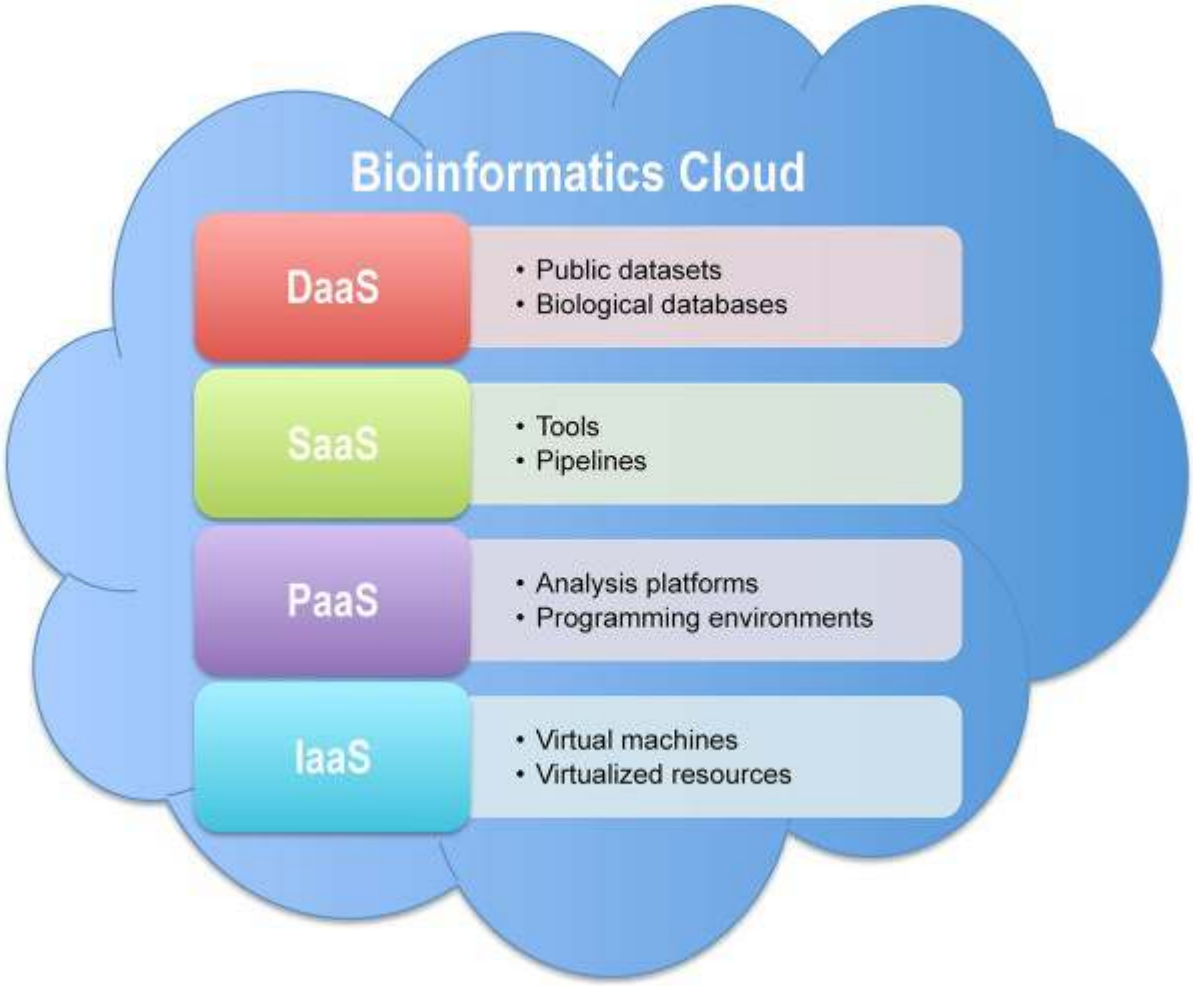
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
BPaaS ($)	0.15	0.23	0.35	0.53	0.80	1.26	1.95	2.93	4.28	6.00	7.66	9.08	10.02
SaaS ($)	5.56	8.09	13.40	21.21	33.09	47.22	63.19	78.43	92.75	105.49	116.39	125.52	132.57
PaaS ($)	0.05	0.12	0.31	0.82	2.08	4.38	7.39	9.80	11.26	11.94	12.15	12.10	11.91
IaaS ($)	0.06	0.24	1.02	2.94	4.99	5.75	5.89	5.82	5.65	5.45	5.23	5.01	4.78

58161

Source: Forrester Research, Inc.

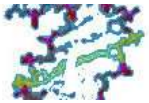


Облачные сервисы в биоинформатике



e.g.

dbSNP
Short Genetic Variations



Crossbow
Genotyping from short reads using cloud computing

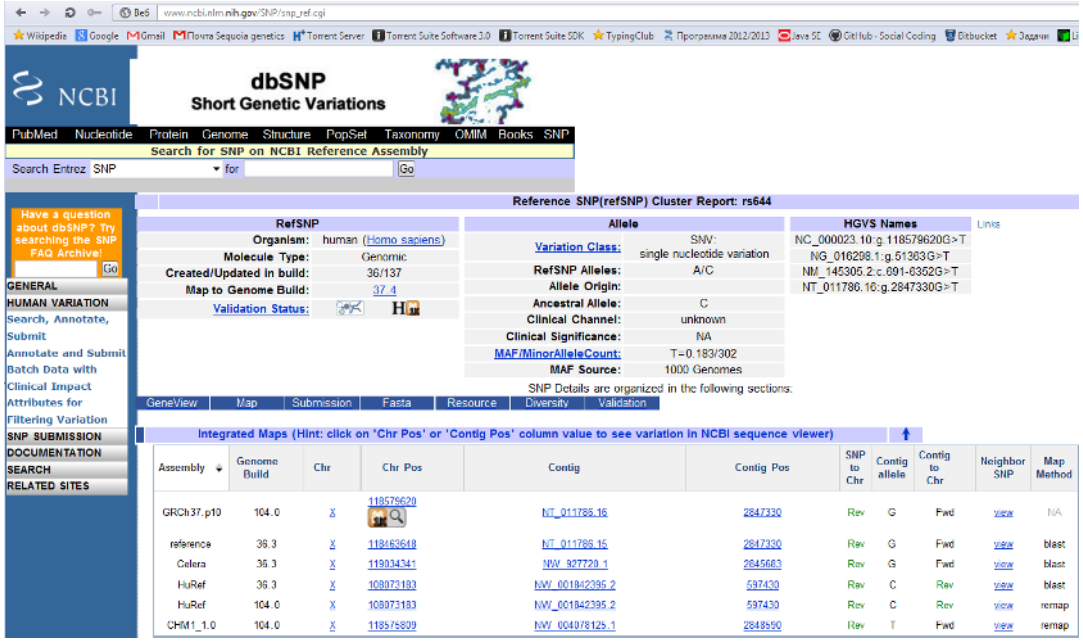
Galaxy

Cloud BioLinux

Lin Dai 2012

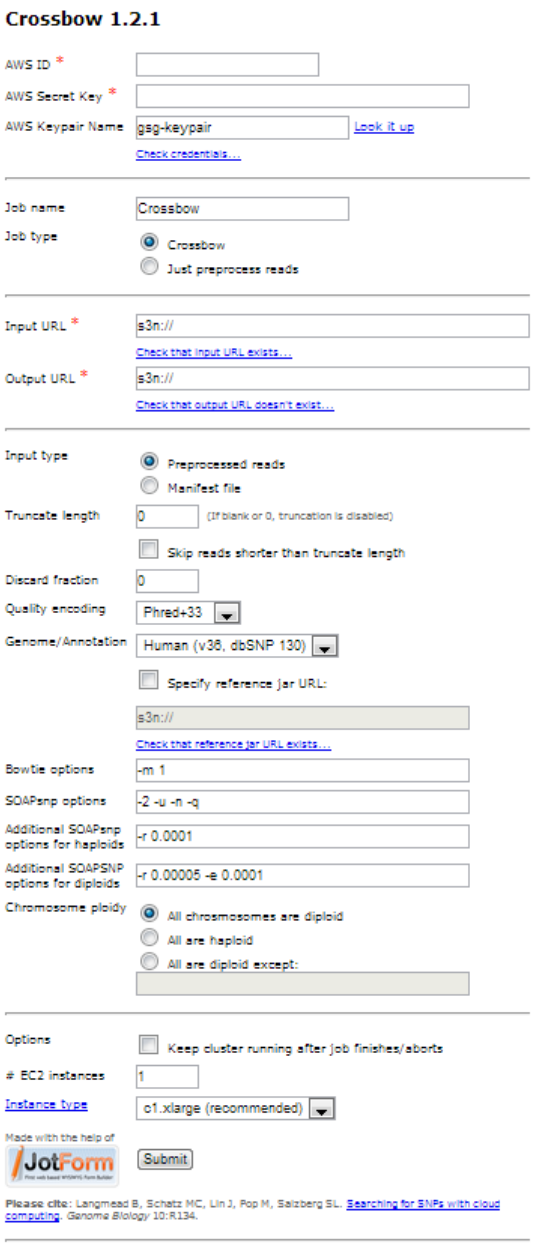
DaaS: dbSNP

- Публично доступные данные (часто с возможностью внесения данных)
- Наличие API, позволяющего осуществлять программный доступ к данным
- Возможность развертывания локального экземпляра базы данных
- Наличие публичных копий БД в облаке (например, AWS)

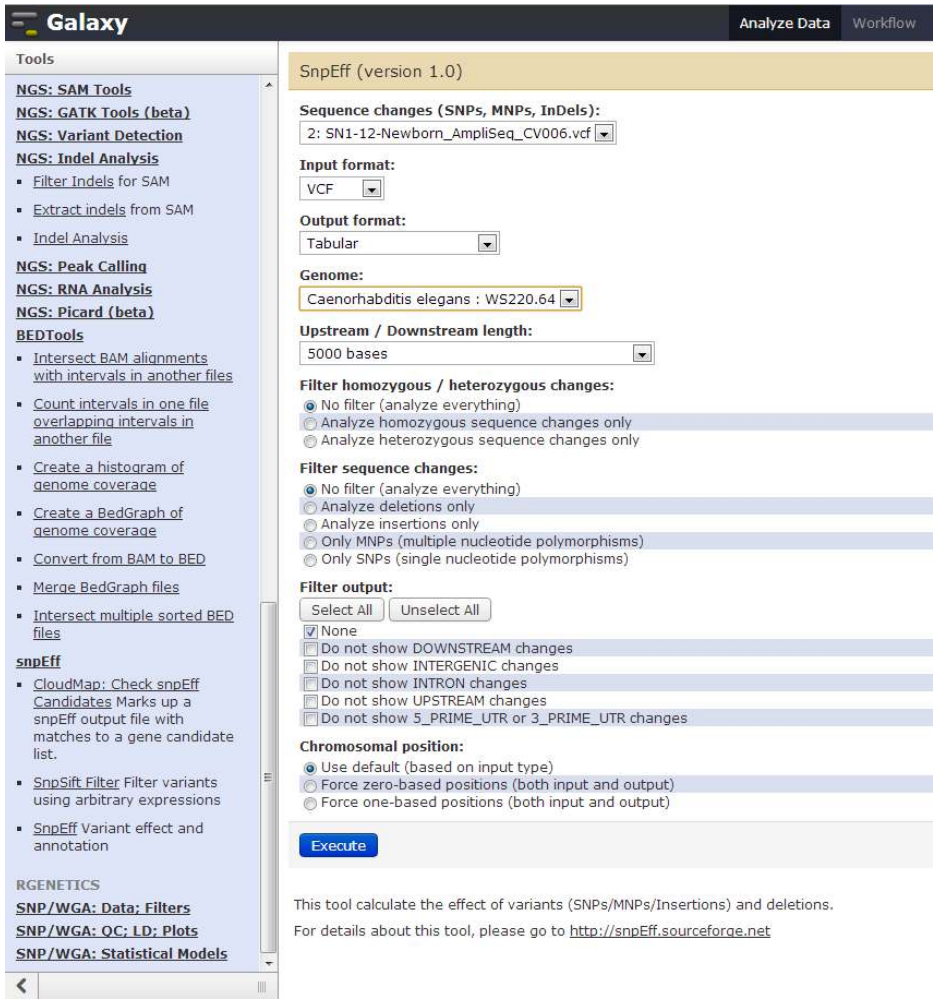


SaaS: Crossbow

- Предназначено для конечного пользователя
- Ограниченный, четко определенный функционал



PaaS: Galaxy

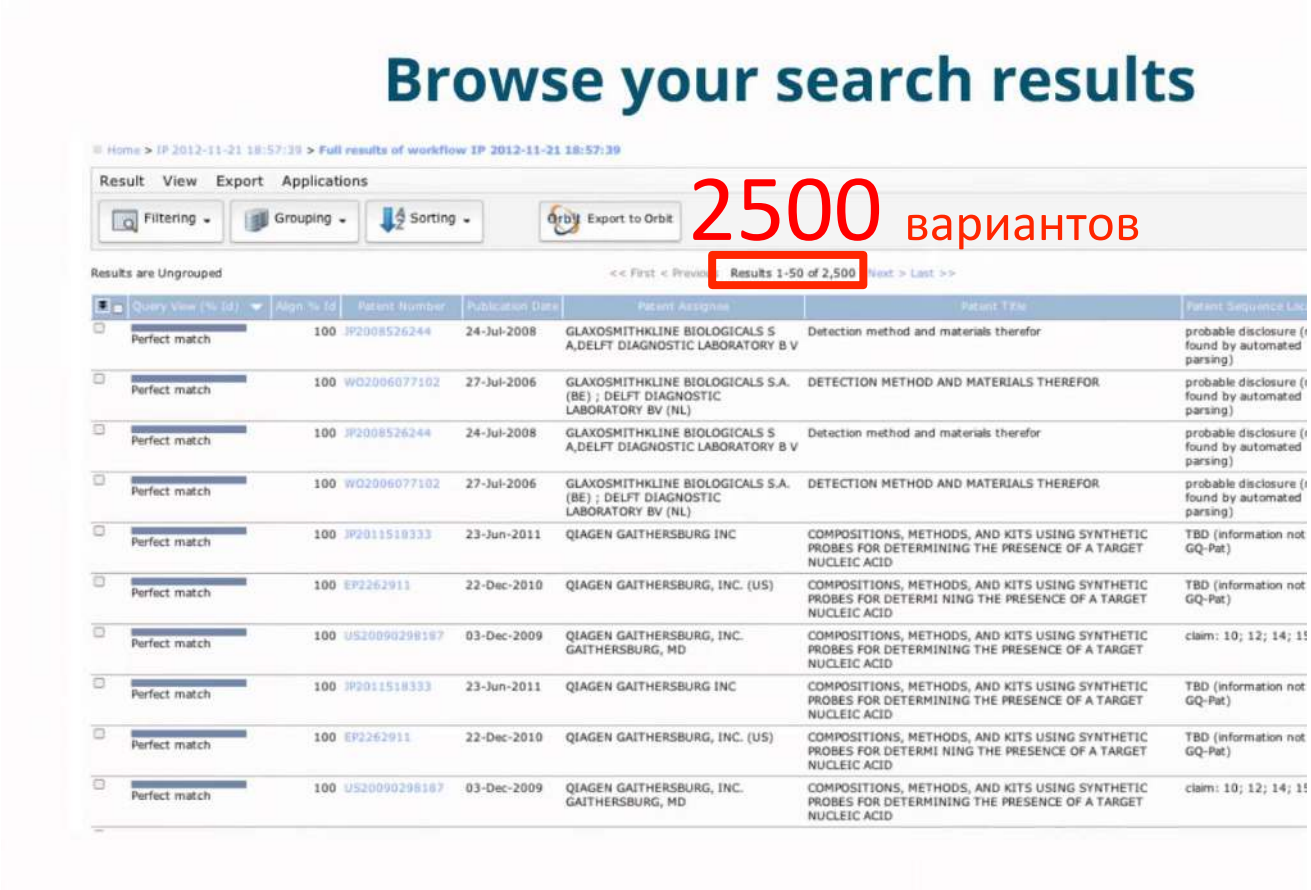


- Возможность построения собственных pipeline’ов анализа
- Расширяющийся функционал
- Интеграция функциональных модулей из разных источников (samtools, snpEff и др.)
- Возможность хранения данных

PaaS: Cloud BioLinux

- Возможность развертывания в пределах произвольной IT-инфраструктуры
- Расширенные возможности настройки
- Необходимость построения собственных pipeline’ов анализа
- Интеграция функциональных модулей из разных источников
- Возможность хранения данных в облаке

Сложность процесса не всегда коррелирует с полезностью результатов



ПО для анализа данных секвенирования следующего поколения (NGS)

Свойства, которыми должно обладать программное обеспечение для анализа клинического генетических данных:

1. Глубина анализа и акцент на клинически-значимых аспектах
2. Дружественность к пользователю и простота работы
3. Стабильность
4. Соответствие предоставляемой информации актуальным данным
5. Кроссплатформенность
6. Локализуемость

Два типа трудностей: технические и концептуальные

Как обозначить генетический вариант?

ATGCGTGGACTGATGCGAGTCGGATGTAATAGGTGCTGAGAGG
ATGCGTGGACTGATGCGAGTCGCATGTAATAGGTGCTGAGAGG

Идентификация генетических полиморфизмов в общедоступных базах данных

Несмотря на несколько попыток формализации, идентификация мутаций в настоящее время основана на *de facto* стандартах, таких как форматы файлов и структура баз данных

HGVS
HUMAN GENOME VARIATION SOCIETY

Mutation Nomenclature Extensions and Suggestions to Describe Complex Mutations: A Discussion

Johan T. den Dunnen[1*] and Stylianos E. Antonarakis[2*]
[1]MGC-Department of Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
[2]Division of Medical Genetics, University of Geneva Medical School, Geneva, Switzerland

Consistent gene mutation nomenclature is essential for efficient and accurate reporting, testing, and curation of the growing number of disease mutations and useful polymorphisms being discovered in the human genome. While a codified mutation nomenclature system for simple DNA lesions has now been adopted broadly by the medical genetics community, it is inherently difficult to represent complex mutations in a unified manner. In this article, suggestions are presented for reporting just such complex mutations. Hum Mutat 15:7–12, 2000.

KEY WORDS: complex mutation; mutation detection; mutation database; nomenclature; MDI

Letters to the Editor

GERALDINE A. MCDOWELL*
AND MIRIAM G. BLITZER†
*Section on Human Biochemical Genetics, Human Genetics Branch, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda; and †Division of Human Genetics, Departments of Pediatrics and Obstetrics and Gynecology, University of Maryland School of Medicine, Baltimore

Reference

Deiler JH (1909) The settlement of the German coast of Louisiana and the Creoles of German descent. Americana Germanica, Philadelphia

0002-9297/93/5303-0027$02.00

Am. J. Hum. Genet. 53:783–785, 1993

The Designation of Mutations

To the Editor:

Many different conventions are used for the primary designation of mutations. Commonly, the amino acid change that has been deduced from the nucleotide substitution is employed, but often the cDNA number, the genomic number, or even a "nickname" based on a restriction site or a patient's name has been used. These notations are not, of course, mutually exclusive, and several of them are used in first describing a mutation.

Структура VCF-файлов

##fileformat=VCFv4.1 *заголовок*
##fileDate=20090805
… *варианты*

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
4	14370	rs6054257	G	A	29	PASS	…
16	87465	rs6040355	ATGC	A	30	PASS	…
21	14489	rs5896654	G	GAT	28	PASS	…

Преимуществом VCF является возможность добавления структурированной информации в поля INFO

dbSNP представляет собой наиболее полный общедоступный источник информации о генетической вариации

Представленная в виде VCF информация о генетических полиморфизмах человека представляет из себя файл размером около 8 Гб, который содержит:

53 417 385 записей, среди которых

SNP/MNP (86,4%)	46 177 146
Инсерции	3 088 167 (5,8%)
Делеции	4 152 072 (7,8%)

75 126 записей аннотированы как клинически-значимые

VCF в качестве формата для хранения данных генетических полиморфизмов

Пример 1:

GATTTCGA… -> GATTTTCGA…

Это изменение последовательности может быть представлено четырьмя различными записями VCF

Pos	Ref	Alt
2	A	AT
3	T	TT
4	T	TT
5	T	TT

VCF в качестве формата для хранения данных генетических полиморфизмов

Пример 2:

TCAGAGCAT… -> TCAGCAT…

Это изменение последовательности может быть представлено тремя различными записями VCF

Pos	Ref	Alt
2	CAG	C
3	AGA	A
4	GAG	G

VCF в качестве формата для хранения данных генетических полиморфизмов

Пример 3:

ACGTCCTCAG… -> ACGTCAG…

Это изменение последовательности может быть представлено тремя различными записями VCF

Pos	Ref	Alt
3	GTCC	G
4	TCCT	T
5	CCTC	C

VCF в качестве формата для хранения данных генетических полиморфизмов

VCF допускает двусмысленность в обозначении одних и тех же генетических вариантов, что приводит к драматическим последствиям при его использовании для аннотации:

• Одно и то же фактическое изменение последовательности может иметь несколько аннотаций, противоречащих друг другу.

• При сравнении обнаруженных при секвенировании полиморфизмов с записями базы данных возможна потеря данных аннотации (известные мутации могут быть классифицированы как неизвестные).

Для того, чтобы оценить масштабы вырожденности, была написана программа, анализирующая файлы VCF dbSNP

Оценка вырожденности записей dbSNP

Среди записей dbSNP:

Инсерции 3 088 167 (5.8%) в том числе 385437 синонимов
(0.7% от общего числа мутаций, 12.5% от числа инсерций)

Делеции 4 152 072 (7.8%) в том числе 317437 синонимов
(0.6% от общего числа записей, 7.6% от числа делеций)
Общее число мутаций 53417385

Среди 75 126 клинически-значимых записей:

Инсерции 9 синонимов среди 3970 вариантов (0.2%)

Делеции 109 синонимов среди 1380 вариантов (7.9%)

Оценка вырожденности записей dbSNP

NCBI

dbSNP
Short Genetic Variations

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive! Go

GENERAL
HUMAN VARIATION
Search, Annotate, Submit
Annotate and Submit Batch Data with Clinical Impact Attributes for Filtering Variation NEW
SNP SUBMISSION
DOCUMENTATION
SEARCH
RELATED SITES

Reference SNP(refSNP) Cluster Report: rs113993960 ** With pathogenic allele **

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	132/137
Map to Genome Build:	37.3
Validation Status:	
Citation:	PubMed

Allele	
Variation Class:	DIV: deletion/insertion variation
RefSNP Alleles:	-/CTT
Allele Origin:	-:germline CTT:germline
Ancestral Allele:	Not available
Clinical Channel:	VarView OMIM
Clinical Significance:	With pathogenic allele [detail]
MAF/MinorAlleleCount:	NA
MAF Source:	

HGVS Names	Links
NC_000007.13:g.117199646_117199648delCTT	
NG_016465.1:g.84630_84632delCTT	
NM_000492.3:c.1521_1523delCTT	
NP_000483.3:p.Ile507_Phe508delinsIle	
NT_007933.15:g.55232489_55232491delCTT	

SNP Details are organized in the following sections:
GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence view

Assembly	Genome Build	Chr	Chr Pos	Contig	Contig Po
GRCh37.p5	37.3	7	117199646:117199648	NT_007933.15	55232489:5523
HuRef	37.3	7	139743898:139743900	NW_001839071.2	14089384:1408
CRA_TCAGchr7v2	37.3	7	116595083:116595085	NT_079596.2	16630651:1663

GeneView

GeneView via analysis of contig annotation: CFTR *cystic fibrosis transmembrane conductance regulator (ATP-bindi*

View more variation on this gene (click to hide).

Clinical Source: in gene region cSNP has frequency double hit Go

Primary Assembly Mapping

Assembly	SNP to Chr	Chr	Chr position	Contig
GRCh37.p5	Fwd	7	117199646:117199648	NT_007933.15

RefSeqGene Mapping

RefSeqGene	Gene (ID)	SNP to RefSeqGene
NG_016465.1	CFTR (1080)	Fwd

Gene Model(s)

Function	mRNA				
	SNP to mRNA	Accession	Position	Allele change	Acces
cds-indel	Fwd	NM_000492.3	1653:1655	NA ⇒ AT	NP_000

NC_000007.13: 117M..117M (3.0Kbp) | Find on Sequence:

117,197,500 | 117,198 K | 117,198,500 | 117,199 K

dbSNP
Short Genetic Variations

Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

for Go

Reference SNP(refSNP) Cluster Report: rs199826652

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	137/137
Map to Genome Build:	37.3
Validation Status:	

Allele	
Variation Class:	DIV: deletion/insertion variation
RefSNP Alleles:	-/TCT
Allele Origin:	
Ancestral Allele:	Not available
Clinical Channel:	unknown
Clinical Significance:	NA
MAF/MinorAlleleCount:	-=0.005/12
MAF Source:	1000 Genomes

HGVS Names	Links
NC_000007.13:g.117199645_117199647delTCT	
NG_016465.1:g.84629_84631delTCT	
NM_000492.3:c.1520_1522delTCT	
NP_000483.3:p.Ile507_Phe508delinsIle	
NT_007933.15:g.55232488_55232490delTCT	

SNP Details are organized in the following sections:
GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

Assembly	Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh37.p5	37.3	7	117199645:117199647	NT_007933.15	55232488:55232490	Fwd	TCT	Fwd	view	remap

GeneView

GeneView via analysis of contig annotation: CFTR *cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)*

View more variation on this gene (click to hide).

Clinical Source: in gene region cSNP has frequency double hit Go

Primary Assembly Mapping

Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position	Allele
GRCh37.p5	Fwd	7	117199645:117199647	NT_007933.15	55232488:55232490	TCT

RefSeqGene Mapping

RefSeqGene	Gene (ID)	SNP to RefSeqGene	Position	Allele
NG_016465.1	CFTR (1080)	Fwd	84629:84631	TCT

Gene Model(s)

Function	mRNA				Protein		
	SNP to mRNA	Accession	Position	Allele change	Accession	Position	Residue change
cds-indel	Fwd	NM_000492.3	1652:1654	NA ⇒ A	NP_000483.3	507	NA ⇒ I [Ile]

NC_000007.13: 117M..117M (3.0Kbp) | Find on Sequence: | Tools | Configure

117,197,500 | 117,198 K | 117,198,500 | 117,199 K | 117,199 | rs199826652 | 117,200 K

SNP

Формат VCF не подходит для хранения полиморфизмов в базе данных

1. Какой формат использовать для однозначной идентификации полиморфизмов?

2. Как обеспечить импорт данных из VCF, который является *de facto* стандартом?

Какой формат использовать для однозначной идентификации полиморфизмов?

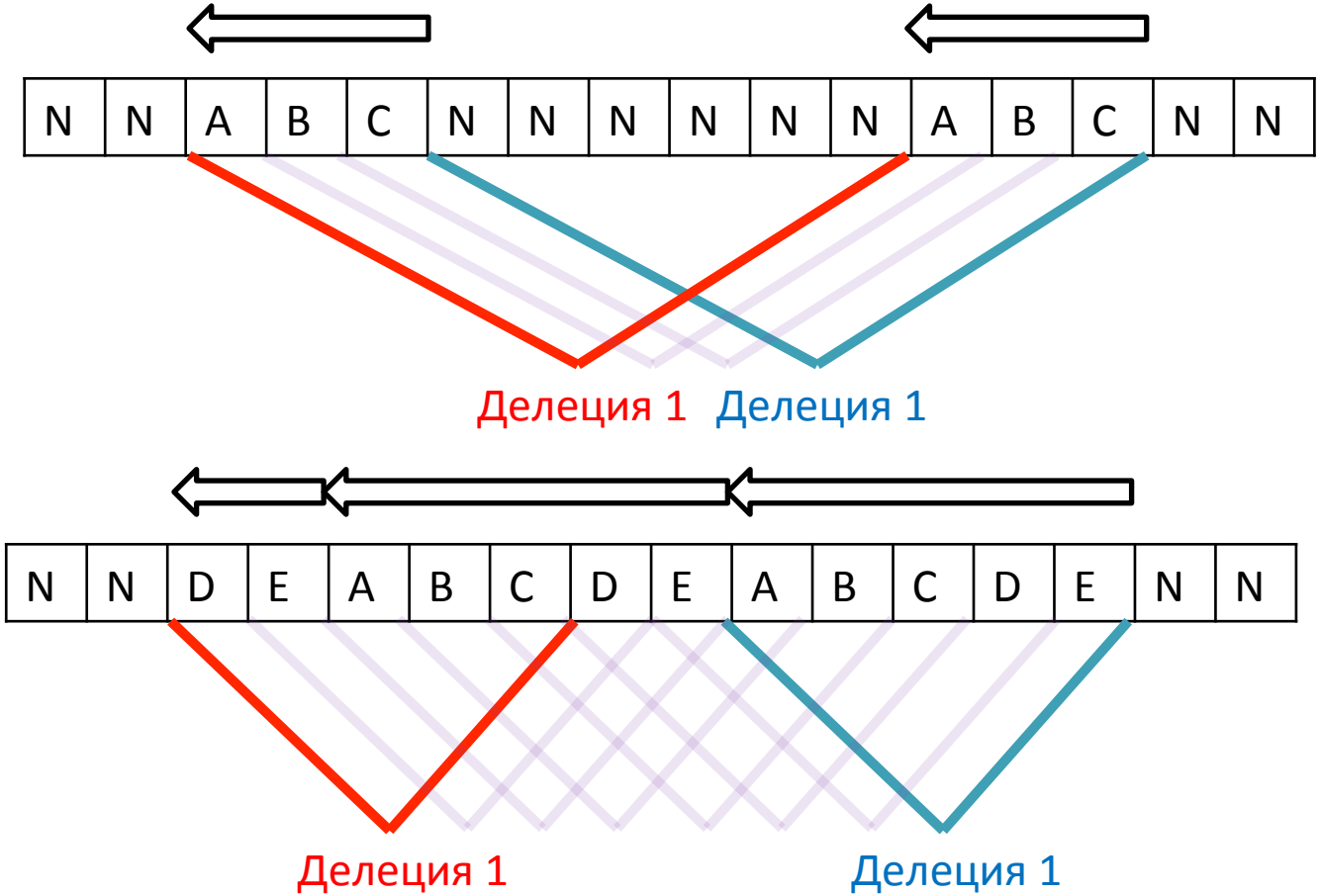
GATTTCGA… -> GATTTTCGA…

TCAGAGCAT… -> TCAGCAT…

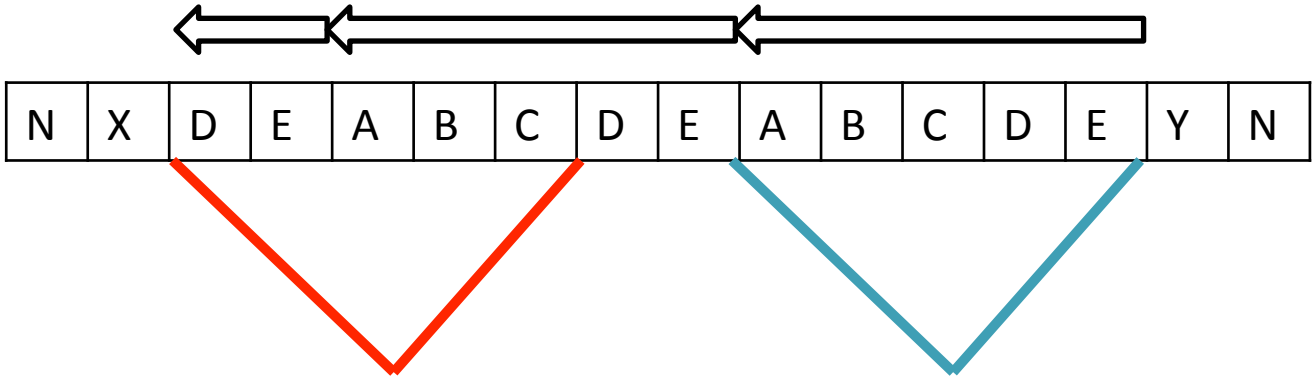
ACGTCCTCAG… -> ACGTCAG…

Вариант однозначно идентифицируется с помощью двух якорных оснований и последовательности, лежащей между ними в измененном варианте.

Как обеспечить импорт данных из VCF, который является *de facto* стандартом?



Как обеспечить импорт данных из VCF, который является *de facto* стандартом?



Ключ мутации

X	D	E	A	B	C	D	E	Y

Все варианты (как аннотированные, так и полученные в результате проведения секвенирования) перед помещением в базу данных преобразуются в недвусмысленный формат

Алгоритм преобразования VCF

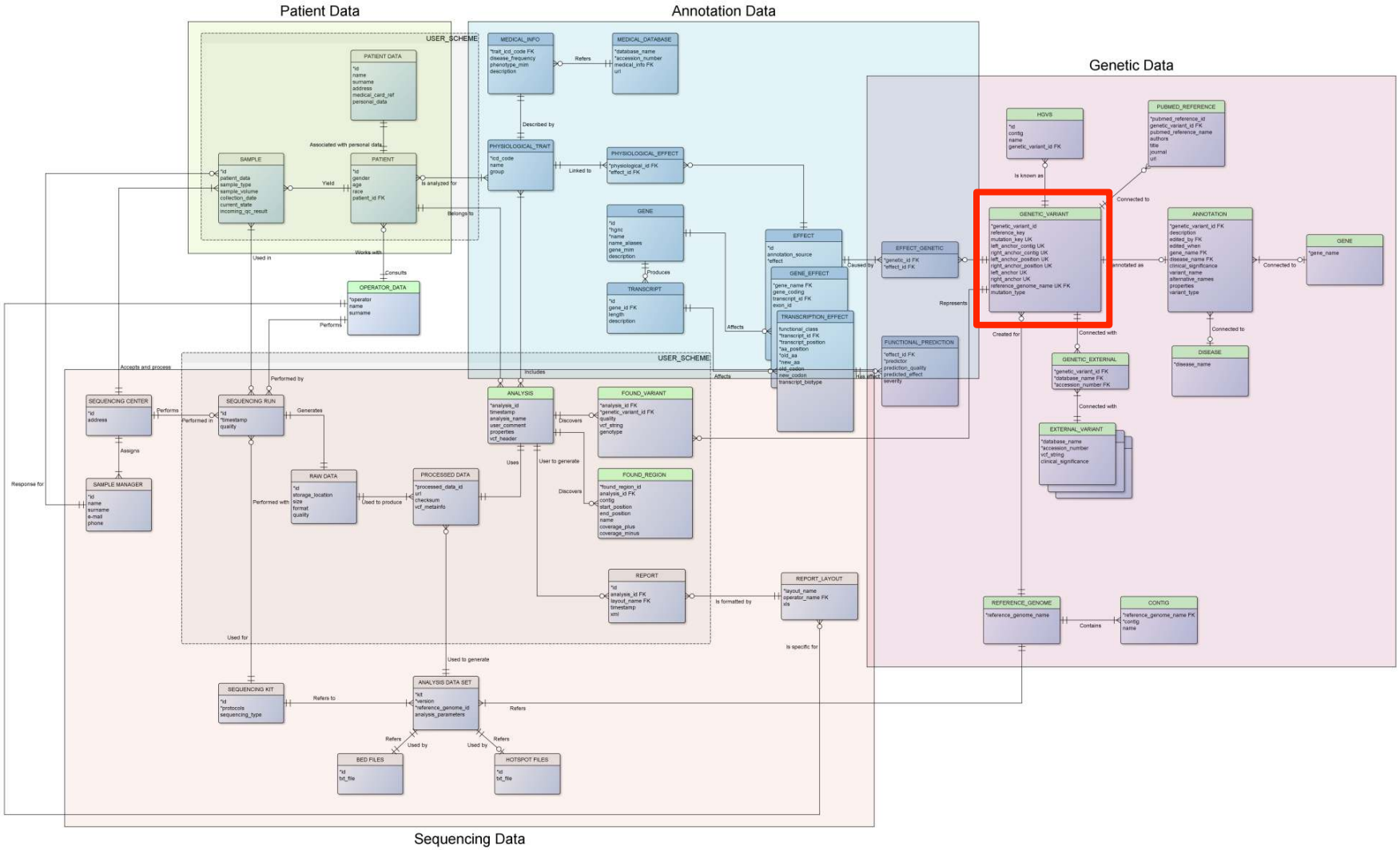
Algorithm 1 Creation of AKA identifier

Require: ref corresponds to $reference$ in appropriate positions
Require: ref and alt do not have similar bases in the same positions

```
procedure CreateIdentifier(ref[], alt[], coord, reference[])
    la ← coord − 1                                   ▷ Left anchor
    ra ← coord + length(ref)                         ▷ Right anchor
    change ← alt                                     ▷ Sequence change
    if length(alt) = 0 then                          ▷ This is deletion
        la, ra ← ExpandBorders(ref, la, ra, reference)
        change ← reference[la + 1, ra − length(ref)]
    else if length(ref) = 0 then                     ▷ This is insertion
        la, ra ← ExpandBorders(alt, la, ra, reference)
        change ← reference[la + 1 : coord] + alt + reference[coord : ra]
    end if
    return la, ra, change
end procedure

procedure ExpandBorders(seq[], la, ra, reference[])
    l ← length(seq)
    L ← length(reference)
    i ← la
    j ← ra
    while i ≥ 0 and seq[(l − (la − i − 1)) mod l] = reference[i] do
        i ← i − 1
    end while
    while j < L and seq[(j − la − 1) mod l] = reference[j] do
        j ← j + 1
    end while
    return i, j
end procedure
```

Секвенирование – это не только GATC



Спасибо за внимание!

info@sequoiag.com