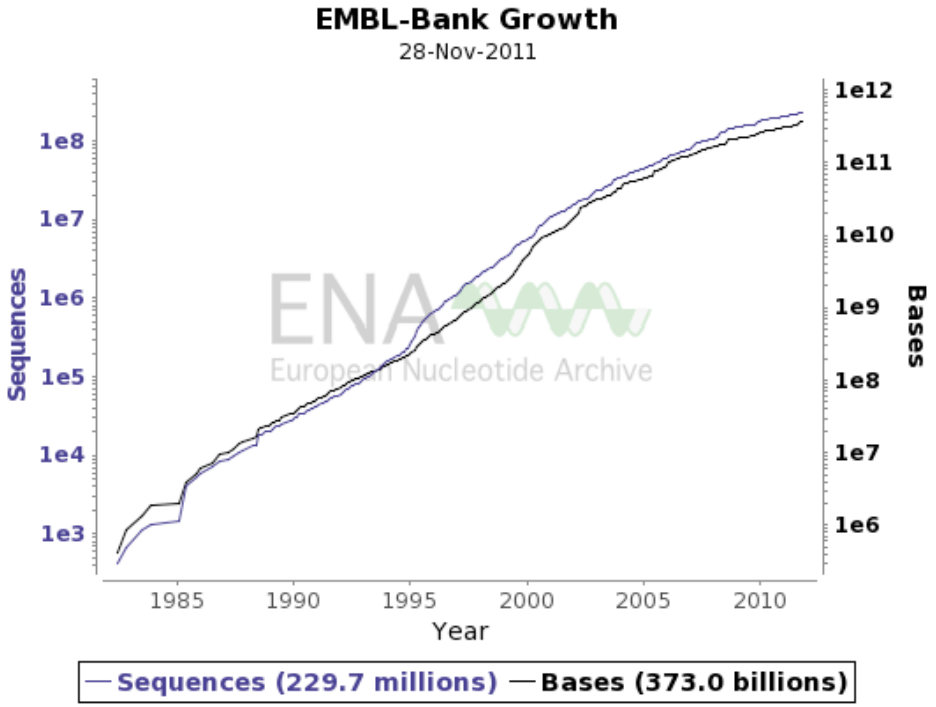
Big Data kihívások a genomikámban

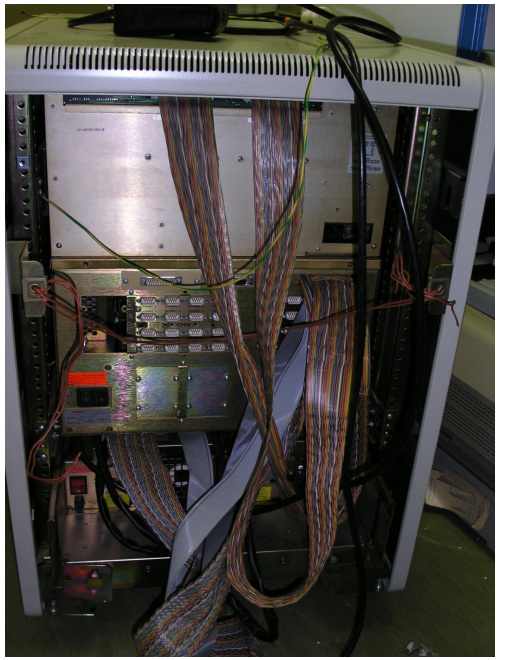
Barta Endre
DEOEC, BMBI, Klinikai Genomikai Központ
MBK, Mezőgazdasági Genomikai és Bioinformatikai Csoport
barta.endre@unideb.hu
barta@abc.hu

Adatbázisok és a tárolókapacitás növekedése (MBK vs EMBL)



- 1990: MicroVax szerver
  - 2x 160 Mbyte HDD – 50 mbp
- 1993: SUN SparcServer 1000
  - 8x 512 Mbyte HDD – 150 mbp
- 1997: SUN Ultra Enterprise II
  - 4x 9 Gbyte HDD – 1 gbp
- 2002: SUN Vire480
  - 8x 180 Gbyte HDD – 23 gbp

  Szekvencia + annotáció + index =~ 100 Gbyte
- 2007: LINUX szerver
  - 6x 500 Gbyte HDD – 100 gbp
- 2011: LINUX klaszter
  - 8x 600 és 24x2000 Gbyte HDD – 370 gbp



MBK külső hálózat

- 1991-ben 2400 bit/s modem Triesztbe
- 1992-ben 9600 bit/s X25 összeköttetés, HUNGARNET!
- 1993-ban 19.2 Kbit/s bérelt vonali összeköttetés
- 1994-ben 2 Mbit/s mikrohullámú összeköttetés
- 2000-ben 155 Mbit/s ATM összeköttetés (SZIE)
- 2002-ben 1Gbit/s optikai összeköttetés

Adat transzfer szekvenátor és a szerver között

1 Gbit

~1.4 Tbyte adat 11 nap futás alatt

Debreceni Egyetem Orvos-és Egészségtudományi Centrum
Klinikai Genomikai és Személyre Szabott Orvoslási Központ

Vezető: **Nagy László**
Laborvezető: **Bálint L. Bálint**
Szekvenálás: **Gyuris Tibor és Póliska Szilárd**
Szerver: **Horváth Attila és Barta ENdre**
Bioinformatika: **Nagy Gergely, Horváth Attila, Jónás Dávid és Barta Endre**

Prices:

Description	Qty	Unit Price	Discount	Total Price
Library preparation for HiSeq, Genomic Shotgun	1	800.00	10%	720.00
High-throughput sequencing on the illumina HiSeq, Paired-reads channel 2x100 bp	8	6,468.00	30%	36,220.80
Data delivery on external hard drive **(OFFERED)** (up to 500 GB, including handling and shipping)	1	250.00	100%	0.00

Sub-total:	**36,940.80**
Swiss VAT: EXPORT	0.00
TOTAL (CHF)	**36,940.80**

Prices are in Swiss Francs. In Switzerland, VAT of 8% is charged.
Fasteris VAT number: 579 029 or CHE-110.264.966 TVA.
By sending us samples you agree with our Terms and Conditions and you are placing an order.

Genomika, funkcionális genomika

- **Genomika**: A teljes örökítő anyag (magi, mitokondriális és ha van a kloroplasztisz DNS) szekvenálása, a szekvencia annotálása, elemzése
- **Funkcionális genomika**: A gének összességének, vagy egy-egy csoportjának a vizsgálata nagy áteresztőképességű kísérletekkel, amelyek vagy a genomszekvencia alapján lettek megtervezve (microarray, SNP), vagy szekvenálást használnak (ChIP-seq, RNA-seq stb.)

http://genomika.blog.hu/

Új generációs szekvenálás első publikáció 2005.

•A főbb modellszervezeteket már megszekvenálták hagyományos elsőgenerációs térképezős módszerekkel
•Különböző genomprogramok folyamatban

Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[2], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[3], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[1], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

454 Life Sciences Corporation

Nature 437 (7057)
376-380

Moore törvény, Moore genom

- Újabb és újabb technológiák
  - Complete Genomics (hibridizálás)
  - Pacific Biosciences, Helicos egy molekula szekvenálása
  - Ion torrent: microchip alapú szekvenálás
  - Nanopore technológia

Los Angeles Times

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH LIVING TRAVEL OPINION Search GO

BREAKING CRIME L.A. APPS WEATHER TRAFFIC OBITS COMMUNITY CROSSWORDS COMICS

YOU ARE HERE: LAT Home → Collections → Chip

ADS BY GOOGLE

Need Next Gen Sequencing?
We're an Experienced Sequencing Lab that can Help Advance Your Research
www.ambrygen.com/NextGenSequencing

NextGen Sequencing
Ion Torrent PGM Sequencing Service Low Introductory Rate, Try Now!
www.EpochLifeScience.com

BOOSTER SHOTS: Oddities, musings and news from the health world

Cheaper DNA sequencing possible with computer chip technology

Recommend 1 0 Tweet Submit 0 +1

July 22, 2011 | By Marissa Cevallos, HealthKey / For the Booster Shots blog

The race to decode a person's genome on the cheap got tighter this week. The sequencing company Ion Torrent announced this week in Nature that it used a $49,500 machine, based on computer chip technology, to unravel a full human genome–aptly, using the DNA of Gordon Moore, co-founder of Intel.

ADS BY GOOGLE

inside HPC

Big Data in Healthcare – Saving Lives with Personalized Medicine

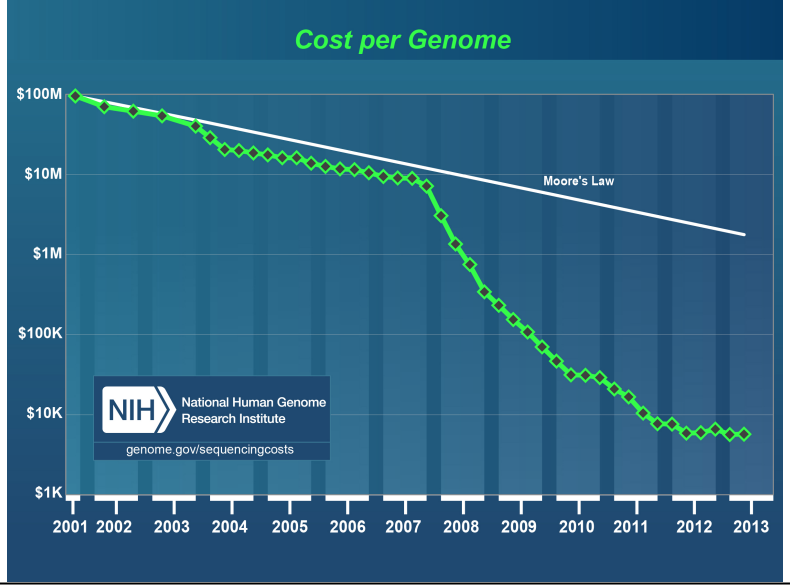
July 28, 2013 by Rich Brueckner

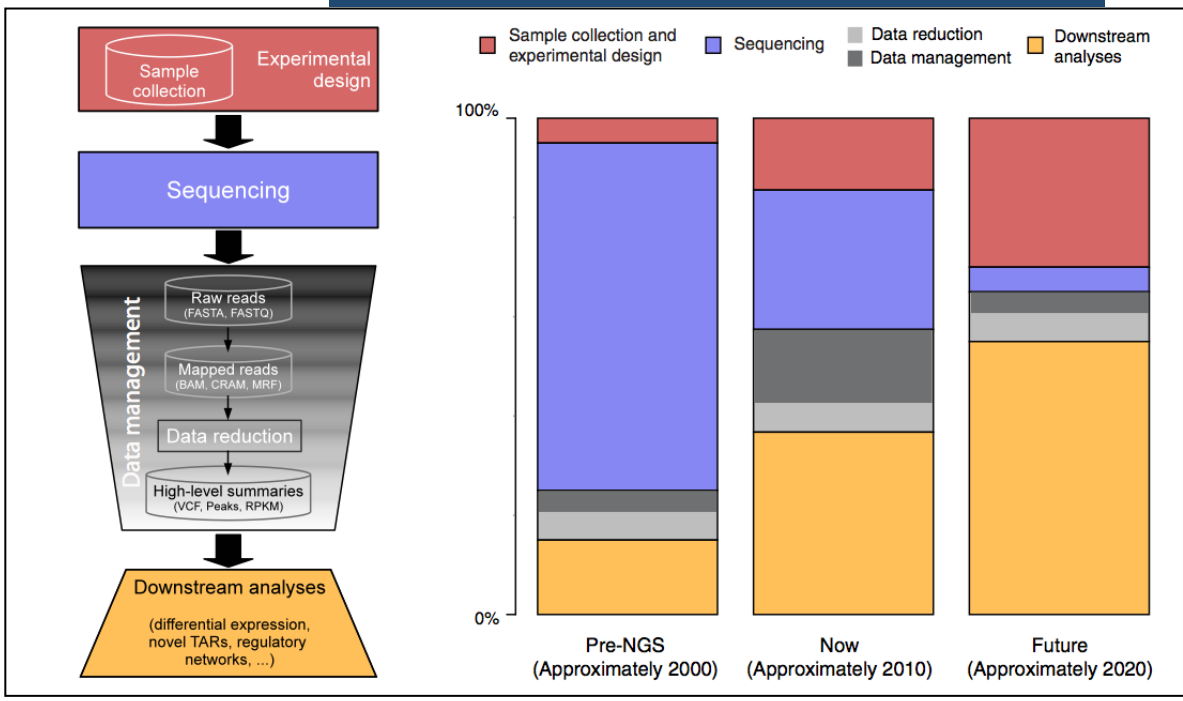
In this video, Intel Fellow Eric Dishman shares a personal story of how his 25 year struggle with kidney cancer was finally resolved through Big Data and personalized medicine. Dishman's doctors were able to treat him successfully after sequencing his complete genome.

For personalized medicine, only 50,000 people on earth have had their entire genome sequenced. As Dishman describes, this used to take months of supercomputing time and was simply out of reach for most people. This is starting to change thanks in part to advances in computing powered by Moore's Law.

Paradigmaváltás a kutatásban

- A „hagyományos” kutatási módszer hipotézis alapú
- A genomika lehetővé teszi az adatalapú kutatásokat
  - A kísérlet megtervezése a fenotípus alapján
  - Adatgenerálás
  - Adatfeldolgozás, szisztematikus analízis
  - Új hipotézisek felállítása
  - Kísérleti validáció

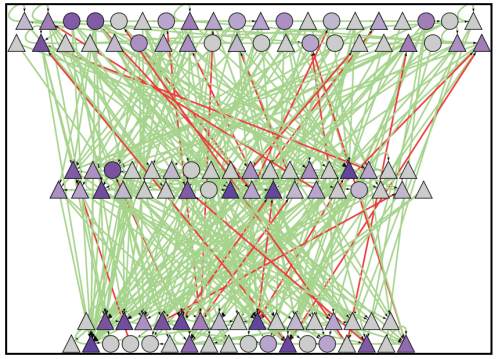


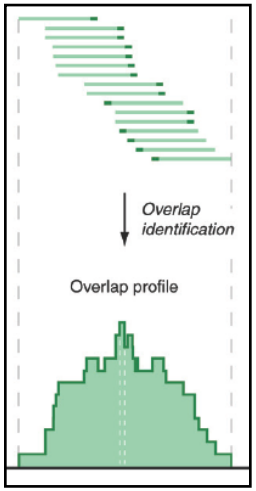


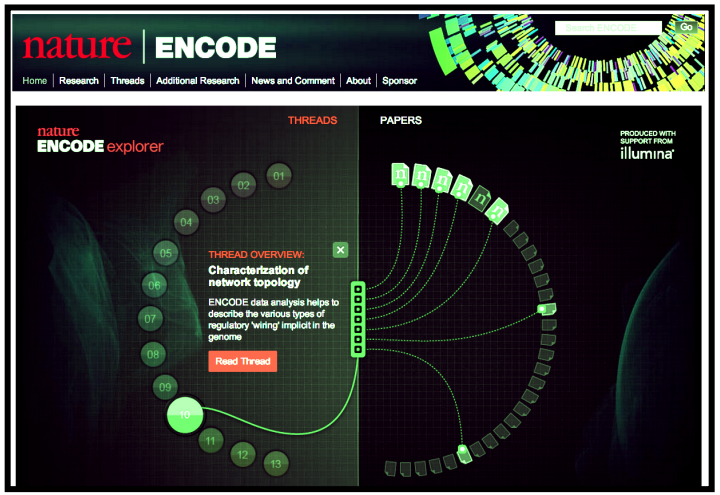
[Sboner et al. (‘11) GenomeBiology]

Szekvenálás megjelenítése genomböngészőben

Hierarchikus ENCODE adatmegjelenítés

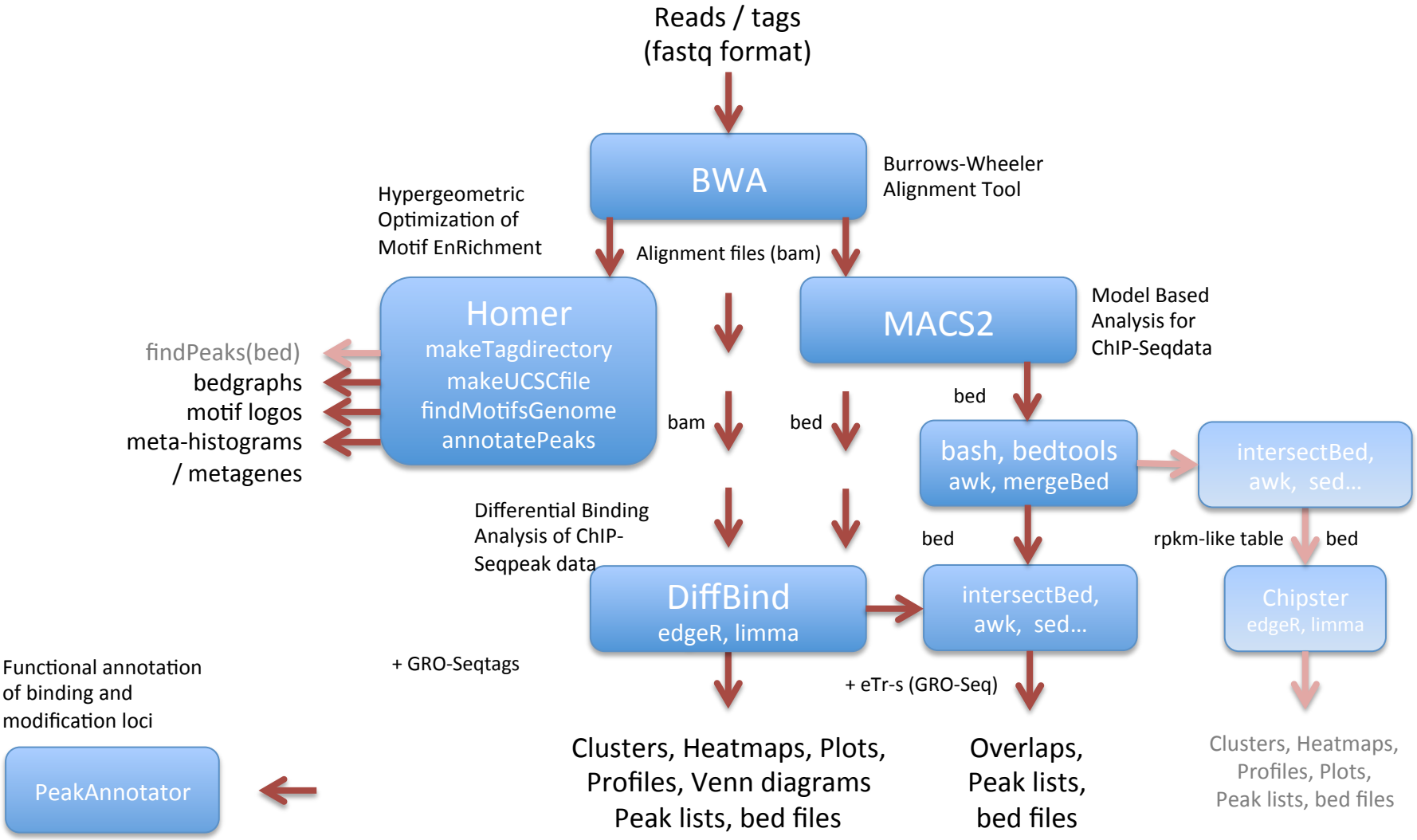






[PLOS CB 4:e1000158;
Nature.com/encode + *Nature* 489: 208]

ChIP-Seq analízis pipeline



Barta E: Command line analysis of ChIP-seq results. EMBNET JOURNAL 17:(1) pp. 13-17. (2011)

Big Data kihívások a genomikában

- Elsődleges szekvenálási adatok tárolása és feldolgozása
- Szekvencia archívumok (nincs adatmentés, párhuzamos adatbázisok vannak)
- Adatok feldolgozása 10-100x-os tárterületet igényel
- Szekvenálás berobbanása a humán diagnosztikába
- Adatok vizualizálása
- Speciális hardverigény (óriási I/O igény, esetenként nagy memóriaigény)

Genomika és Big Data a kínai BGI-ban

Table 1. BGI Computing Platform

Location	Cores	Memory	Storage	T flops
Shenzhen	11,000	22TB	8.88PB	117T flops
Hong Kong	7,776	9.7TB	8.515PB	83T flops
Beijing	300	500GB	162TB	1.5T flops
Wuhan	1760	500GB	1PB	4T flops
US-CHOPS	300	800GB	500TB	2T flops
US-UC Davis	300	800GB	500TB	2T flops
EUR-Denmark	600	1.6TB	1PB	4T flops
Total	22,036	35.9TB	20.6PB	214T flops

Project

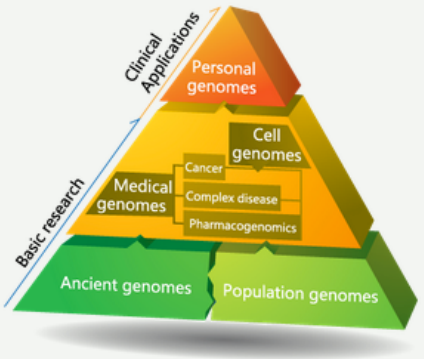
- Million Plant & Animal Genomes Project
- Million Human Genomes Project
- Million Micro-ecosystem Genomes Project
- Article Published

Industry

- BGI Tech
- Clinical Analysis Center
- Reproductive Health Center
- Agriculture Service

Contact Us Investors Relationship Site map

Project Overview



Gímszarvas genomszekvenálása

- Nincsen referencia szekvencia (amihez illeszteni lehetne)
- 4 lane HiSeq 2000-es paired-end és 2 lane mate pair szekvenálás (~4 millió forint)
- 1.1 milliárd 2x100 bp hosszú read -> 220 milliárd bázispár -> ~80x lefedettség (elméletileg a genom minden pontjára 80 leolvasás jut)
- denovo genomösszerakás a pécsi szuperszámítógépen

NIIF szuperszámítógépek, Pécs

- **SGI UltraViolet1000** típusú, SMP (ccNUMA)
- **Intel Xeon X7542**(Nehalem EX) típusú 6 magos processzorok találhatóak, **összesen 1152 mag (96 gép)**
- **6 Tbyte memória**- Bármely processz egyben látja, meg tudja címezni!
- **0,5 Pbyte diszkr**endszer

- **Suse enterprise operációs rendszer,**
- **SGE feladatütemező**

denovo szekvencia összerakás Big Data vonzatai

- 772Gb memória
- 2.8Tb diszk
- Az összerakás 3 hónapja folyik

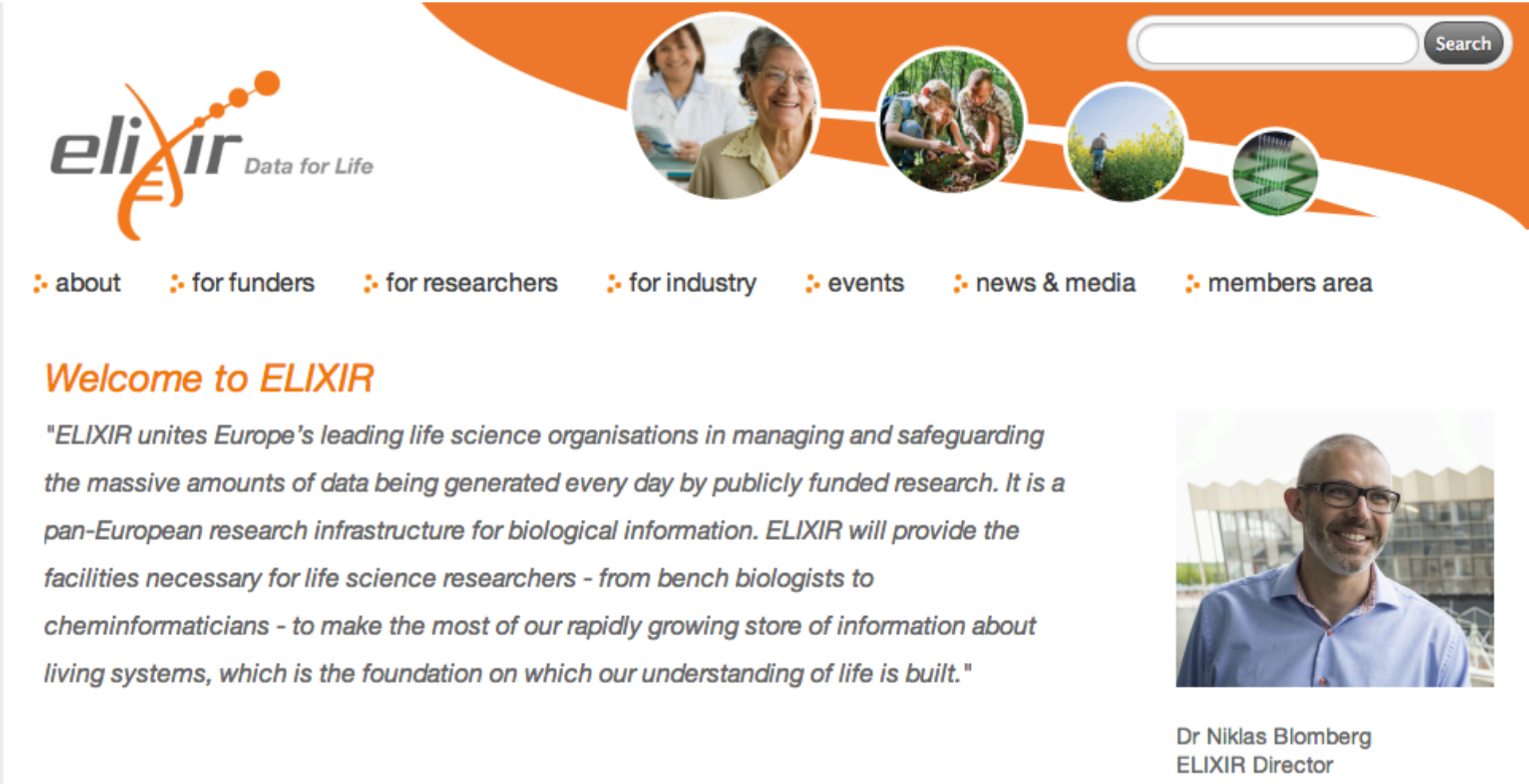
```
[vserver] szarvas (0)$ tail -f ./logs/RunAllPathsLG_build1_8.log
sze szept 11 21:26:25 2013: building extenders_q_jump, memory usage = 61,544,034,304
cs szept 12 07:29:14 2013: now memory usage = 149,702,168,576
cs szept 12 07:29:52 2013: found extenders, memory usage = 45,580,173,312
cs szept 12 07:55:16 2013: reloaded extending data, memory usage = 298,075,394,048
cs szept 12 08:45:14 2013: join data created, memory usage = 17,702,809,600
cs szept 12 08:45:14 2013: considering 11157628 joins
cs szept 12 08:45:15 2013: forking 80 cottage workers to do 11157628 joins.
cs szept 12 08:45:58 2013: farming out work
Farming out 11157628 jobs to sub-processes.
.......... .......... .......... .......... ......
```

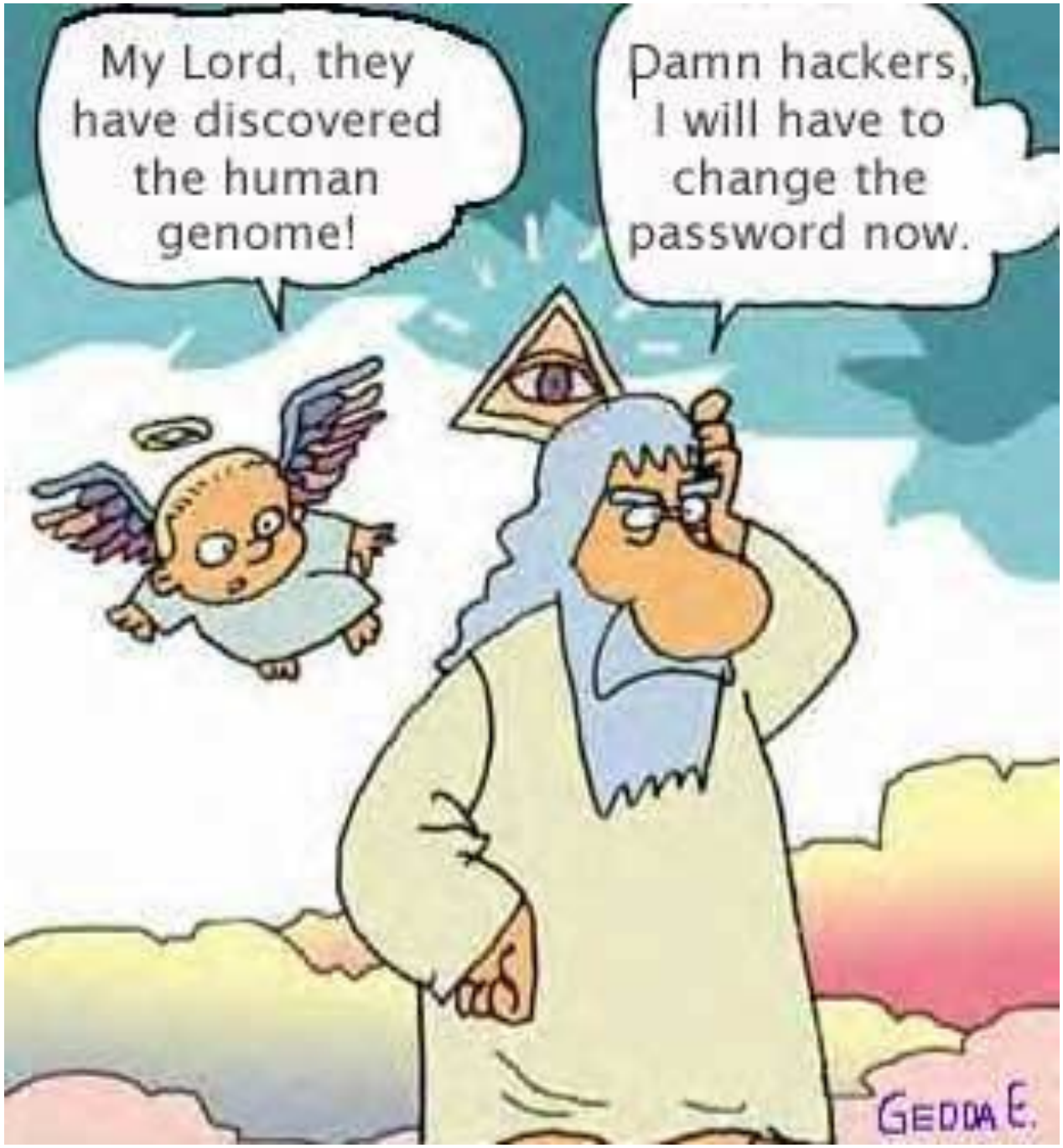
```
[vserver] top (0)$ cat ebarta.top.2013-08-17_22:42:52.out
top - 22:42:55 up 5 days,  6:45,  0 users,  load average: 607.30, 649.18, 642.81
Tasks: 20445 total, 502 running, 19943 sleeping,   0 stopped,   0 zombie
Cpu(s): 41.7%us,  4.4%sy,  0.0%ni, 53.9%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:   6187177M total,   909152M used,  5278025M free,      388M buffers
Swap:   285945M total,     2048M used,   283897M free,   116378M cached

    PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
1173780 ebarta    20   0  772g 713g 2068 S  663 11.8  13745:16 exe
 178350 ebarta    20   0 24032  15m  780 R   66  0.0   0:02.74 top
 178346 ebarta    20   0 12628 1564 1232 S    0  0.0   0:00.10 sh
 574656 ebarta    20   0 13016 1268 1264 S    0  0.0   0:00.02 57390
 574662 ebarta    20   0 24952 2232 2228 S    0  0.0   0:00.16 exe
 574663 ebarta    20   0  5632  636  632 S    0  0.0   0:00.00 tee
 574671 ebarta    20   0 10388 1440  856 S    0  0.0   0:00.11 make
1173779 ebarta    20   0 12632 1228 1224 S    0  0.0   0:00.00 sh
1173783 ebarta    20   0 22220 1656 1532 S    0  0.0  15:46.61 MemMonitor
1173784 ebarta    20   0  5632  664  632 S    0  0.0   0:00.03 tee

[vserver] top (0)$
```

```
[vserver] ebarta (0)$ uname -a
Linux vserver 2.6.32.59-0.7-default #1 SMP 2012-07-13 15:50:56 +0200 x86_64 x86_64 x86_64 GNU/Linux
[vserver] ebarta (0)$ pwd
/scratch/ebarta
[vserver] ebarta (0)$ du -hs szarvas/
2,8T    szarvas/
[vserver] ebarta (0)$
```

Összefoglalás

- A biológia, molekuláris biológia mindig is élen járt a számítástechnika alkalmazásában
- A genomika a közeljövőben az egyik legnagyobb Big Data generáló tudomány lesz
- Megoldás: Infrastruktúra fejlesztés (ELIXIR)





Köszönöm a figyelmet!