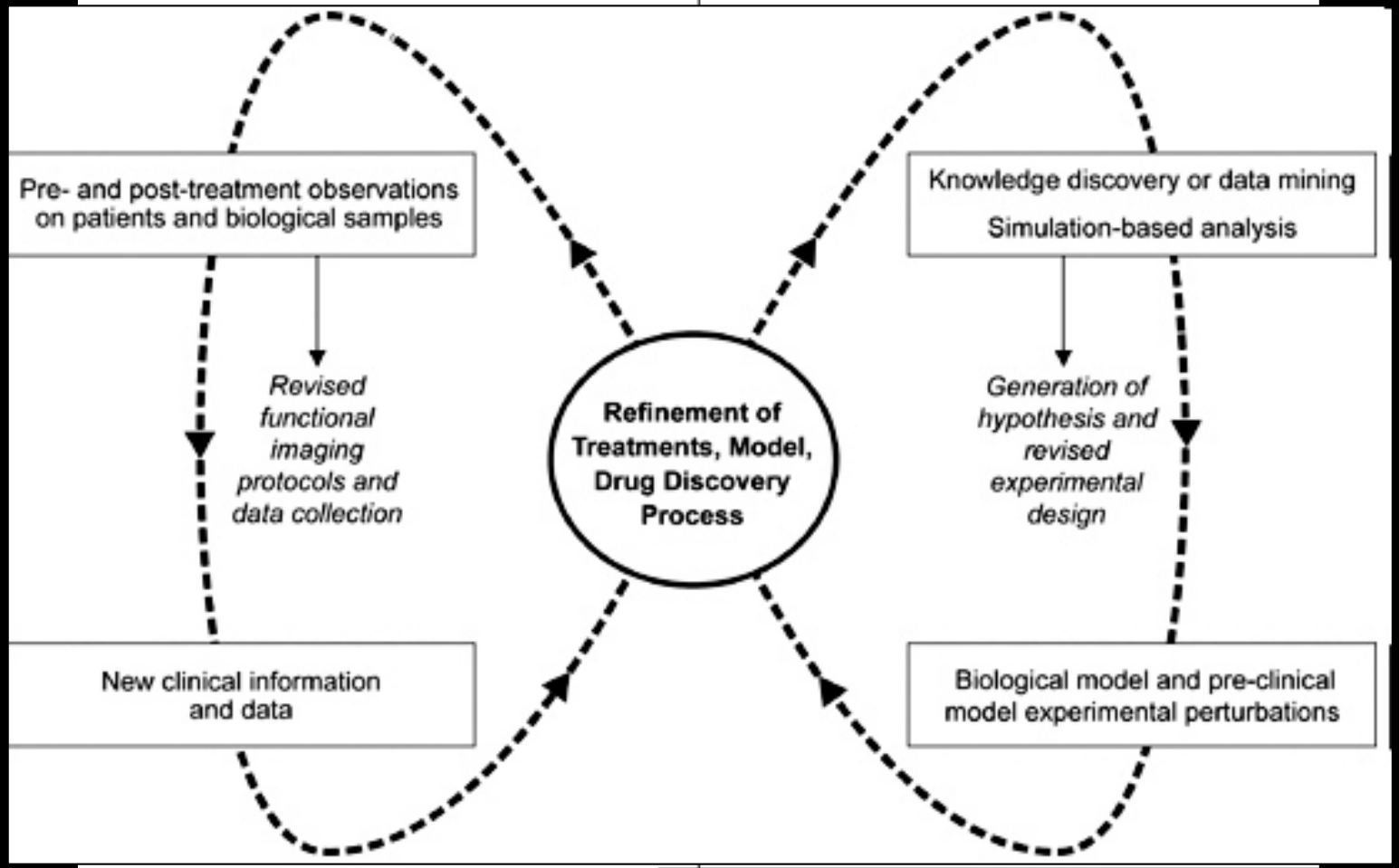
HPC and Translational Medicine

Arthur Thomas
Proteus Associates
July 1 2009

Translational Research

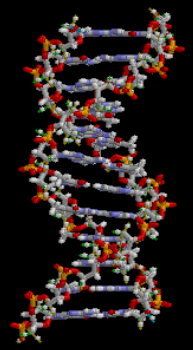


Source: A.C. Ahn et al "The Clinical Applications Of a Systems Approach," *PLOS Medicine* 3:956 (2006)

Source: J. Costa "Systems Medicine in Clinical Oncology" *Nature Clinical Practice Oncology* 5: 117 (2008)

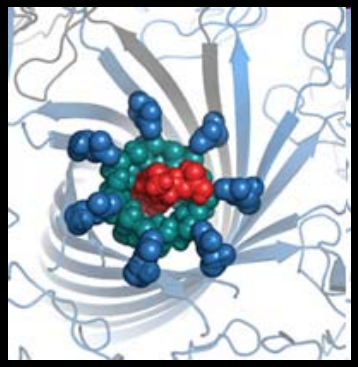
Translational Research

- Understanding disease at the molecular, cellular and organ level
- Diagnosisw
  - Genetic profiling for early detection, mitigation and (eventually?) prevention
  - New biomarkers for diagnosis/staging
  - Disease sub-typing
- Treatment
  - Drug discovery and development
  - Pharmacogenomics (personalized medicine)
  - New biomarkers (therapeutic monitoring)
- Translational medicine
  - Moving beyond “evidence-based”
  - Moving faster from research to patient care
  - Collaboration across traditional disciplines
  - Integration of information across traditional “silos”
- “P4” medicine (Lee Hood)
  - Predictive
  - Personalised
  - Preventative
  - Participatory



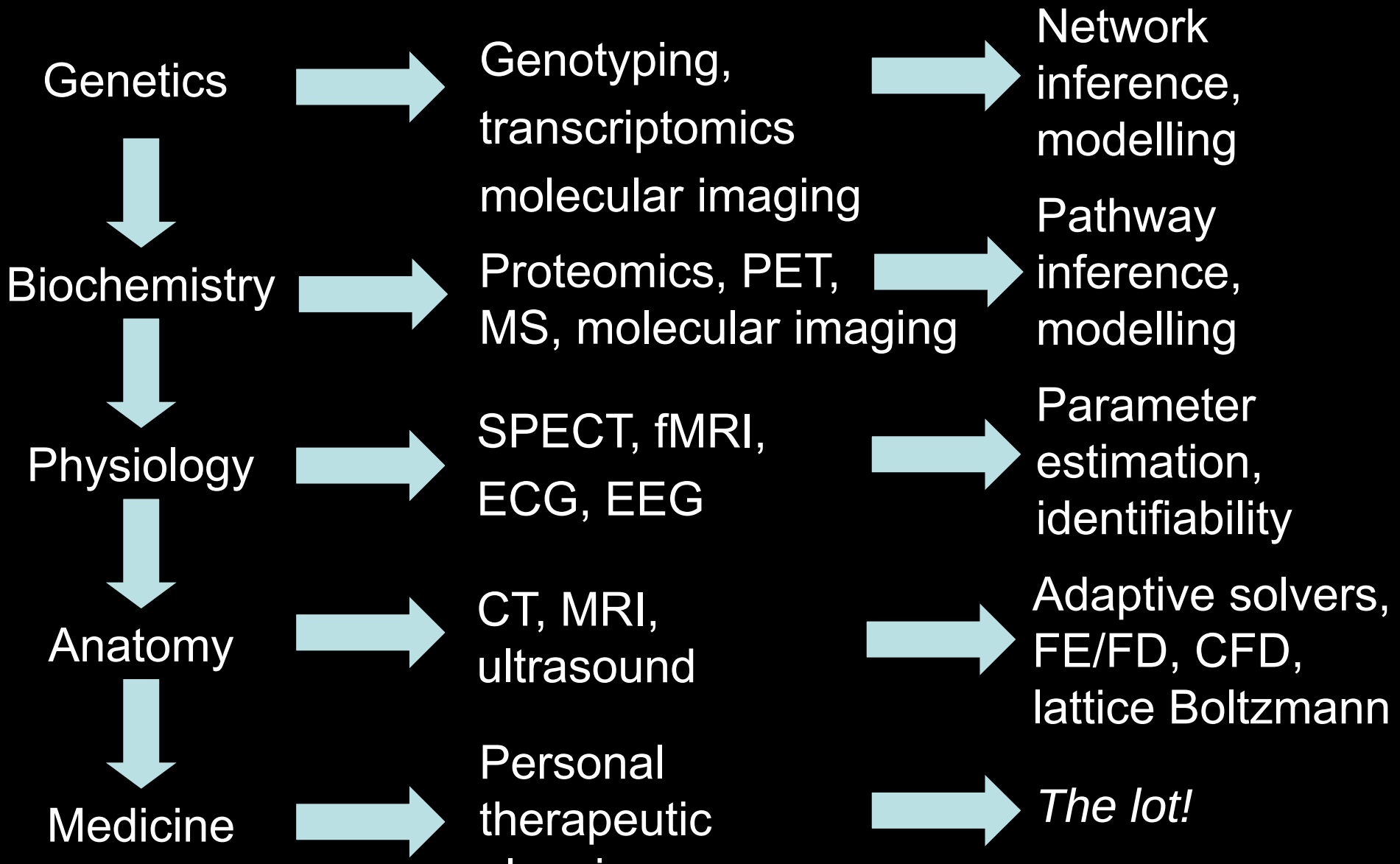
Challenges

- High throughput technologies
  - Genotypes
    - From PCR to single molecule
    - Throughput up from 2x/year to 10x/year
    - $100k/genome => $1k/genome
  - Phenotypes
    - Proteomics: entire serum profiles using MS, antibody arrays
    - Metabolomics: MS
    - Imaging: molecular, functional, anatomical
    - Classical: morphological, cognitive, psychological
- Modelling
    - Molecular dynamics and drug interactions
    - Network/pathway models
    - Multi-scale, multi-physics models of physiology
- Data Integration
  - Naming and ontological issues
  - Security and privacy



Source: Oxford Nanopore Technologies

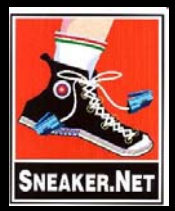
Domains and Methods



High-throughput genotyping

Genotyping: Big Data!

wellcome trust sanger institute



Proposed Data Íslandia archive

Sequence data:
- 115k TIFF images/run (Illumina GA II)
- *80 TB/week/sequencer raw*
- 50 GB/week/sequencer processed
- 40 sequencers = 2TB/week processed
- *2 PB/year in 2008*
- *3-4 PB/year in 2010*
- Trace archive doubling time: 11 months

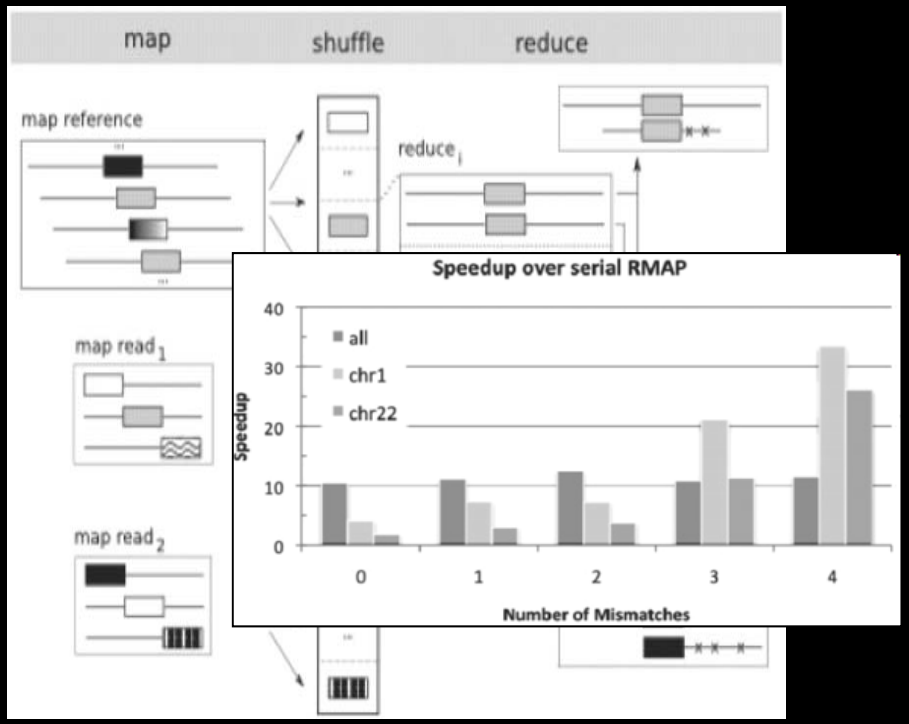
Challenges:
- Storage/backup
- Computation (analysis & modelling)
- Comms Networks (private lambdas)

ELIXIR EUROPEAN LIFE SCIENCES INFRASTRUCTURE FOR BIOLOGICAL INFORMATION

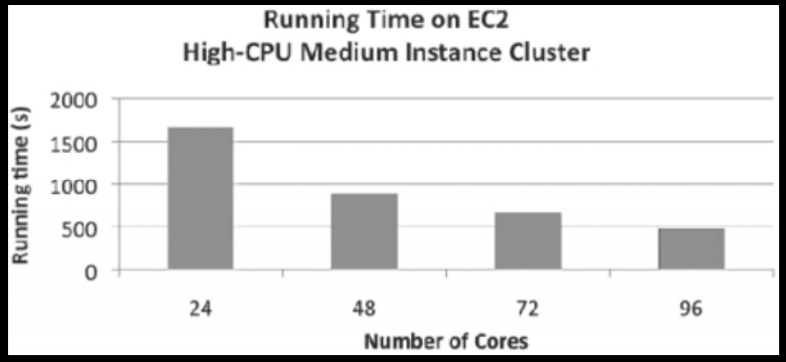
- Preparation: 2007-2010 4.5 M€
- Construction: 2011-2018 470 M€
- Operation: ~100 M€/year

Move from individual to population data:
- *1000 Genomes project*
- *Intl. Cancer Atlas*
- *> 25,000 human genomes within 3 years*

CloudBurst: sequence read mapping with *MapReduce*

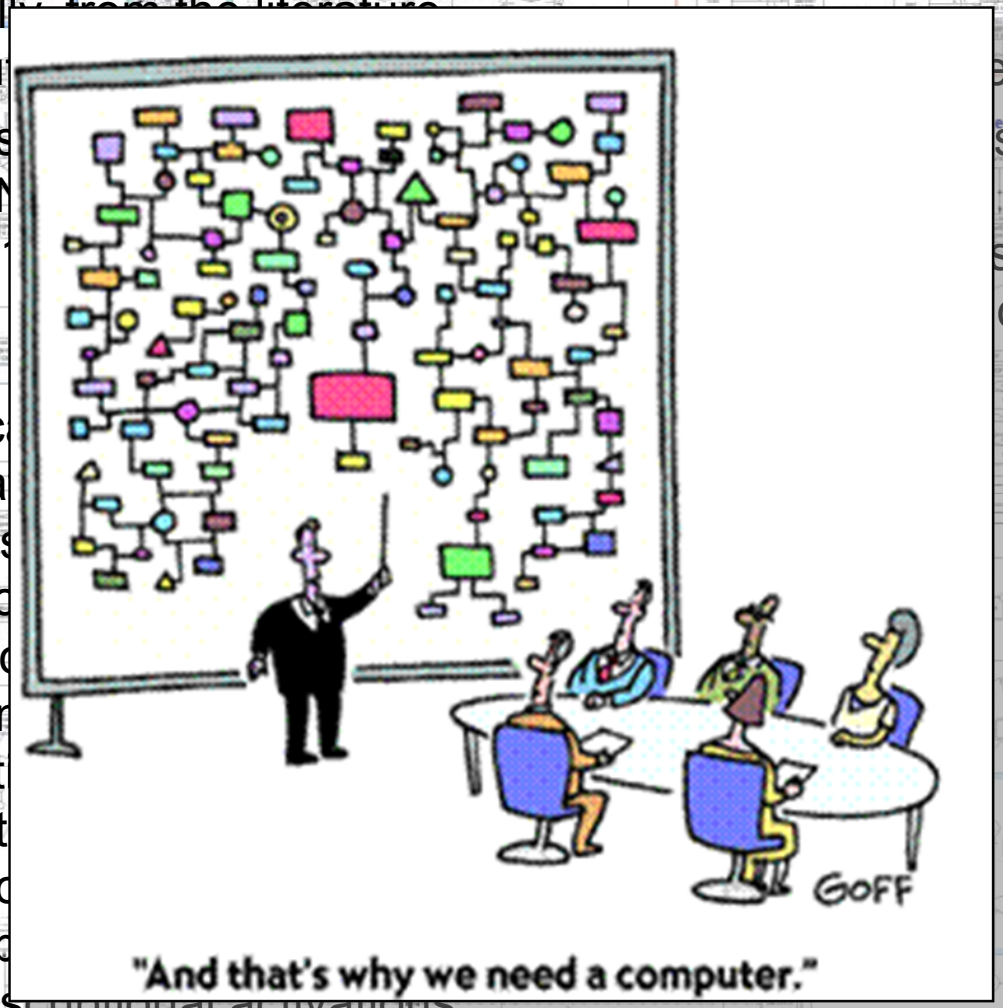


- Map short (25-250bp) reads to a reference genome
- Used for SNP discovery, genotyping, gene expression, comparative genomics and personal genomics
- Up to 4 billion reads/experiment
- *Hadoop* on local 24-core local cluster and 24-96 core Amazon EC2 “High core medium AMIs”
- Near linear in number of cores and number of reads



M.C. Schatz *Bioinformatics* 25:1363-1369 (2009)

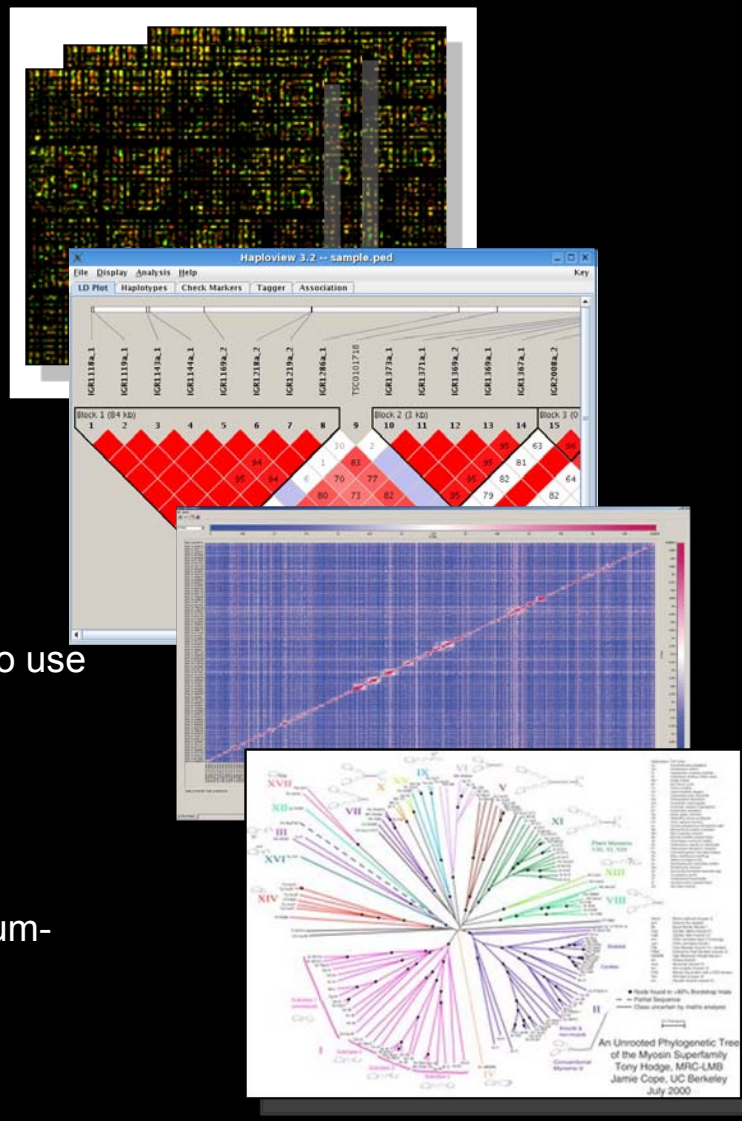
- Pathway inference:
  - Manuall[illegible] from the literature
  - Comput[illegible]ein interaction data
- 202 proteins[illegible]s, seven genes, and seven RN[illegible]
- 10 ligands, 1[illegible]s), three ion channels, 10 [illegible]d 22 adaptor proteins.
- Reactions c[illegible]
  - 131 sta[illegible]
  - 34 trans[illegible]
  - 32 assoc[illegible]
  - 11 disso[illegible]
  - two trun[illegible]
- 247 interact[illegible]
  - 206 cat[illegible]
  - 9 unkno[illegible]
  - 16 inhib[illegible]
  - 12 transcriptional activations
  - 4 transcriptional inhibitions.



www.sbgn.org

Kitano et al *Molecular Systems Biology* (2005)

Genome-wise Association/ Linkage Analysis



Large-scale analysis of microarray or deep sequencing data

- Search for gene variants that may cause disease
- Model dynamics of gene evolution
- Deduce evolutionary history by comparing species

Challenges:

- 500-1000k SNPs/chip; 1000-10,000s of individuals
- E.g. PLINK (Broad/MGH):

Compare allele frequencies between cases and controls. Also use
- Cochran-Armitage trend test
- Fisher's exact test
- different genetic models (dominant, recessive and general)
- tests for stratified samples (e.g. Cochran-Mantel-Haenszel, Breslow-Day tests)
- multilocus tests, using either Hotelling's T(2) statistic or a sum-statistic approach

Memory-intensive (TB of raw data)
Visualisation of very large datasets difficult
Move from binary to graded associations and CNV, will scale problem 10x

GWAS: Epistasis testing

- N(2,3)-way interactions between gene variants

Table 1: Estimated single-processor computing time o processor, and the total number of tests for two-locus

Number of SNPs (N)	
500,000	Computing time (T) Number of tests (M)
1,000,000	Computing time (T) Number of tests (M)

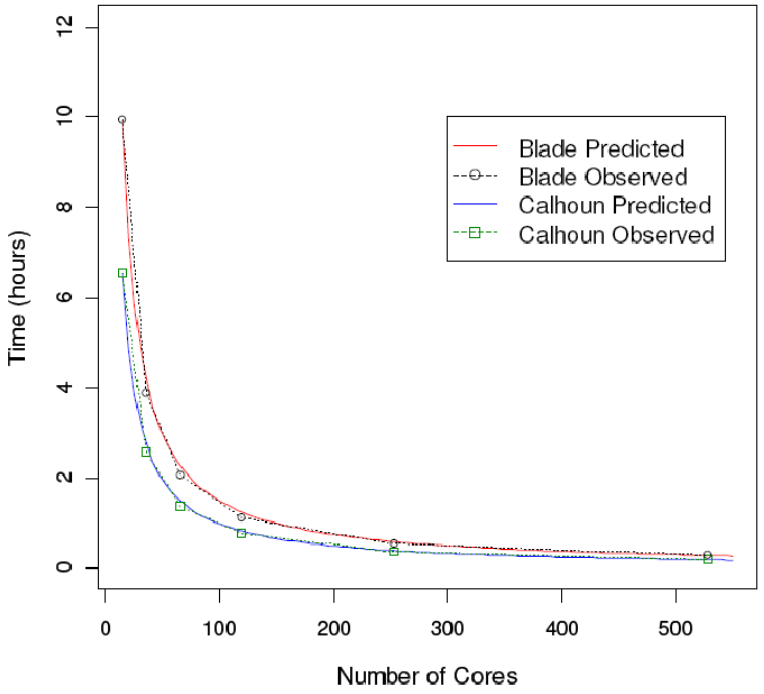
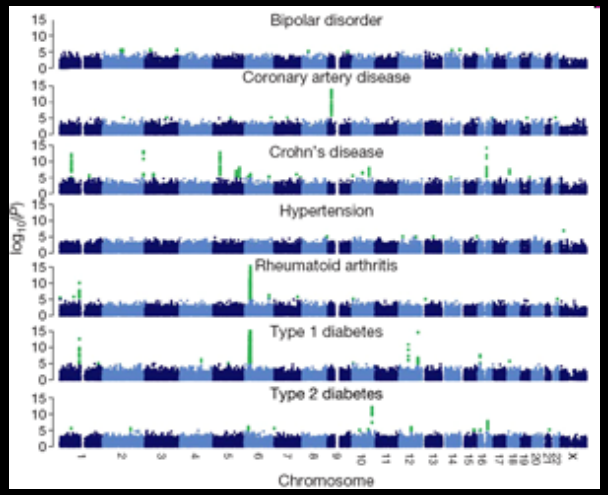


Figure 1
Observed and predicted run times of the EPISNPmpi program on Minnesota Supercomputing Institute's 2.6 GHz IBM BladeCenter Linux cluster (Blade) and the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor (Calhoun). The observed run times (circles representing Blade and squares representing Calhoun) matched well with the predicted run times under ideal speedup and scalability (solid line representing Blade and dotted line representing Calhoun). Analyses in this figure used a hypothetical GWAS data set with 50,000 SNPs and 2000 individuals.

Pairwise comparison of 10^6 SNPs on 2048 core cluster: 20 hours

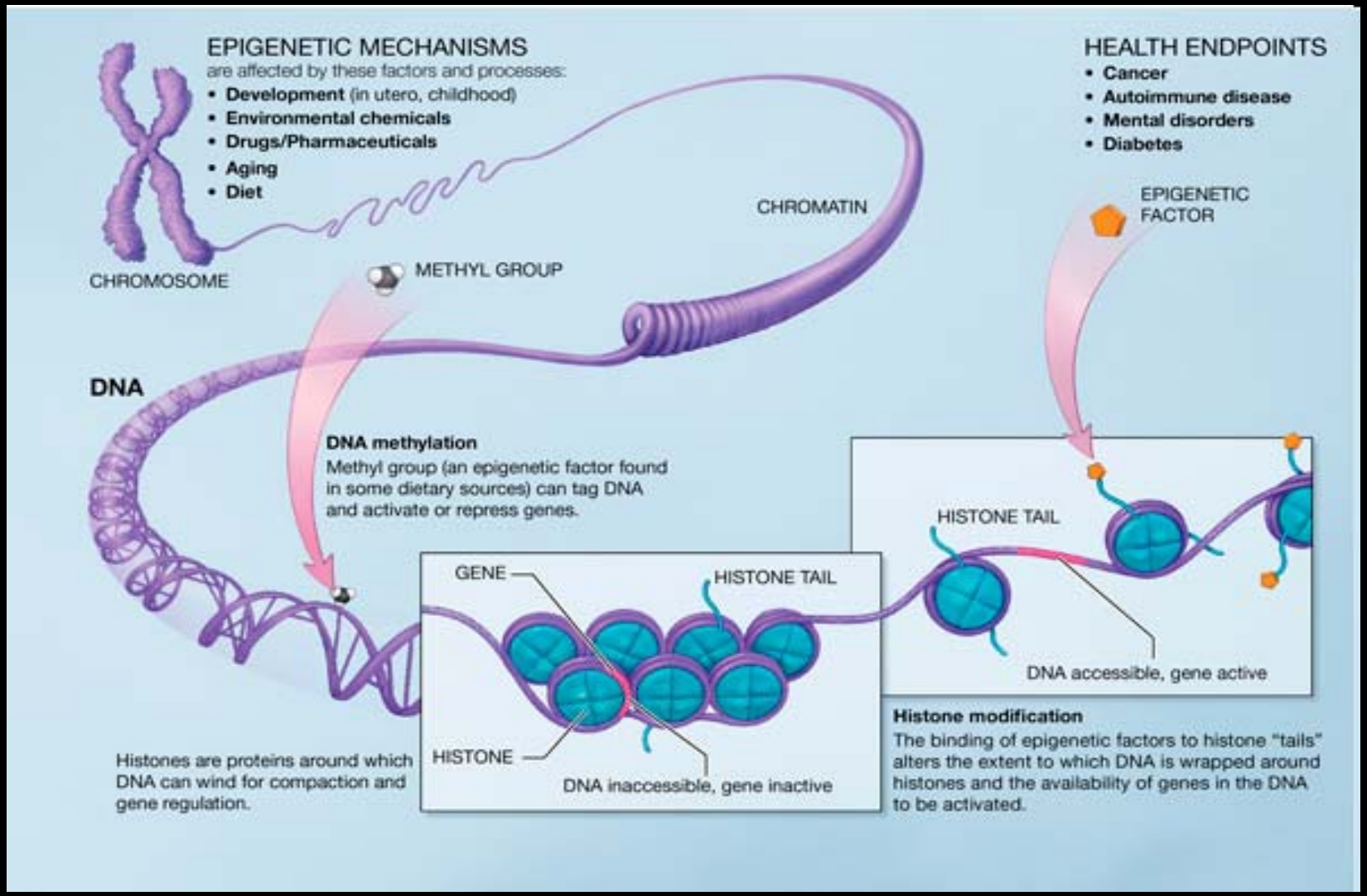
Source: Ma et al BMC Informatics 2008

Wellcome Trust Case Control Consortium



- WTCCC1 (2005-2007)
  - 50 studies, 19,000 samples, 7 diseases
  - Affymetrix 500k chip and Infinium 15k SNP chip
  - Identified 90 new significant gene variants; also studying CNV
- WTCCC2 (2008-)
  - 27 studies, 120,000 samples, 13 diseases
  - Affymetrix v6.0 or Illumina 660k chip

Oh! But…



(1) Feytmans, Noble & Peitsch, Trans. Comp. Sys. Biol. 2005 Source: NIH Roadmap

High-throughput phenotyping

Proteomics Analysis

- Data from 38 mass spectrometers
- 6908 files (256 Gb)
- 2.4 million MS/MS peak lists.
- Analysis time: 10.75 days on Swiss Multi Science Computing Grid
- 5.43% tasks were “lost” during computation
- 241,637 peptide matches
- 1,350 identified proteins
  - 290 “secreted”
  - 225 cytoplasmic
  - 321 membrane
  - 225 nuclear
- Just the beginning!

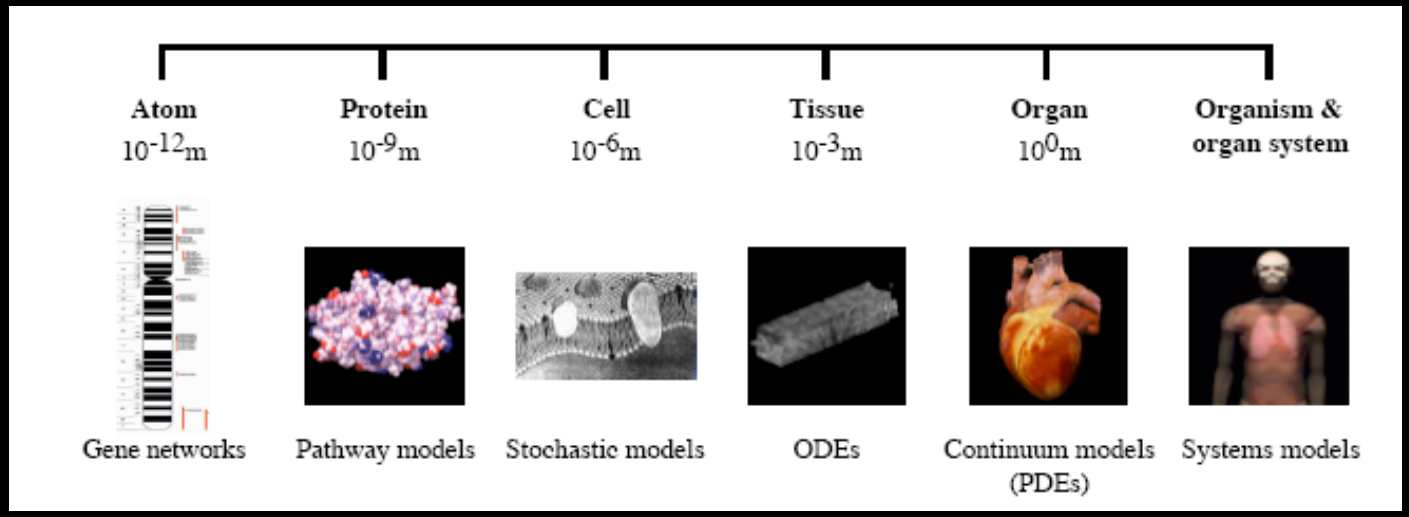
Figures and data courtesy of A. Quandt, SIB and ISMB/ETHZ

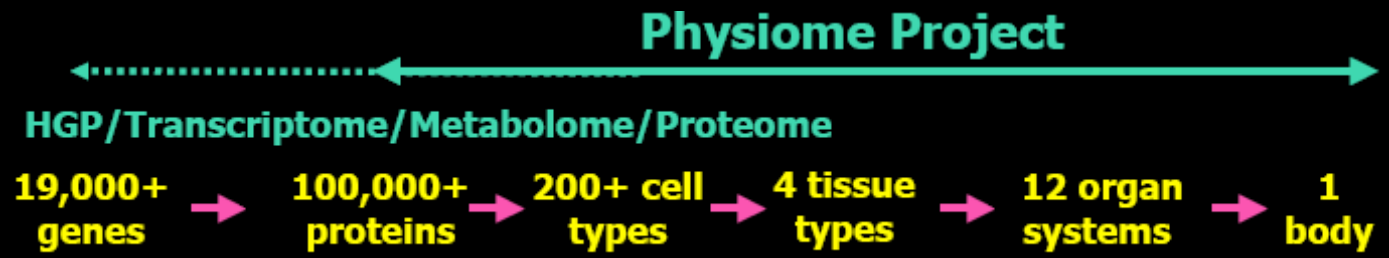
BioBanks: integrating genotypes and phenotypes

- e.g. UK Biobank
  - 500,000 participants
  - 20 million samples
  - SNP analysis and selective phenotyping
  - 250 million data points
  - 20 year timescale
  - Secure storage and communication
  - Oracle Heath Transaction Base (HL-7)
  - Tagged using ICD-9-CM + free text

Modelling

Multi-scale Modelling





Adapted from P. Hunter, P. Robbins, D. Noble "The IUPS human physiome project" *Eur J Physiol* 445:1–9 (2002) and P. Hunter "An Update on the Human Physiome Project," in *Proc. IUPS Satellite Workshop on Computational Physiology, San Diego, CA (2005)*

Virtual Physiological Human

- **Goals**
  - *Patient-specific* computational modelling and simulation
  - Data integration and new knowledge extraction demonstrated on clinical applications
  - Medical simulation for surgery
  - Prediction of disease or early diagnosis
  - Assessment of safety of drugs
- **Impacts**
  - New environments for *predictive*, *individualised*, more effective and safer healthcare.
  - Improved safety through simulation of adverse drug effects in *patient-specific models*
  - Improved semantic interoperability and knowledge management of biomedical information

VPH Projects

Acronym	Topic	Project type
VPH NoE	Networking	NoE
VPHOP	Osteoporosis	IP
euHeart	Heart/CV disease	IP
ARTreat	CV/Atheroschlerosis	IP
preDiCT	Heart/CV disease	STREP
ContraCancrum	Cancer	STREP
ARCH	Vascular/AVF & haemodialysis	STREP
PASSPORT	Liver/surgery	STREP
PredictAD	Alzheimers/BM & diagnosis	STREP
NeoMARK	Oral cancer/BM, D & T	STREP
VPH2	Heart/LVD surgery	STREP
IMPPACT	Liver cancer/RFA therapy	STREP
HAMAM	Breast cancer/diagnosis	STREP
Action-Grid	Grid access EU – LA & Balkans	CA
RADICAL	Security and privacy in VPH	CA

Source: P. Coveney, UCL

Computational Requirements

Project	Biggest Job	Average Job	Wall Time	# of Jobs P/A	Disk Space	Total CPU time required	Persistent Access	Other requirements
Seed Exemplar 4	100 core/1GB	32 core/1GB	24 hours	250	300GB	200000	Short bursts is fine	None
VPHOP (1)	10000 cores/100 GB RAM	5000 cores/24GB RAM	120 mins	20	1 TB	200,000 hours	Bursts, but not of duration inferior to the average job duration (120 min)	Planning a lot of MonteCarlo, need gang scheduling to ensure calculation threads are scheduled together. Emergency jobs are not planned for 2009; may be necessary in following years.
VPHOP (2)	15 cores/50 GB	5 cores/12 GB	500 mins	20	1 TB	10000 hours	Bursts, but not of duration inferior to the average job duration (500 min)	Planning a few MonteCarlo analyses, need gang scheduling to ensure that all calculation threads are scheduled together.
Action-Grid	2 core/4GB	2 core/2GB	20 mins	1000	100GB	500 hours	Persistent access	none
Seed Exemplar 3	NEURO 200 cpus RAM: 1GB ATHERO 100cpus RAM: 1GB	NEURO 100 cpus RAM: 1GB ATHERO 40 cpus RAM: 1GB	NEURO 20 hours, ATHERO 5 hours	NEURO 100-200, ATHERO 10000-20000 (can obtain more different model types)	NEURO 100GB, ATHERO 20GB	NEURO 200000 hours, ATHERO 50000 hours	It could be done in both ways. Maybe better in shorts bursts of time and then analyze data.	Not really any special one

Table 11: VPH-I computational resource requirements in the next 12 months.

Source: VPH Requirements and Technology Assessment Exercise (2008)

VPH Toolkit

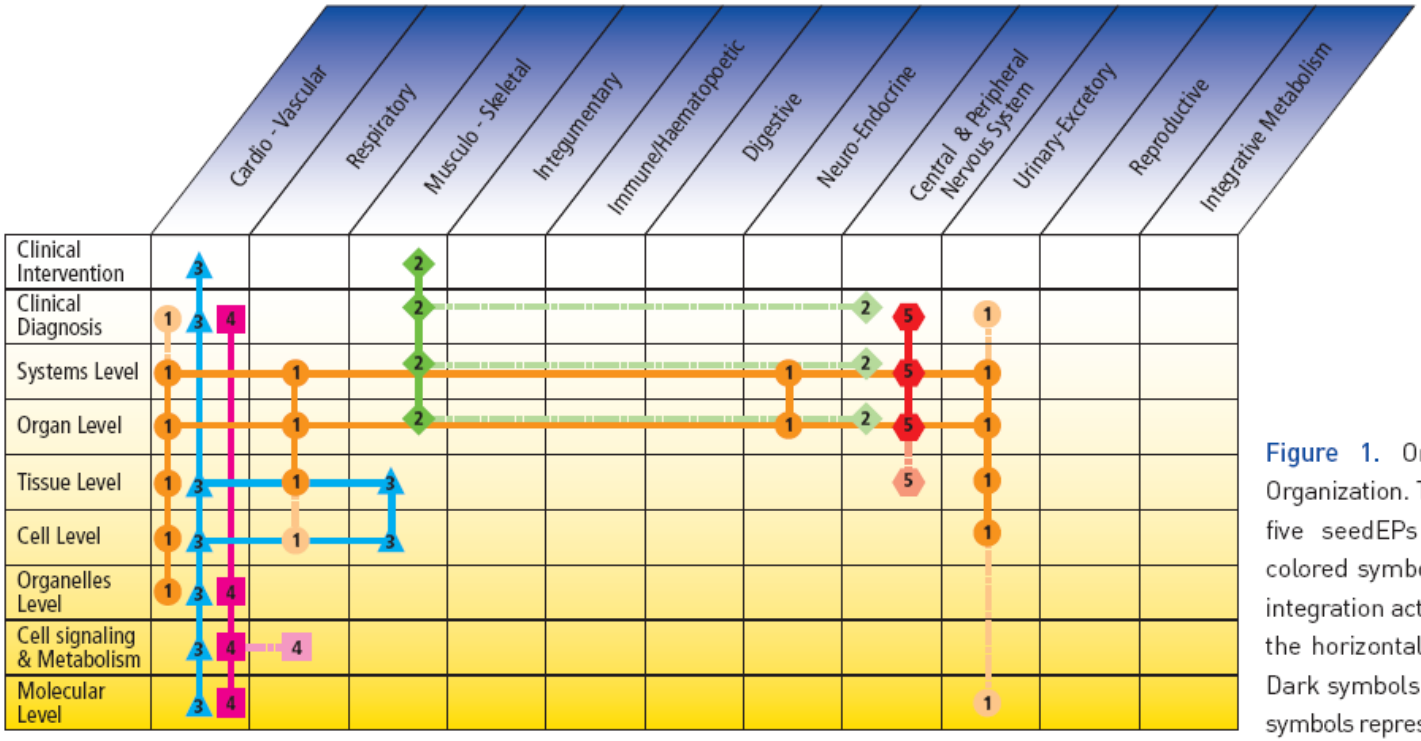


Figure 1. Organ Systems vs. Levels of Organization. The figure shows, for each of the five seedEPs (indicated by the numbered colored symbols), the vertical and horizontal integration activities. Note that emphasis is on the horizontal and vertical links themselves. Dark symbols indicate existing activities; pale symbols represent activities under development.

Source: VPH NoE

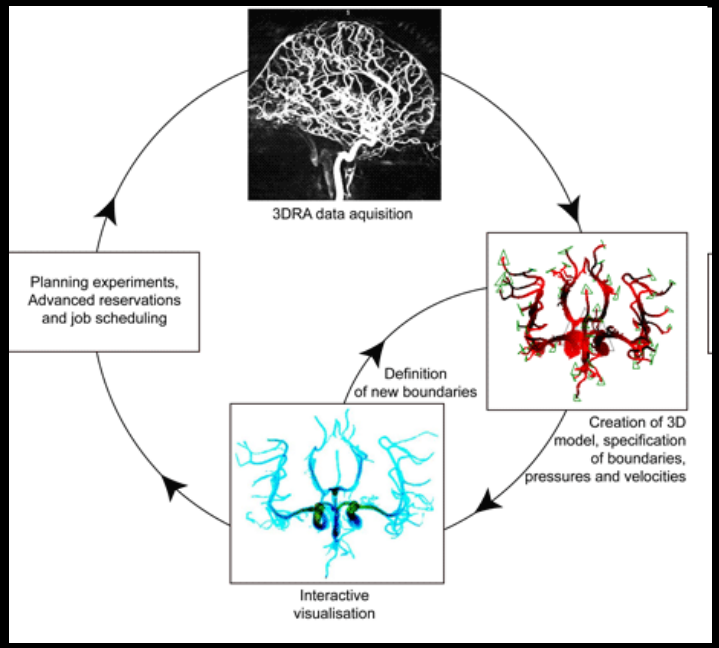
VPH Visualisation

Display wall: C. Goodyear, K. Brodlie, Leeds; data provided by P. Kohl, Oxford

- 11.7 T MRI, 25µm voxels
- 1576 1k x 1k TIFF images = 1.6 GB/study
- *VTK* Isosurfacing ->typical mesh of 45 million triangles
- 28 1600x1200 pixel displays (53 Mpixels)
- Driven by 14 *nVidia 7800* GPUs running *VRJuggler*

Personalised therapeutic planning

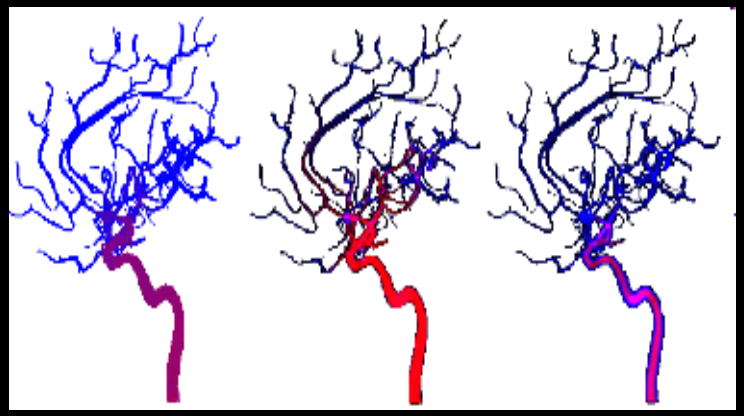
GENIUS: Grid Enabled Neurosurgical Imaging Using Simulation



- Patient-specific surgical planning for treating intra-cranial aneurysms (AVMs)
- 3-D volume models (10^9 200μm voxels) from MRI or CT
- New lattice Boltzmann code (optimized for grids) to model blood flow
- Real-time visualisation
- Simulation steered by clinician using *RealityGrid* steering library

- Real-time simulation doesn't fit well into traditional HPC job scheduling
  - Reservation
  - Preemption

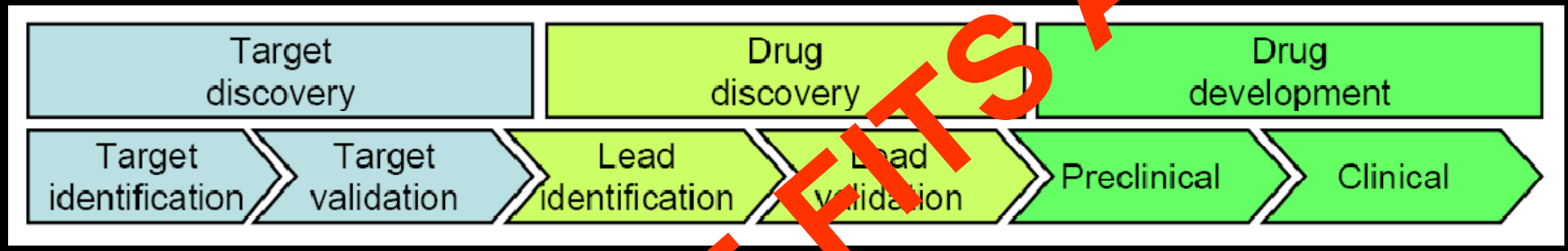
within TeraGrid, UK NGS and DEISA



Source: Peter Coveney, UCL

Drug Discovery & Development

12+ years, $1-1.25 billion



ONE SIZE FITS ALL

Sequence Homology, Gene Expression, Proteomics,

'Omics

System & Disease Modelling

Comb. Libraries

HTS

QSAR

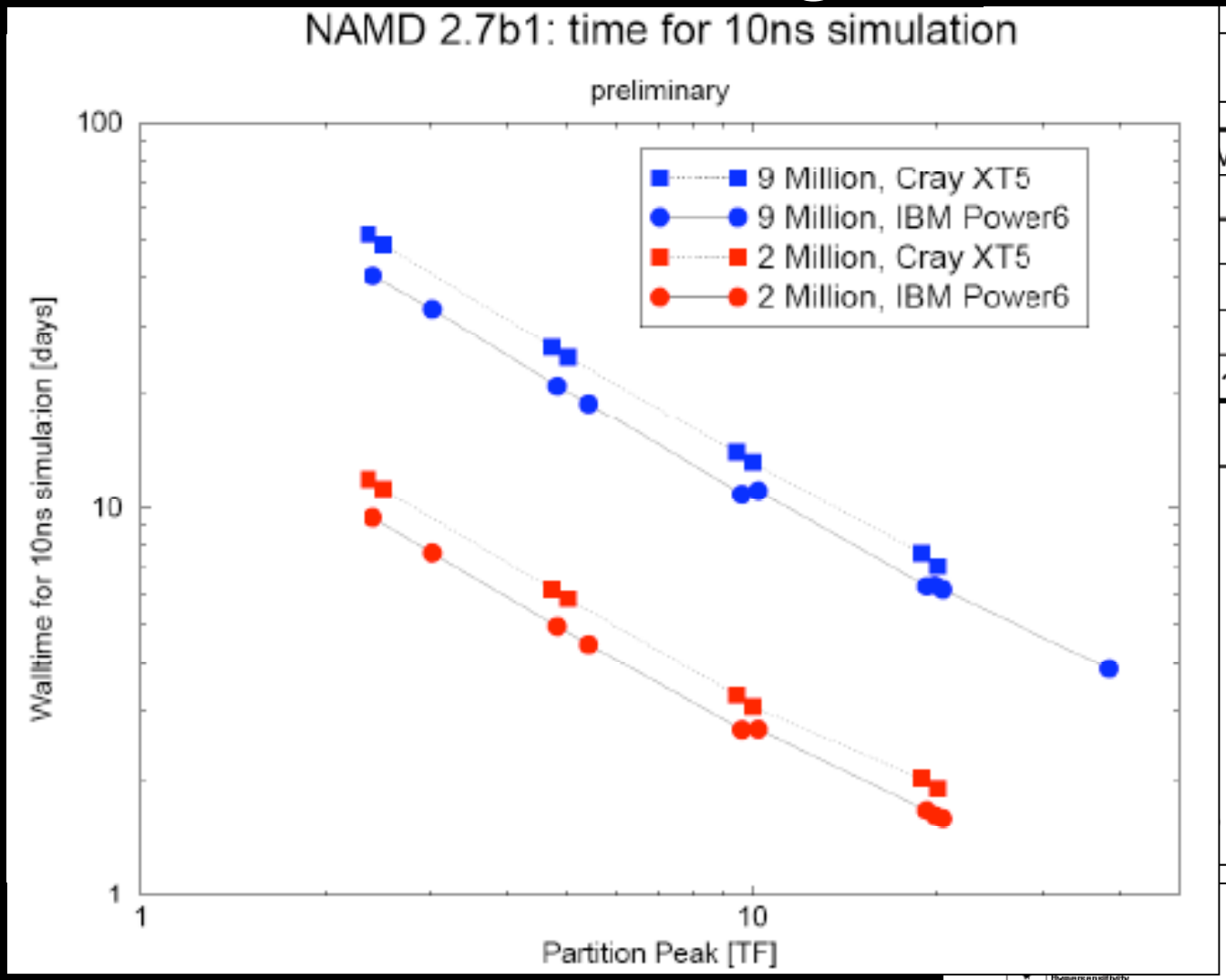
ADME/Tox

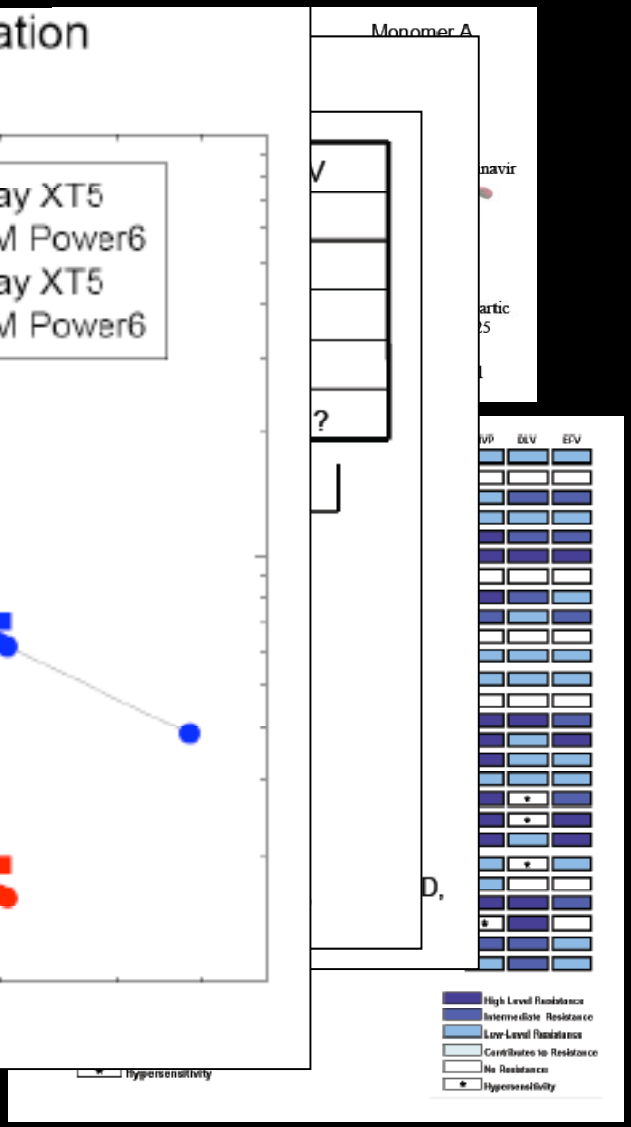
Trial Design

Paradigm Change

Old Science	*New Science*
Classical chemistry	**Combinatorial chemistry**
Basic biology	**'Omics, Biotechnology**
Experimentation	**Computation**
Low throughput	**High throughput**
Animal studies	**Molecular imaging**
Unidirectional	**Bidirectional**

Personalized Drug Treatments





Source: P. Coveney, UCL

One day…

High-throughput genotyping

High-throughput phenotyping

(Molecular, cellular, anatomical)

High-throughput physiology

Evidence-based medicine (static analysis)

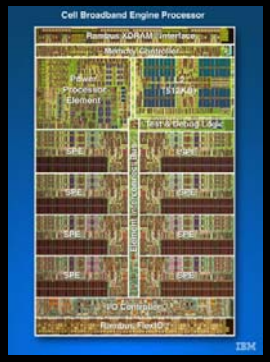
Computation-based medicine (dynamic analysis)

Personalised, predictive therapy

Thank you!

Accelerator Technologies

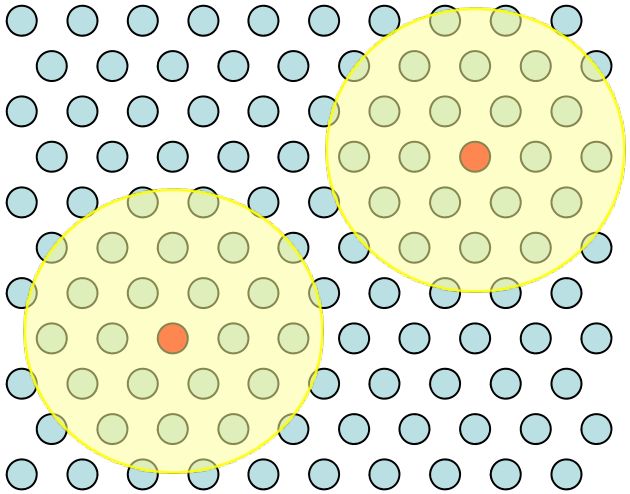
- ASIC-based
  - Paracel *GeneMatcher*®
  - NEC Molecular Dynamics Server
    - Port of CHARMM (M. Karplus, Harvard, T. Shuepbach, Vital-IT): 10-40x speedup
- FPGA-based
  - Time Logic *DeCypher*®
  - Mitrionics™ *Virtual Processor*
  - Progeniq *BioBoost*
  - XLBiosim
- GPU-based
  - NVidia GPU/Tesla
  - Intel Larrabee
  - AMD Vision
- CellBE-based (e.g. Sony PS3®)



Molecular Dynamics (MD) Modelling

- Calculate the realistic dynamics of massively multi-body systems
- Time evolution is deterministic and derived from Newton's equations of motion.
- 90% of CPU time spent computing the N^2 non-bonded interactions arising from Coulomb and Van der Waals potentials
- Good candidate for acceleration using special-purpose computing devices?
  - Computation vs data transfer tradeoffs

Molecular Dynamics: cutoff method



~1970-1980		~1995 on
Cutoff 2-4 Å	Increasing cutoffs	Cutoff 10-15 Å
No LR (long-range) interaction	No LR intereaction	LR interaction calculated with PME (particle-mesh Ewald) approximation

- Cutoff methods neglect interaction beyond a certain distance, thereby significantly reducing the number of interactions.

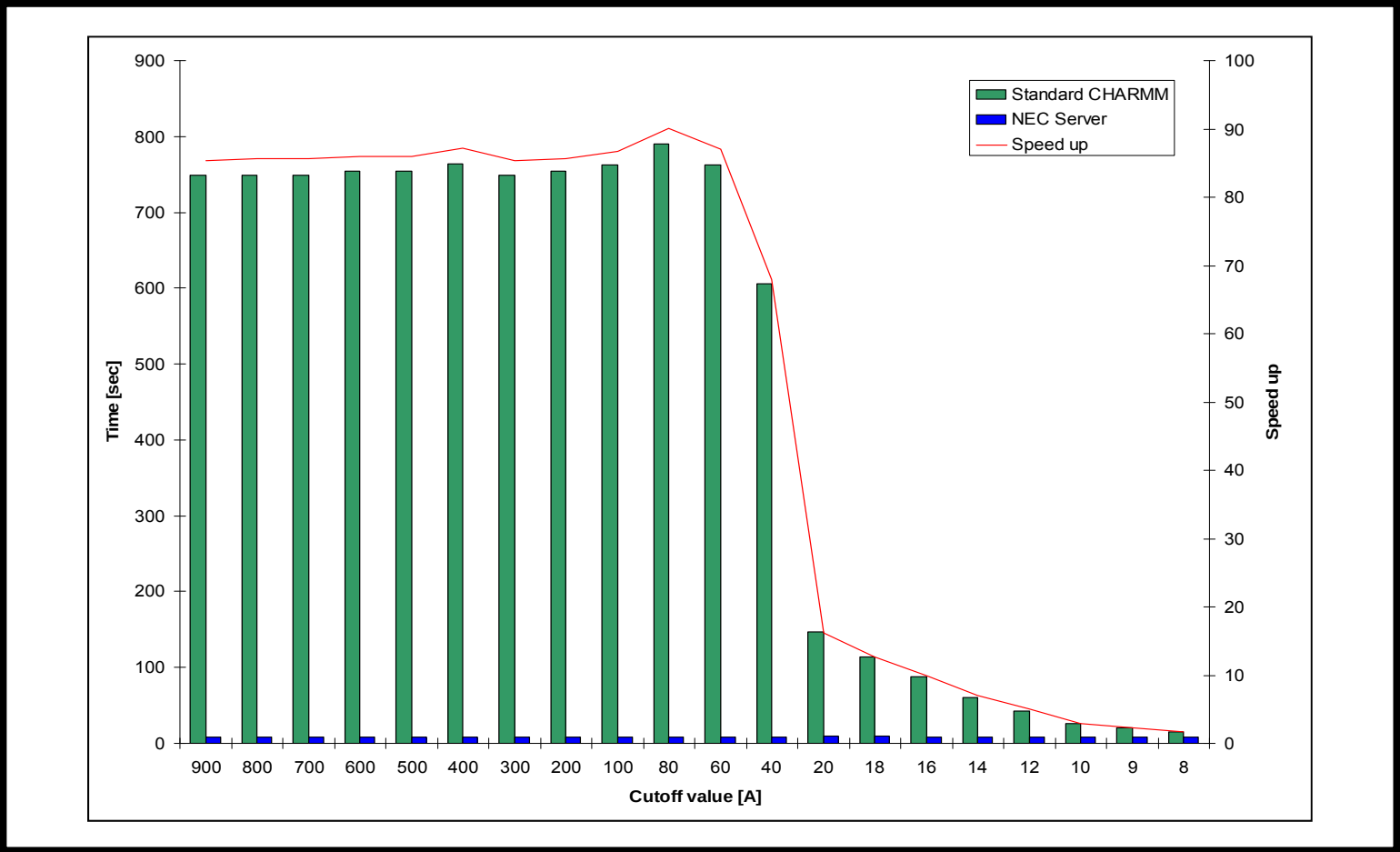
$$O(N^2) \rightarrow O(N)$$

- But speed comes at the cost of inaccurate modelling of long range interactions

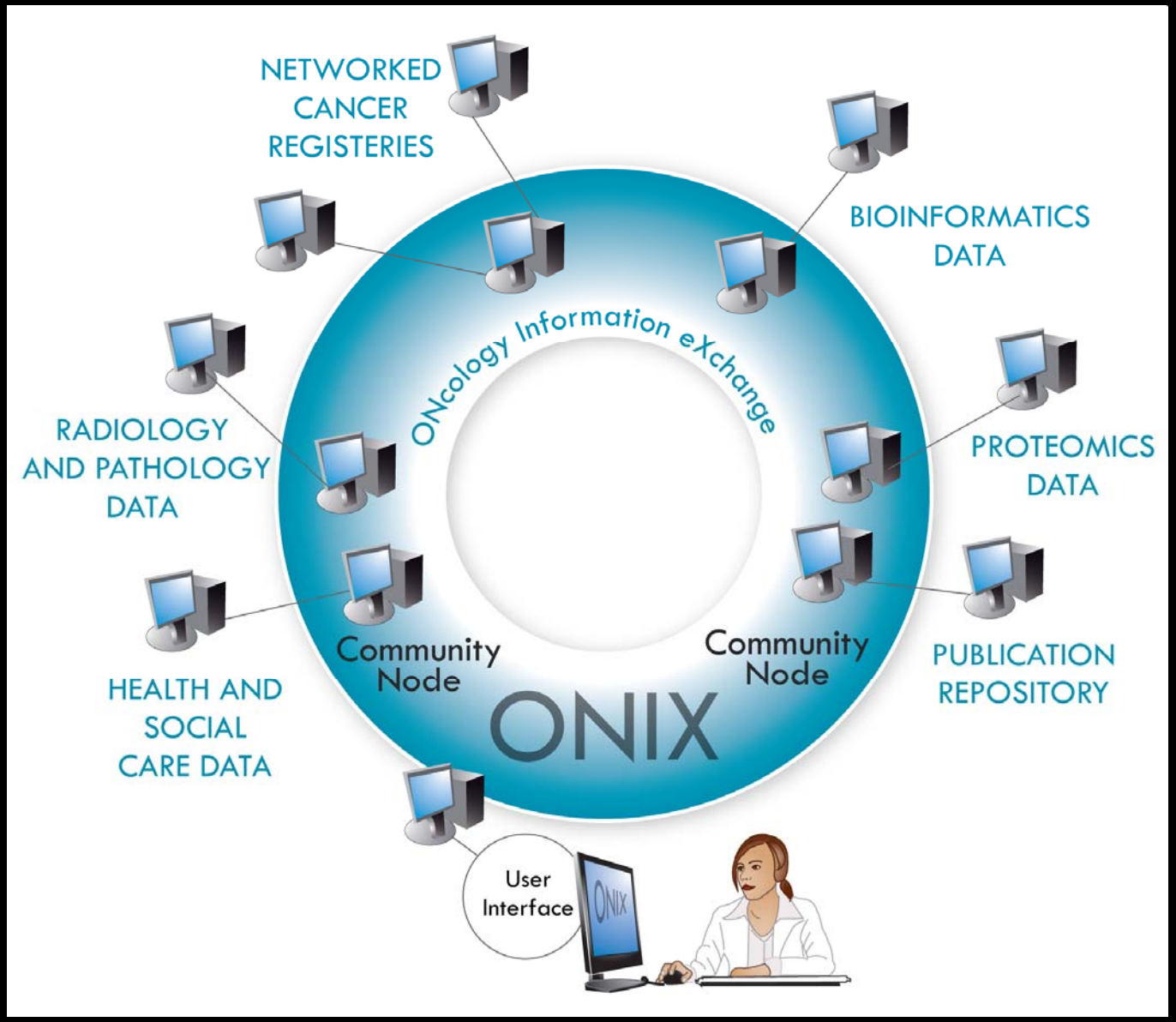
Slide courtesy of T. Shuepbach, SIB

CHARMM performance improvement

System of 10,713 atoms , CHARMM v32b1 compiled with g77 v3.2.6
Wall clock taken from single run on NEC MD server (Xeon 3.8GHz)



Slide courtesy of T. Shuepbach, SIB



Source: NCRI Informatics Initiative

caBIG® Cancer Biomedical Informatics Grid®

Source: caBIG

46 NCI-designated Cancer Centers
16 Community Cancer Centers

Clinical:

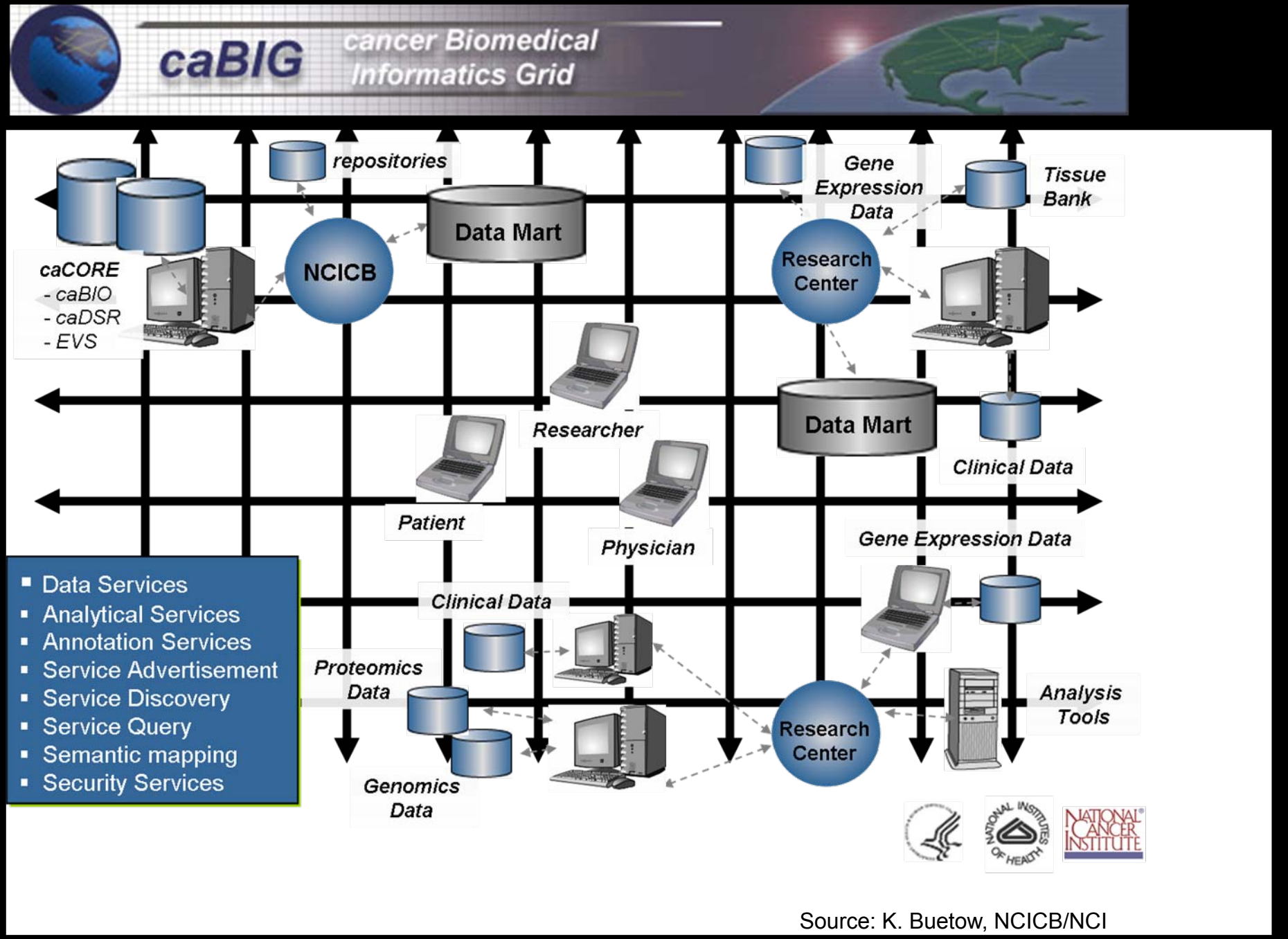
- Track and manage adverse events
- Provide current enrollment statistics
- Schedule and track patient encounters
- Integrate labs results and patient records
- Support regulatory compliance

Discovery

- Manage biospecimen information
- Obtain access to microarray data
- Accelerate analysis of results
- Access data across multiple studies
- Facilitate integration of diverse data types

Health-care delivery

- Facilitate Interoperability
- Promote data-sharing and collaboration
- Leverage common identity and security
- Query across multiple data resources



Source: K. Buetow, NCICB/NCI