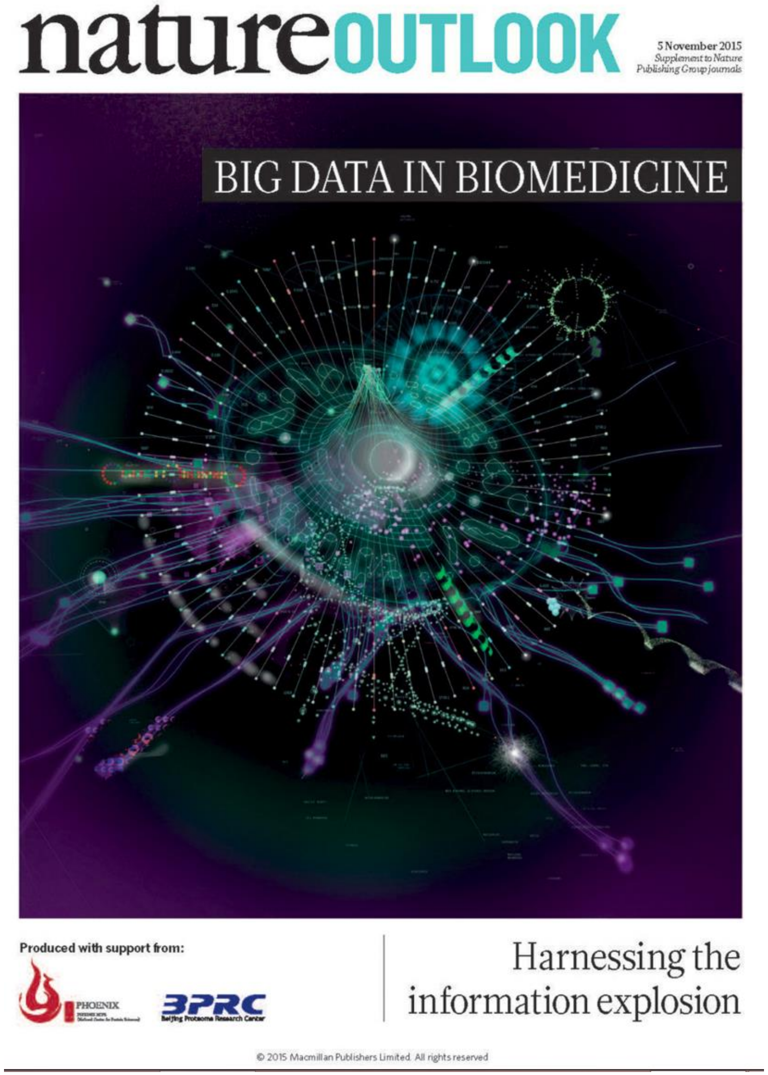
Data Tsunami as a Limiting Step in Using the All Omics Approach

ESMO Asia 2015

Yu Shyr, Ph.D.
Vanderbilt University
yu.shyr@vanderbilt.edu

Big data in biomedicine 2015 Vol. 527 No. 7576_supp ppS1-S19



nature OUTLOOK

5 November 2015
Supplement to Nature
Publishing Group journals

BIG DATA IN BIOMEDICINE

Produced with support from:

PHOENIX

3PRC
Beijing Proteome Research Center

Harnessing the information explosion

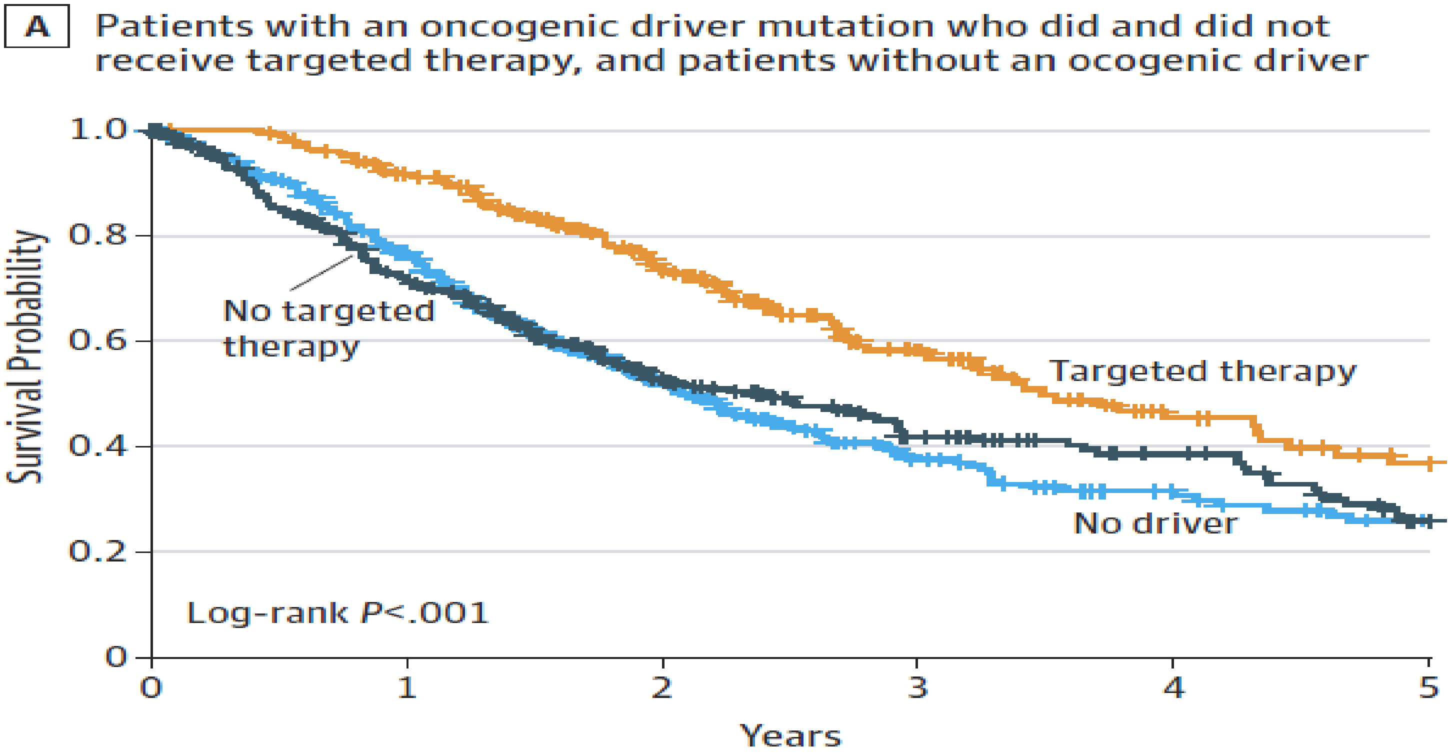
Advances in sequencing technology have triggered a tsunami of genomic data, and these are joined by waves of information from other '-omics' studies, clinical trials and patient records. **Analysis of this big data is launching the era of precision medicine** — **but enormous scientific, engineering and institutional challenges remain**.

JAMA. 2014;311(19):1998-2006. doi:10.1001/jama.2014.3741

Original Investigation

Using Multiplexed Assays of Oncogenic Drivers in Lung Cancers to Select Targeted Drugs

Mark G. Kris, MD , et. al.

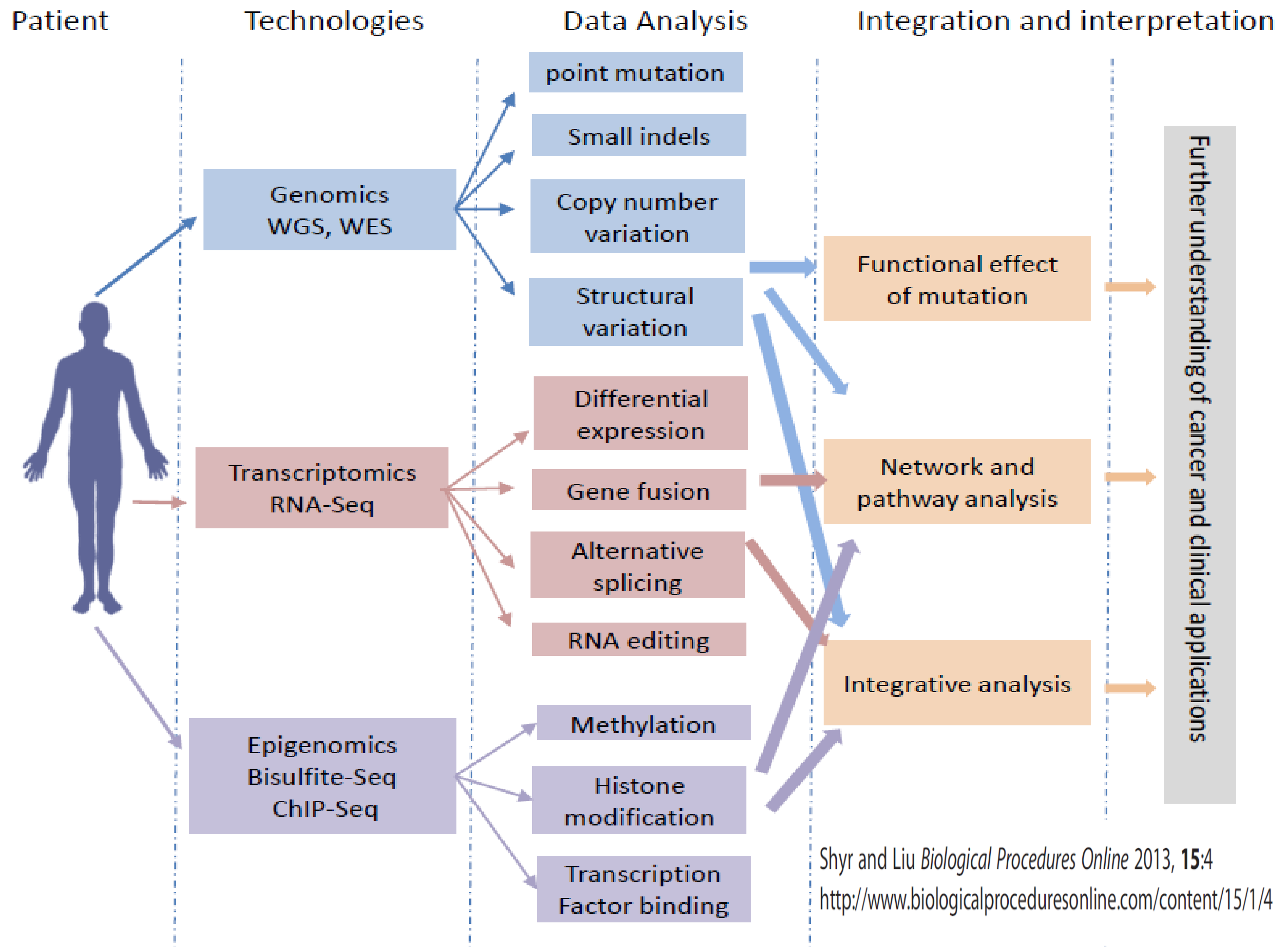


Highlights

- Overview of the BIG data in biomedical research
  - Omics data
  - EHR data
  - Data from patients & other sources
- Analytical challenges & tasks
- Future of the BIG data in biomedical research

Omics biomedical research

- **Microarray: cDNA (about 5,000 variables), Affymetrix U133 Plus 2.0 (about 45,000 variables)**
- **SNPs (about 500,000 – 2,000,000 variables)**
- **Next Generation Sequencing (?)**



Patient

Technologies

Genomics
WGS, WES

Transcriptomics
RNA-Seq

Epigenomics
Bisulfite-Seq
ChIP-Seq

Data Analysis

point mutation

Small indels

Copy number variation

Structural variation

Differential expression

Gene fusion

Alternative splicing

RNA editing

Methylation

Histone modification

Transcription Factor binding

Integration and interpretation

Functional effect of mutation

Network and pathway analysis

Integrative analysis

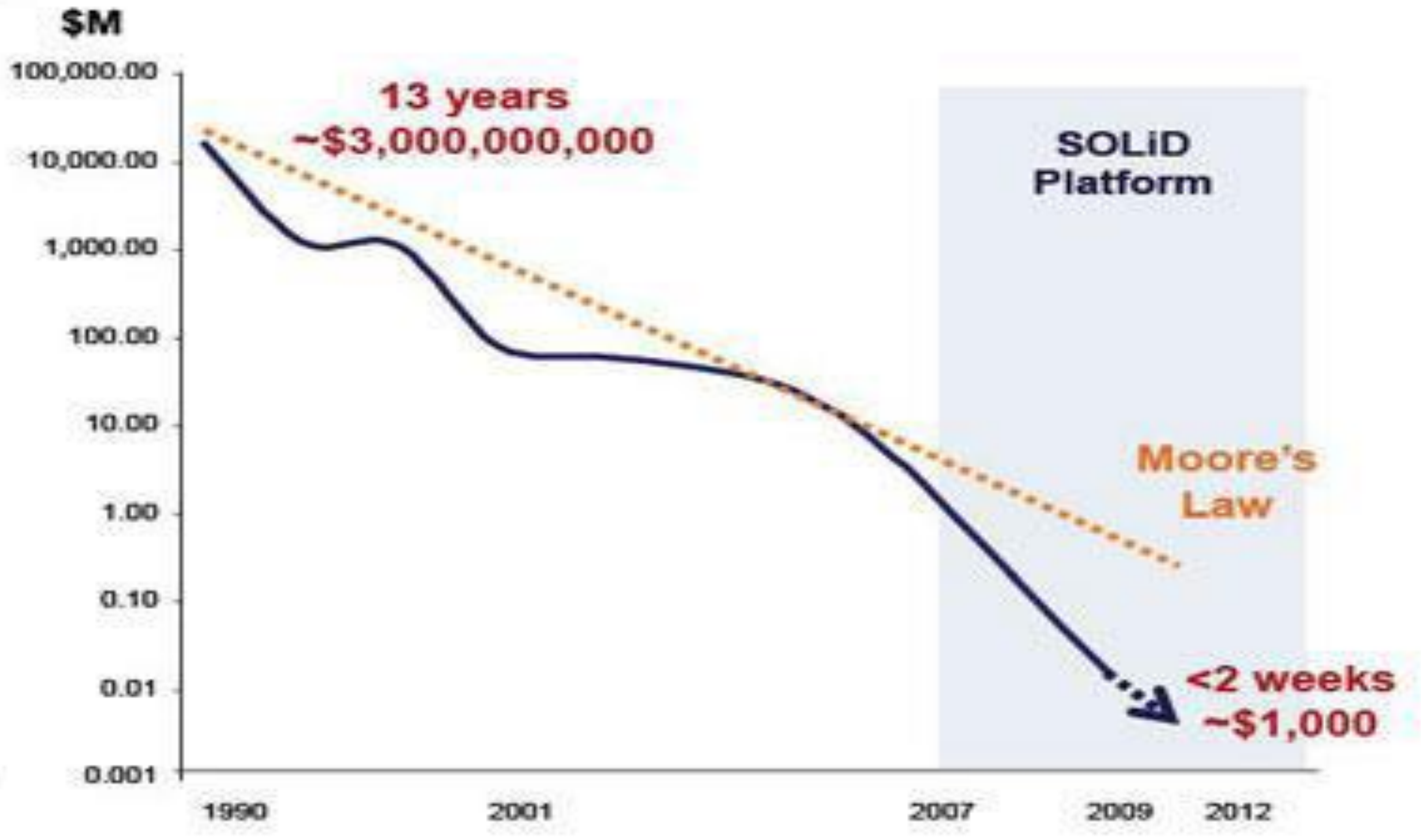
Further understanding of cancer and clinical applications

Shyr and Liu *Biological Procedures Online* 2013, **15**:4
http://www.biologicalproceduresonline.com/content/15/1/4

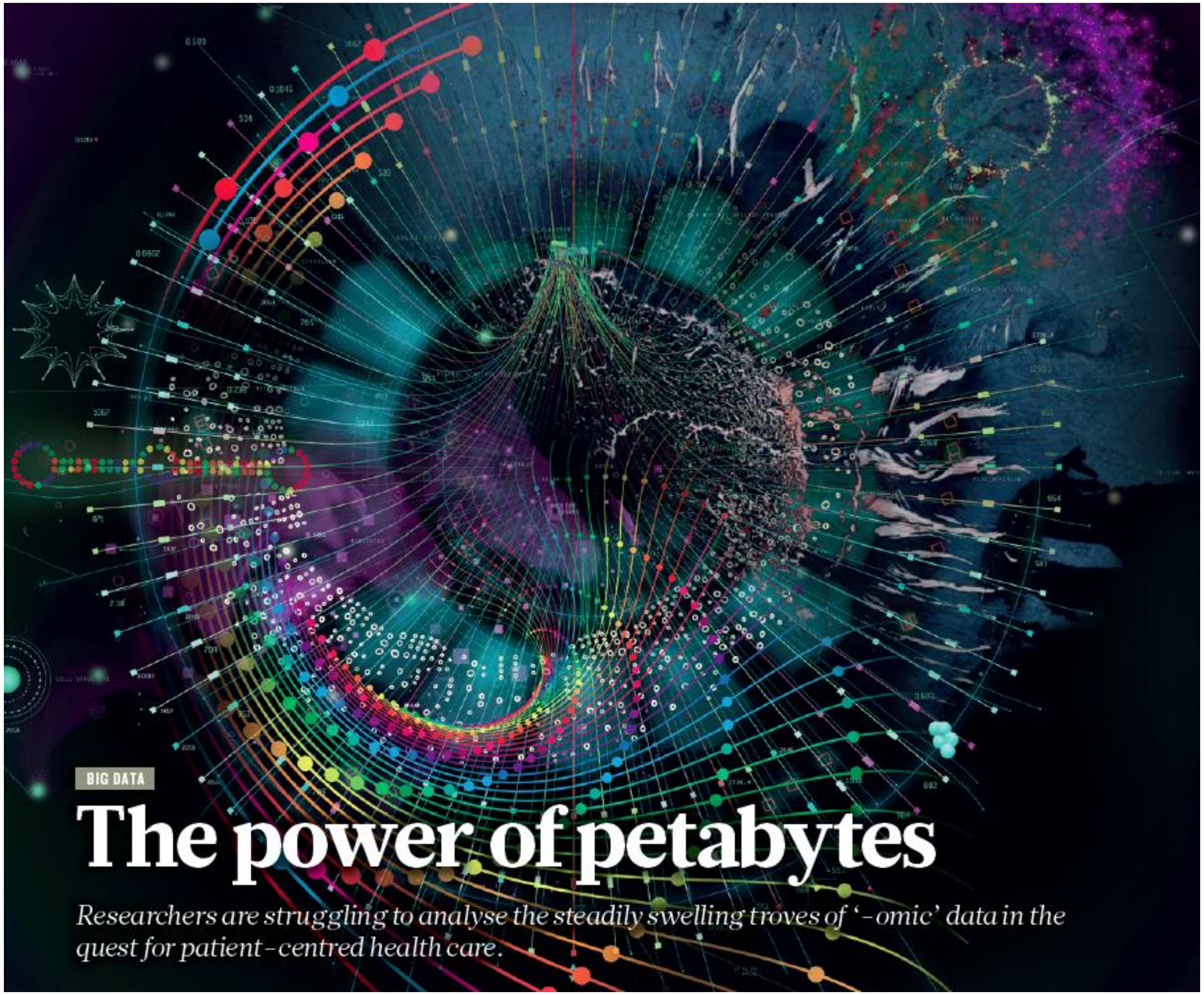
Storage of the Data?

- **cDNA, Microarray, SNPs**
- **NGseq raw imaging data: > 2 TB per sample**
- **RNAseq or Exome seq data: 10 GB per sample (raw data), 30-50 GB during the processing.**
- **Whole genome seq: 200 GB per sample (raw data), 400-600 GB during the processing.**

Cost per Human Genome



Data analysis/mining ?



BIG DATA

The power of petabytes

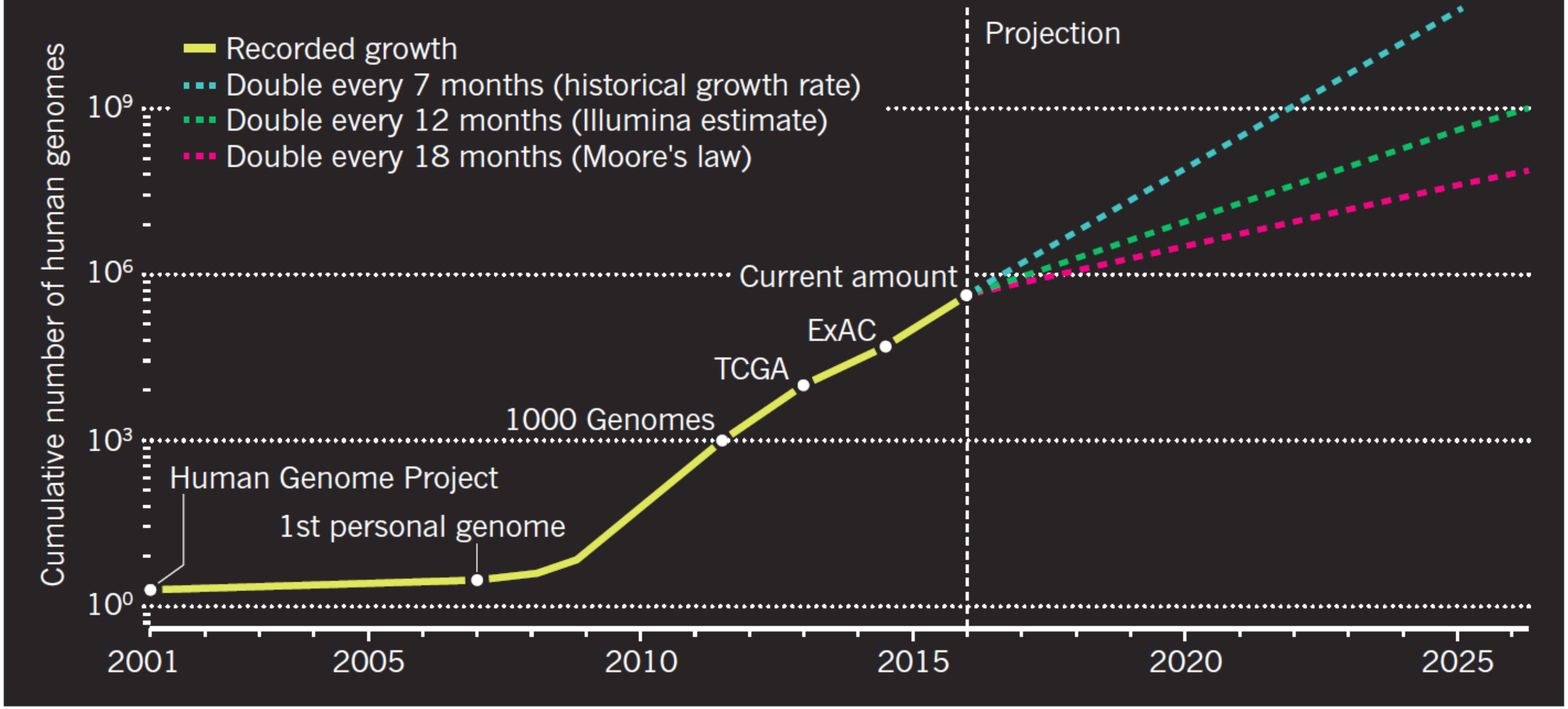
Researchers are struggling to analyse the steadily swelling troves of '-omic' data in the quest for patient-centred health care.

> ***"You're shooting yourself in the foot if you're collecting data you don't know how to interpret."***

S2 | NATURE | VOL 527 | 5 NOVEMBER 2015

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TGCA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.

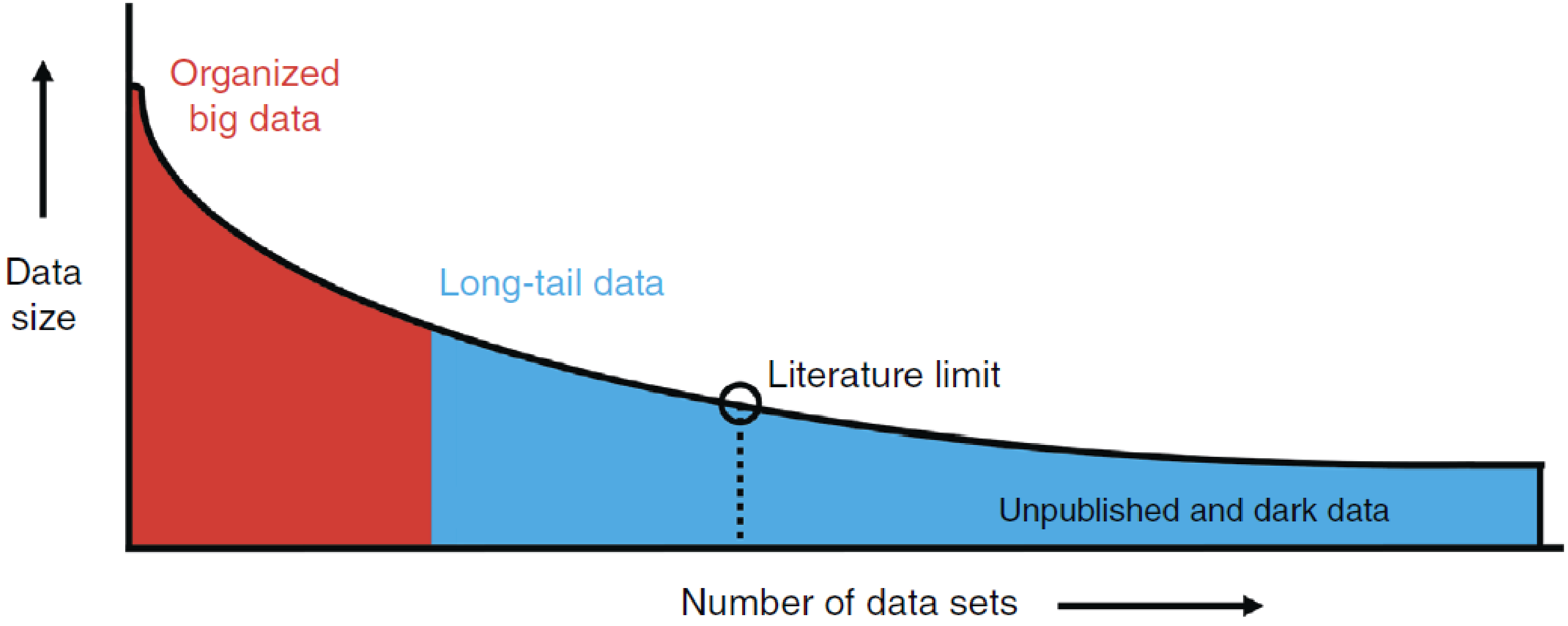


- In 2014 the United Kingdom launched the 100,000 Genomes Project, and both the United States (under the Precision Medicine Initiative) and China (in a programme to be run by BGI of Shenzhen) have unveiled plans to analyze genomic data from one million individuals.

- A partnership between Geisinger Health System, based in Danville, Pennsylvania, and biotech firm Regeneron Pharmaceuticals of Tarrytown, New York, for instance, aims to generate sequence data for more than 250,000 people.

Big data from small data: data-sharing in the ‘long tail’ of neuroscience

Adam R Ferguson[1], Jessica L Nielson[1], Melissa H Cragin[2], Anita E Bandrowski[3] & Maryann E Martone[3,4]



VOLUME 17 | NUMBER 11 | NOVEMBER 2014 NATURE NEUROSCIENCE

CANCER

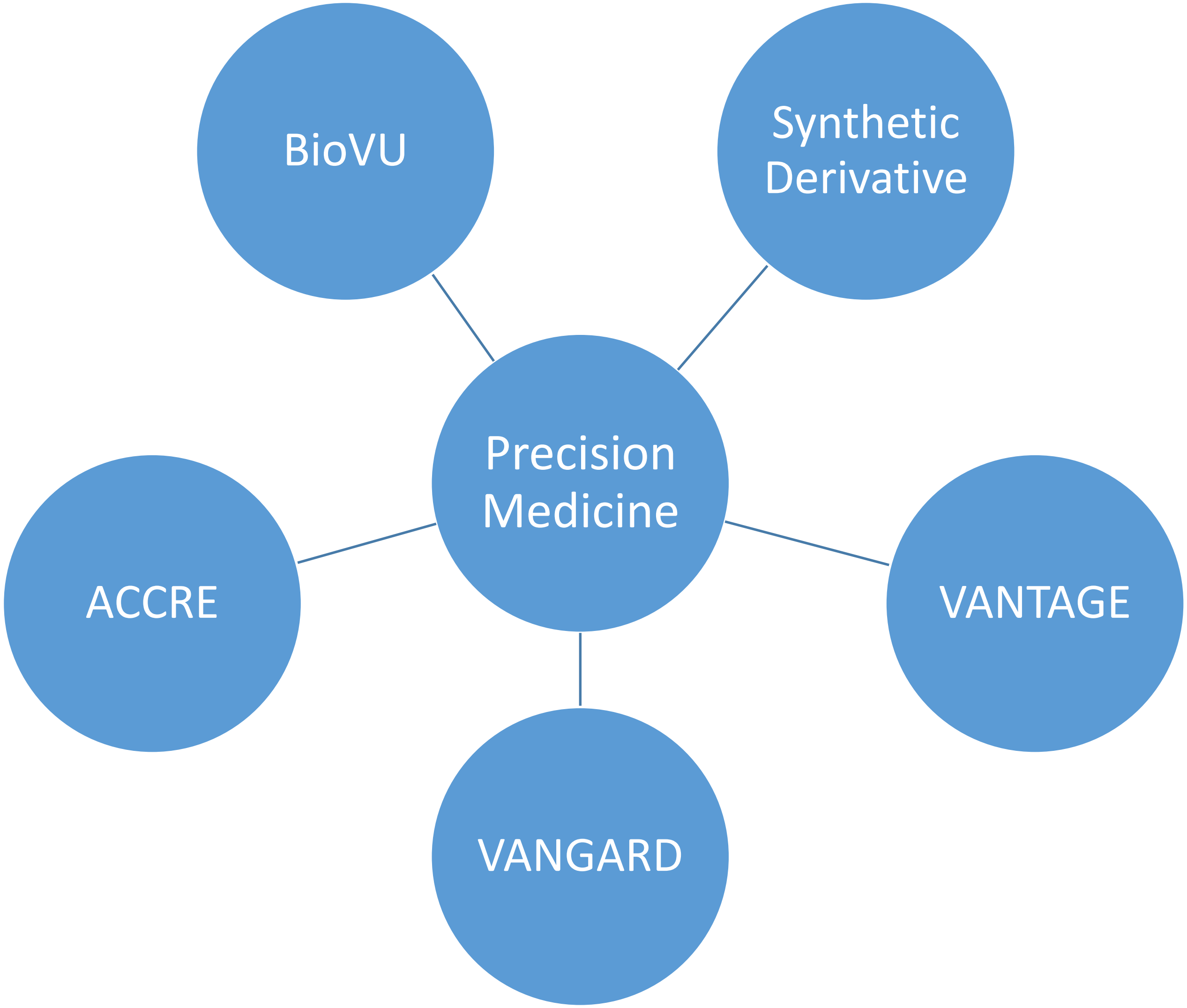
Reshaping the cancer clinic

Big data's war on cancer is still in the early stages, but the front line is advancing.

The Cancer Genome Atlas, which catalogues cancer mutations, contains some 2.5 million gigabytes of data. This giant project, run by the US National Institutes of Health, has vastly improved our understanding of various forms of cancer — but it holds relatively little information on the clinical experience of the patients who supplied the samples.

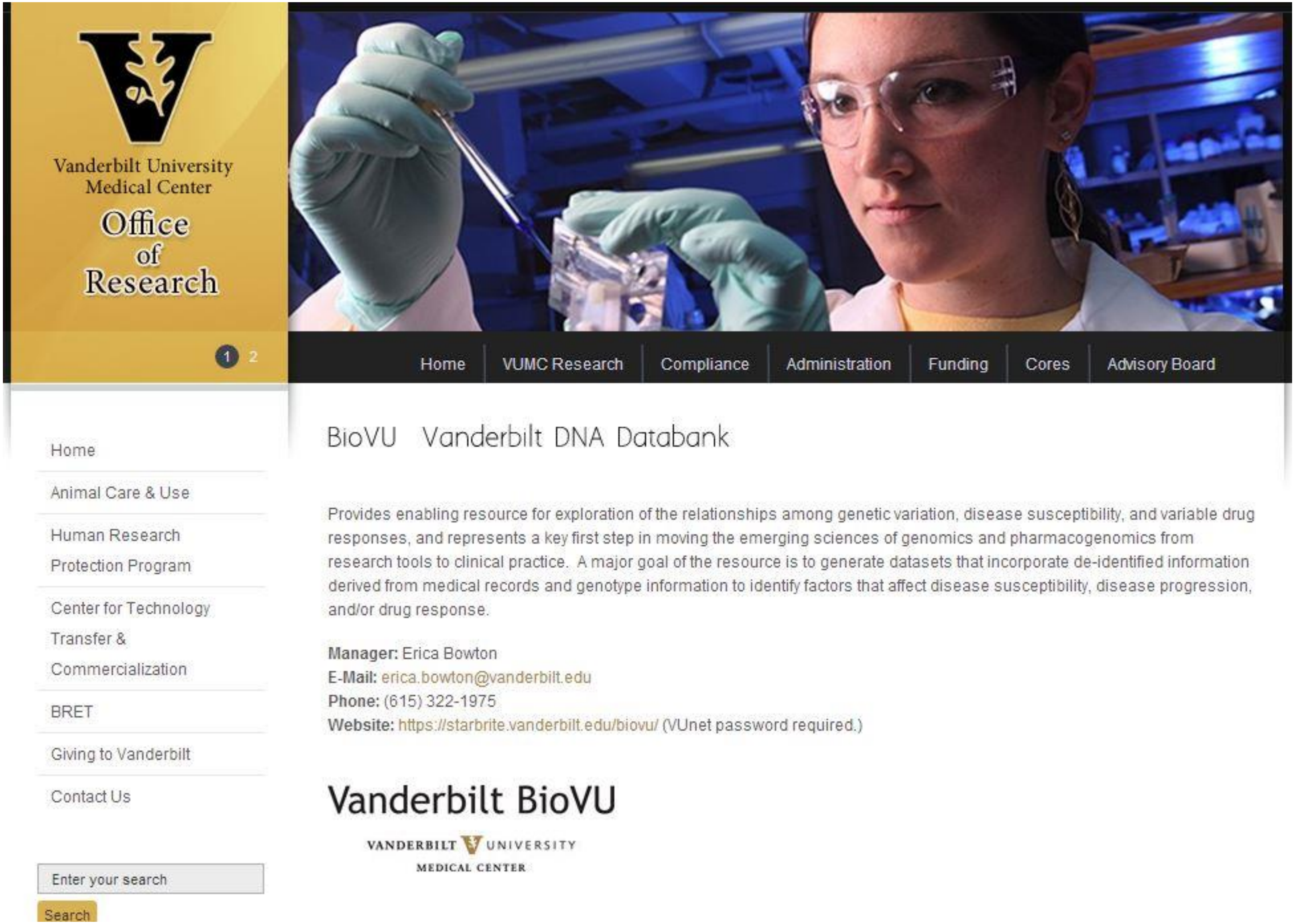
S10 | NATURE | VOL 527 | 5 NOVEMBER 2015

VU Precision Medicine



Precision Medicine

- BioVU
- Synthetic Derivative
- ACCRE
- VANTAGE
- VANGARD



BioVU Vanderbilt DNA Databank

Provides enabling resource for exploration of the relationships among genetic variation, disease susceptibility, and variable drug responses, and represents a key first step in moving the emerging sciences of genomics and pharmacogenomics from research tools to clinical practice. A major goal of the resource is to generate datasets that incorporate de-identified information derived from medical records and genotype information to identify factors that affect disease susceptibility, disease progression, and/or drug response.

Manager: Erica Bowton
E-Mail: erica.bowton@vanderbilt.edu
Phone: (615) 322-1975
Website: https://starbrite.vanderbilt.edu/biovu/ (VUnet password required.)

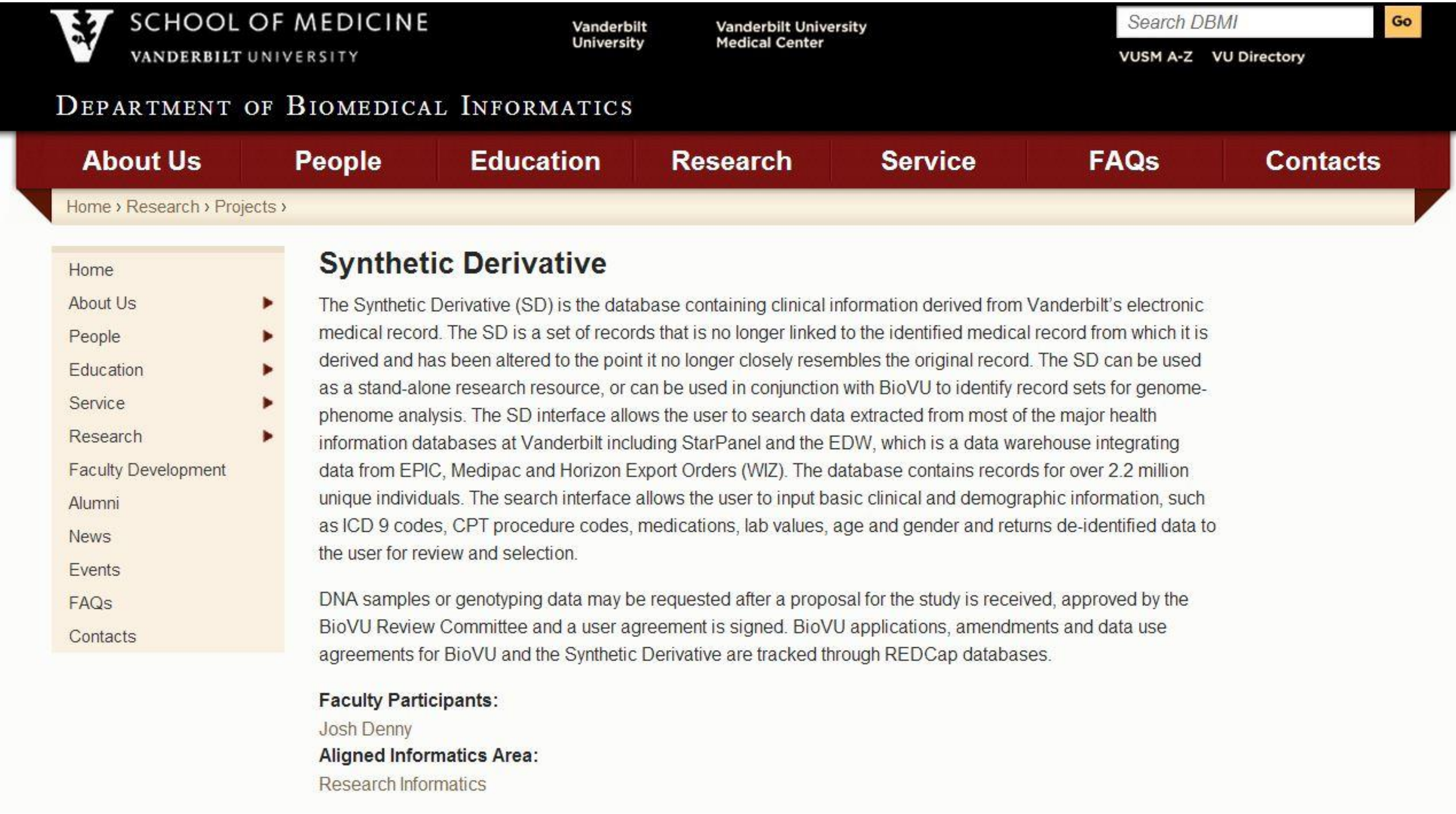
Vanderbilt BioVU

VANDERBILT UNIVERSITY
MEDICAL CENTER

What is BioVU?

- The move towards personalized medicine requires very large sample sets for discovery and validation
- BioVU: biobank intended to support a broad view of biology and enable personalized medicine
- Contains de-identified DNA extracted from leftover blood after clinically-indicated testing of Vanderbilt patients who have not opted out
- A major goal of the resource is to generate datasets that incorporate de-identified information derived from medical records and genotype information to identify factors that affect disease susceptibility, disease progression, and/or drug response.

What is the Synthetic Derivative (SD)?



Synthetic Derivative

The Synthetic Derivative (SD) is the database containing clinical information derived from Vanderbilt's electronic medical record. The SD is a set of records that is no longer linked to the identified medical record from which it is derived and has been altered to the point it no longer closely resembles the original record. The SD can be used as a stand-alone research resource, or can be used in conjunction with BioVU to identify record sets for genome-phenome analysis. The SD interface allows the user to search data extracted from most of the major health information databases at Vanderbilt including StarPanel and the EDW, which is a data warehouse integrating data from EPIC, Medipac and Horizon Export Orders (WIZ). The database contains records for over 2.2 million unique individuals. The search interface allows the user to input basic clinical and demographic information, such as ICD 9 codes, CPT procedure codes, medications, lab values, age and gender and returns de-identified data to the user for review and selection.

DNA samples or genotyping data may be requested after a proposal for the study is received, approved by the BioVU Review Committee and a user agreement is signed. BioVU applications, amendments and data use agreements for BioVU and the Synthetic Derivative are tracked through REDCap databases.

Faculty Participants:

Josh Denny

Aligned Informatics Area:

Research Informatics

What is the Synthetic Derivative (SD)?

- Rich, multi-source database of de-identified clinical and demographic data
- User Interface tool that can be used for access and analysis
- Contains ~2.6 million records
  - ~1 million with detailed longitudinal data
  - averaging 100k bytes in size
  - an average of 27 codes per record

Technology + Policy

De-identification

- **Derivation of 128-character identifier (RUI) from the MRN generated by Secure Hash Algorithm (SHA-512)**

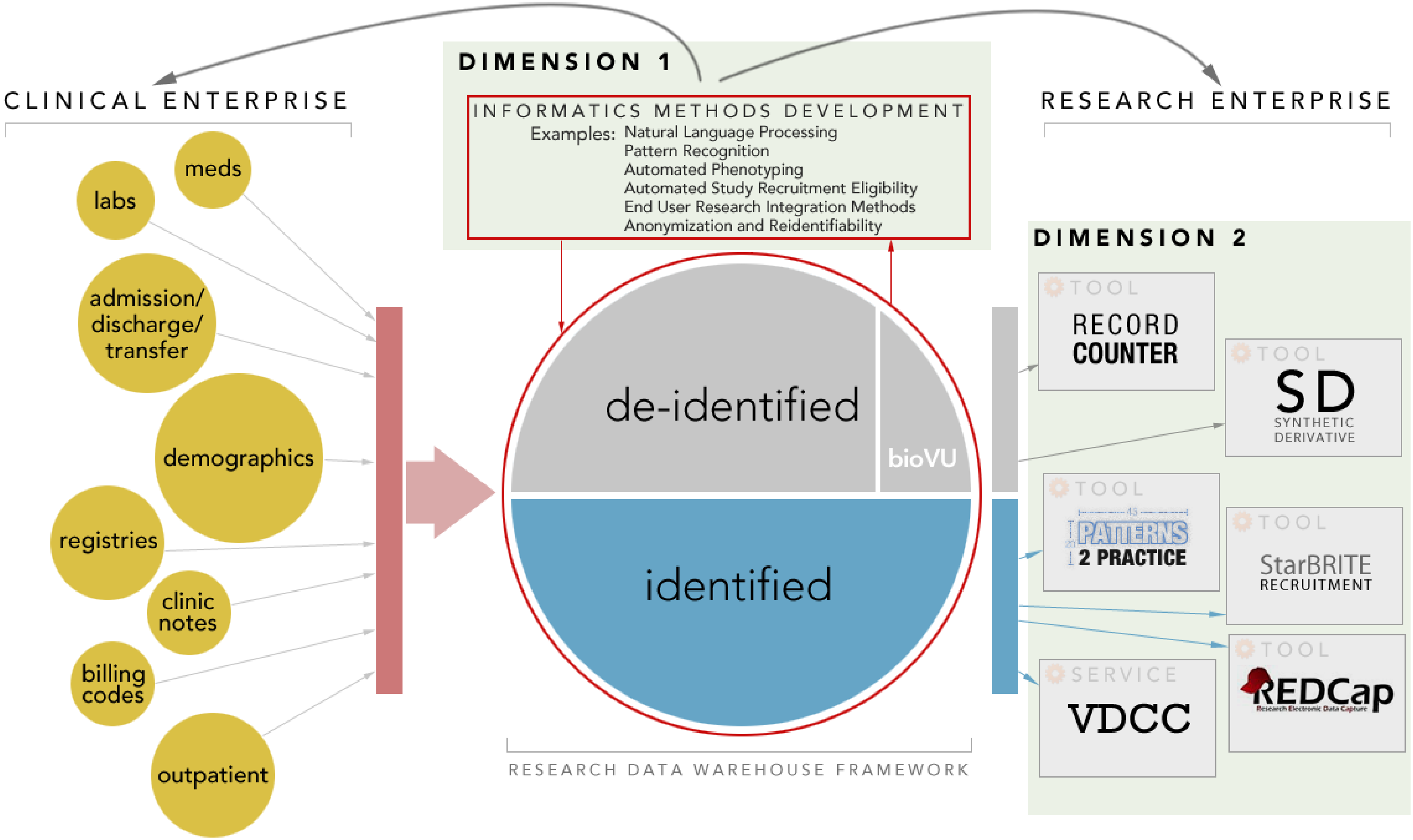
Date Shift

- **Our algorithm shifts the dates within a record by a time period *(up to 364 days backwards)* that is consistent within each record, but differs *across* records**

Restricted access & continuous oversight

- **IRB approval for study (non-human)**
- **Data Use Agreement**
- **Audit logs of all searches and data exports**

Synthetic Derivative (SD)



CLINICAL ENTERPRISE

meds
labs
admission/ discharge/ transfer
demographics
registries
clinic notes
billing codes
outpatient

DIMENSION 1

INFORMATICS METHODS DEVELOPMENT

Examples:
Natural Language Processing
Pattern Recognition
Automated Phenotyping
Automated Study Recruitment Eligibility
End User Research Integration Methods
Anonymization and Reidentifiability

de-identified
bioVU
identified

RESEARCH DATA WAREHOUSE FRAMEWORK

RESEARCH ENTERPRISE

DIMENSION 2

TOOL
RECORD COUNTER

TOOL
SD
SYNTHETIC DERIVATIVE

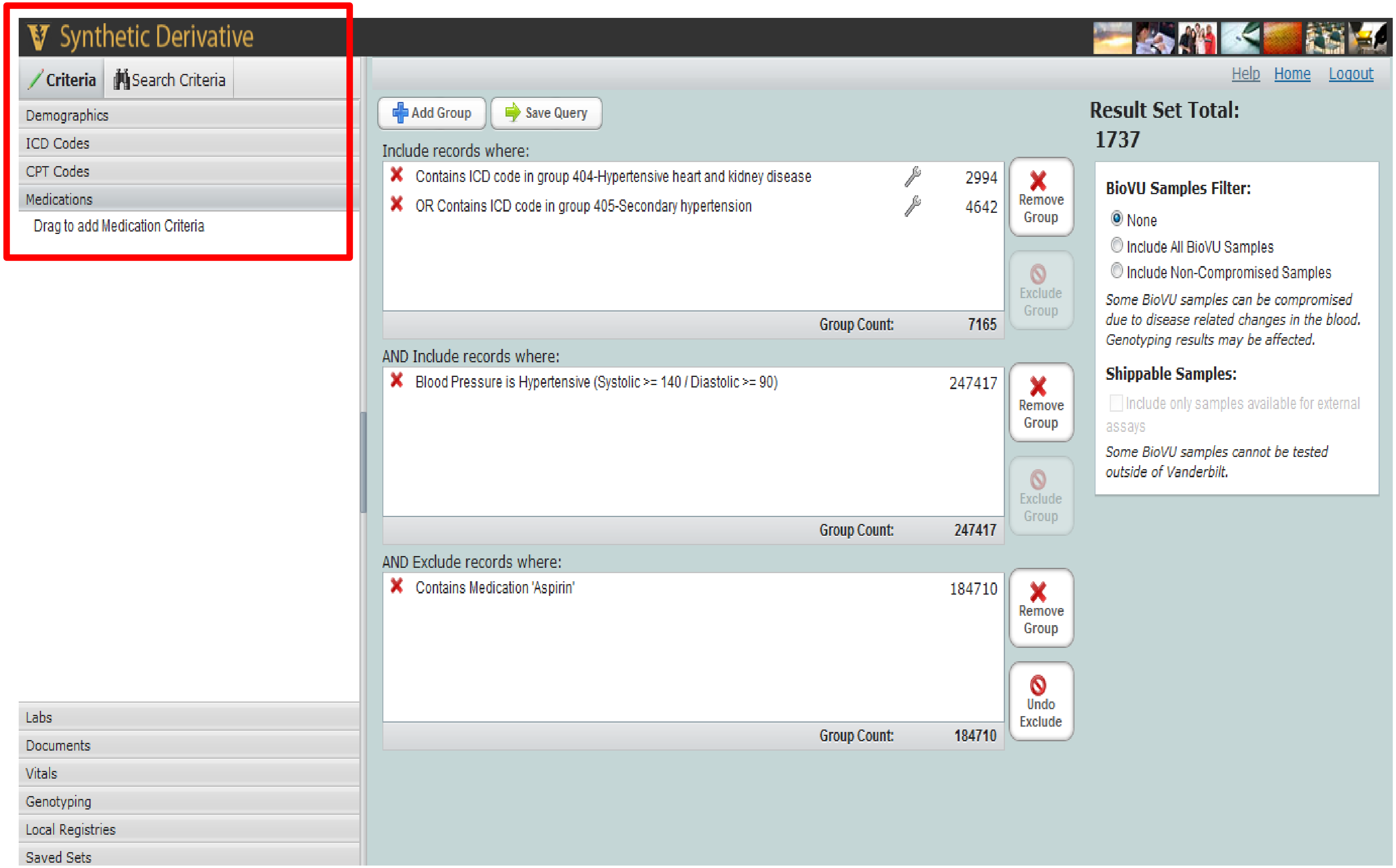
TOOL
PATTERNS 2 PRACTICE

TOOL
StarBRITE
RECRUITMENT

SERVICE
VDCC

TOOL
REDCap
Research Electronic Data Capture

Synthetic Derivative Search Features



Include records where:

Contains ICD code in group 404-Hypertensive heart and kidney disease	2994
OR Contains ICD code in group 405-Secondary hypertension	4642
Group Count:	7165

AND Include records where:

Blood Pressure is Hypertensive (Systolic >= 140 / Diastolic >= 90)	247417
Group Count:	247417

AND Exclude records where:

Contains Medication 'Aspirin'	184710
Group Count:	184710

Result Set Total:
1737

BioVU Samples Filter:

- ● None
- ○ Include All BioVU Samples
- ○ Include Non-Compromised Samples

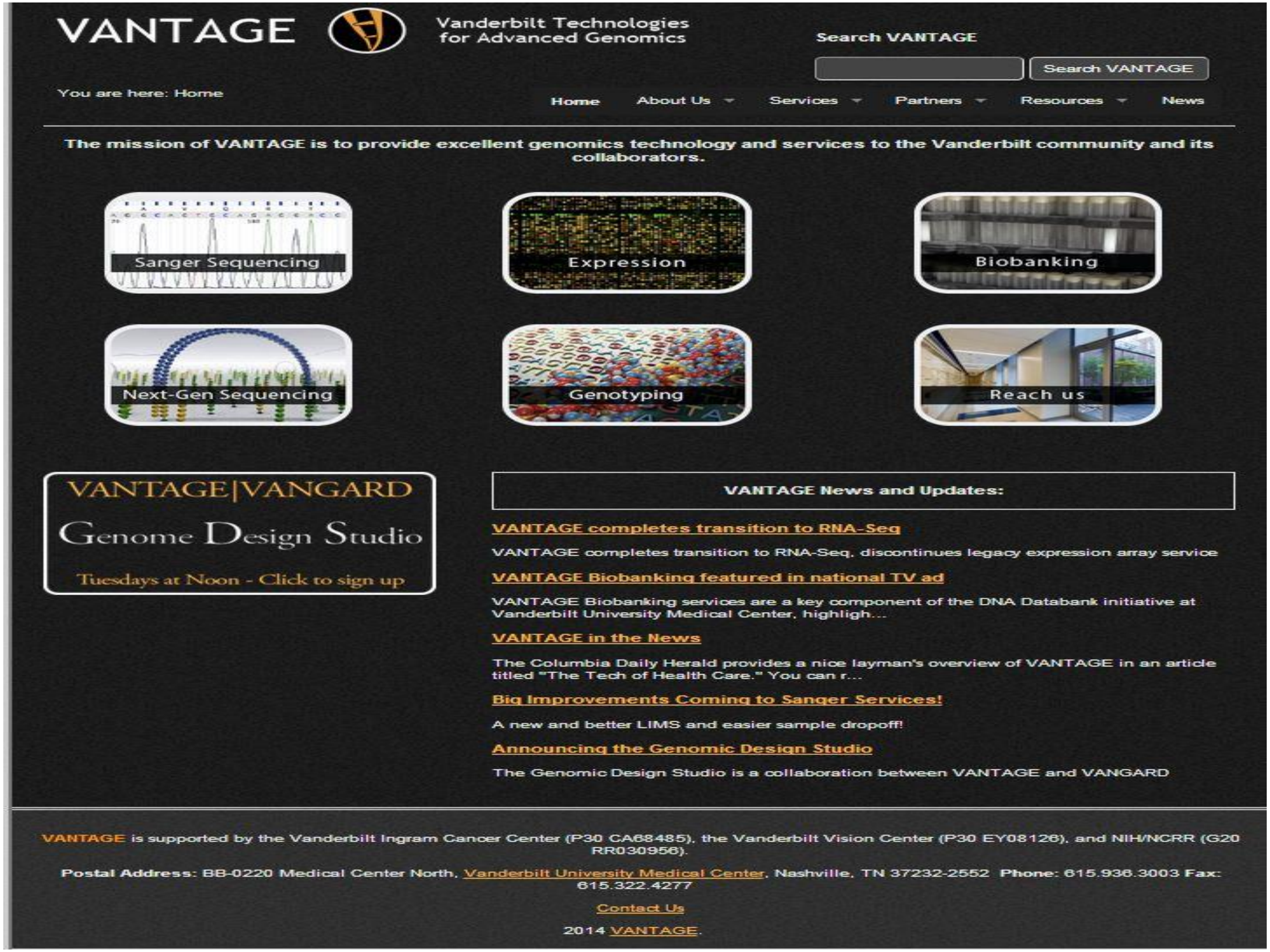
Some BioVU samples can be compromised due to disease related changes in the blood. Genotyping results may be affected.

Shippable Samples:

☐ Include only samples available for external assays

Some BioVU samples cannot be tested outside of Vanderbilt.

VANTAGE



The mission of VANTAGE is to provide excellent genomics technology and services to the Vanderbilt community and its collaborators.

VANTAGE News and Updates:

VANTAGE completes transition to RNA-Seq

VANTAGE completes transition to RNA-Seq, discontinues legacy expression array service

VANTAGE Biobanking featured in national TV ad

VANTAGE Biobanking services are a key component of the DNA Databank initiative at Vanderbilt University Medical Center, highligh...

VANTAGE in the News

The Columbia Daily Herald provides a nice layman's overview of VANTAGE in an article titled "The Tech of Health Care." You can r...

Big Improvements Coming to Sanger Services!

A new and better LIMS and easier sample dropoff!

Announcing the Genomic Design Studio

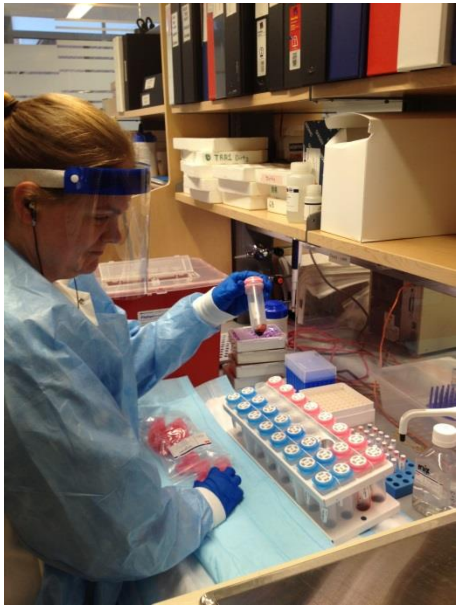
The Genomic Design Studio is a collaboration between VANTAGE and VANGARD

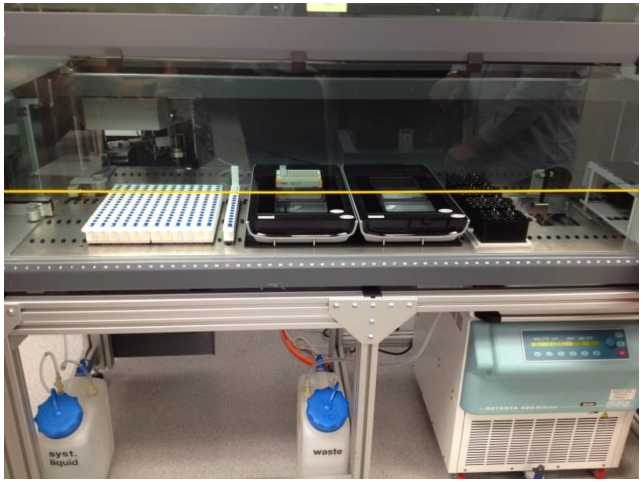
VANTAGE is supported by the Vanderbilt Ingram Cancer Center (P30 CA68485), the Vanderbilt Vision Center (P30 EY08126), and NIH/NCRR (G20 RR030956).

Postal Address: BB-0220 Medical Center North, Vanderbilt University Medical Center, Nashville, TN 37232-2552 Phone: 615.936.3003 Fax: 615.322.4277

Contact Us

2014 VANTAGE.

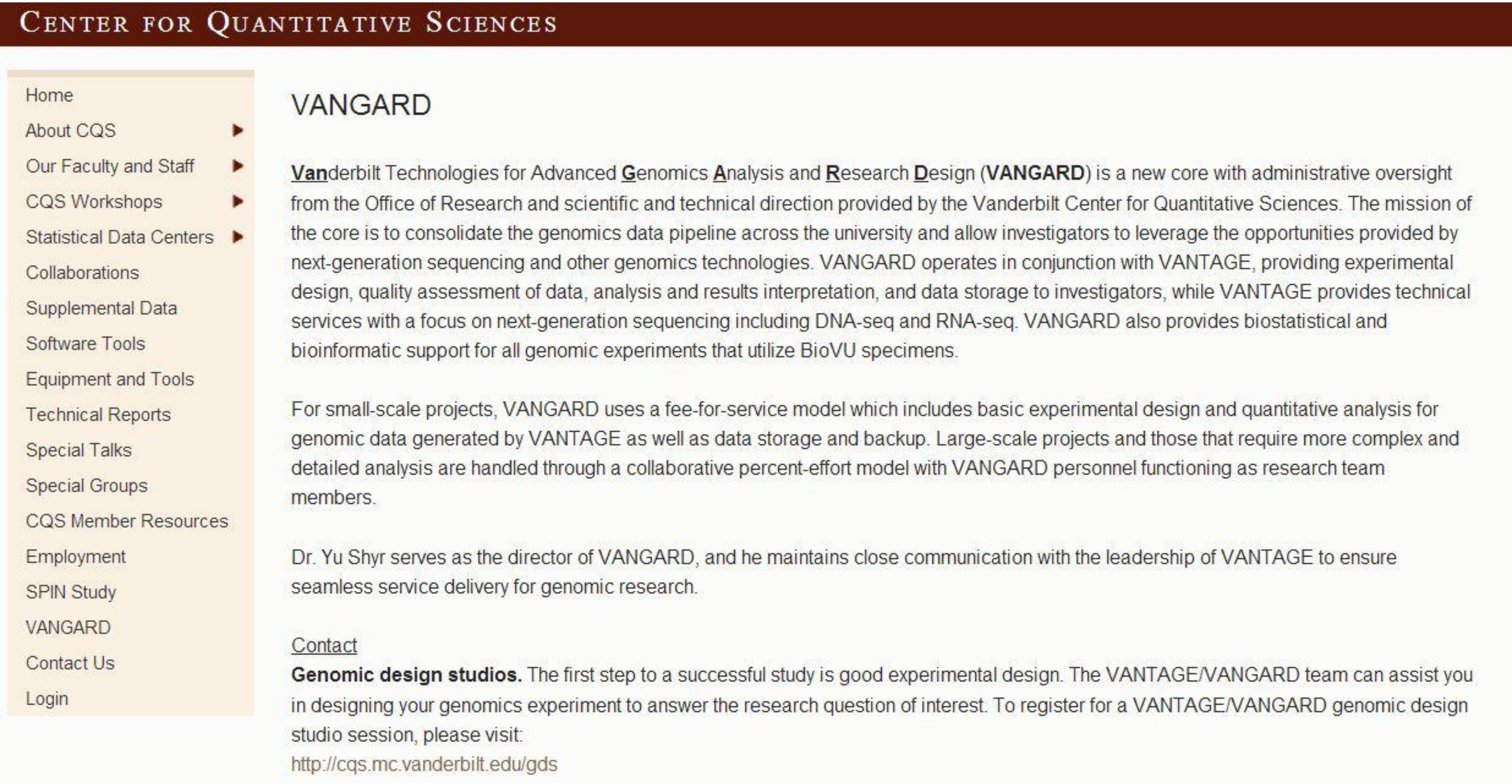




VANTAGE

BioVU

VANGARD



VANGARD

Vanderbilt Technologies for Advanced **G**enomics **A**nalysis and **R**esearch **D**esign (**VANGARD**) is a new core with administrative oversight from the Office of Research and scientific and technical direction provided by the Vanderbilt Center for Quantitative Sciences. The mission of the core is to consolidate the genomics data pipeline across the university and allow investigators to leverage the opportunities provided by next-generation sequencing and other genomics technologies. VANGARD operates in conjunction with VANTAGE, providing experimental design, quality assessment of data, analysis and results interpretation, and data storage to investigators, while VANTAGE provides technical services with a focus on next-generation sequencing including DNA-seq and RNA-seq. VANGARD also provides biostatistical and bioinformatic support for all genomic experiments that utilize BioVU specimens.

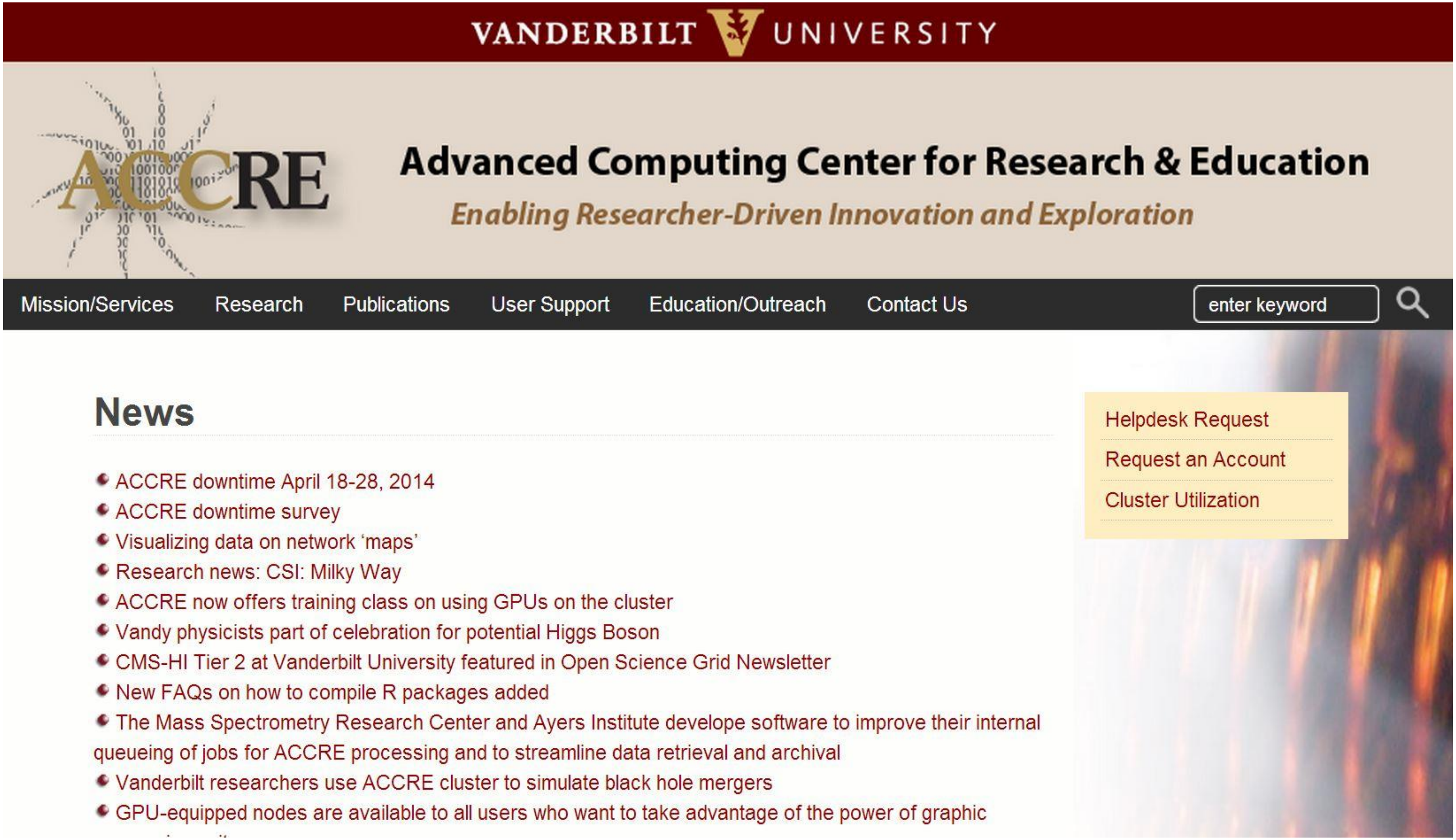
For small-scale projects, VANGARD uses a fee-for-service model which includes basic experimental design and quantitative analysis for genomic data generated by VANTAGE as well as data storage and backup. Large-scale projects and those that require more complex and detailed analysis are handled through a collaborative percent-effort model with VANGARD personnel functioning as research team members.

Dr. Yu Shyr serves as the director of VANGARD, and he maintains close communication with the leadership of VANTAGE to ensure seamless service delivery for genomic research.

Contact

Genomic design studios. The first step to a successful study is good experimental design. The VANTAGE/VANGARD team can assist you in designing your genomics experiment to answer the research question of interest. To register for a VANTAGE/VANGARD genomic design studio session, please visit:
http://cqs.mc.vanderbilt.edu/gds

Vanderbilt ACCRE



Advanced Computing Center for Research & Education

Enabling Researcher-Driven Innovation and Exploration

News

- ACCRE downtime April 18-28, 2014
- ACCRE downtime survey
- Visualizing data on network 'maps'
- Research news: CSI: Milky Way
- ACCRE now offers training class on using GPUs on the cluster
- Vandy physicists part of celebration for potential Higgs Boson
- CMS-HI Tier 2 at Vanderbilt University featured in Open Science Grid Newsletter
- New FAQs on how to compile R packages added
- The Mass Spectrometry Research Center and Ayers Institute develope software to improve their internal queueing of jobs for ACCRE processing and to streamline data retrieval and archival
- Vanderbilt researchers use ACCRE cluster to simulate black hole mergers
- GPU-equipped nodes are available to all users who want to take advantage of the power of graphic

Helpdesk Request

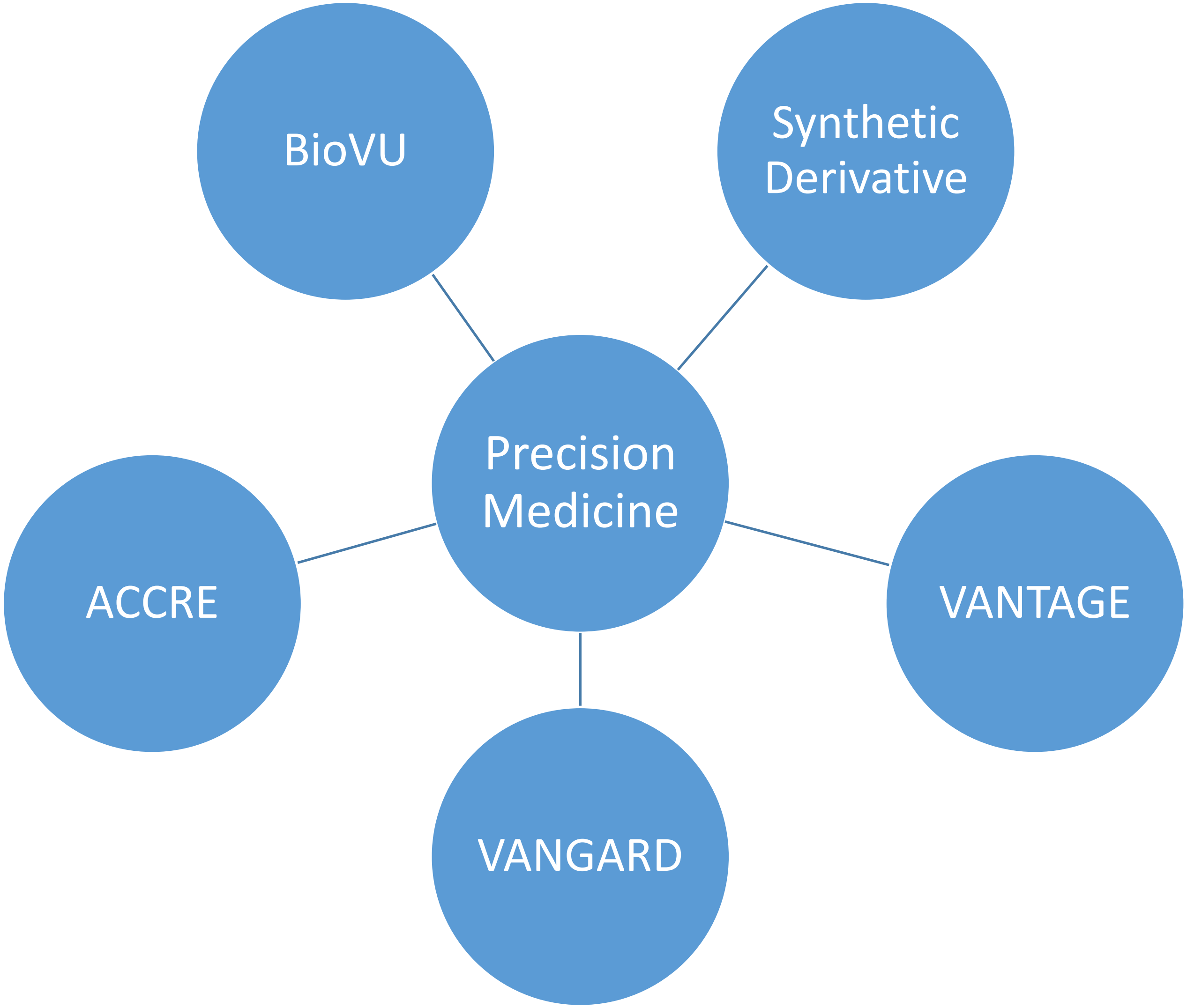
Request an Account

Cluster Utilization

Vanderbilt ACCRE

- The ACCRE high-performance computing cluster has about 6,000 processor cores.
- Compute nodes run 64-bit Linux operating system, with hard drives of 250 GB to 1 TB and dual copper gigabit Ethernet ports.
- Each node is monitored via Nagios, with an integrated scheduling system (Moab/Torque) utilized for resource management, scheduling of jobs, and usage tracking.
- The home directories of all users are backed up daily to tape.

VU Precision Medicine


Precision Medicine

BioVU

Synthetic Derivative

ACCRE

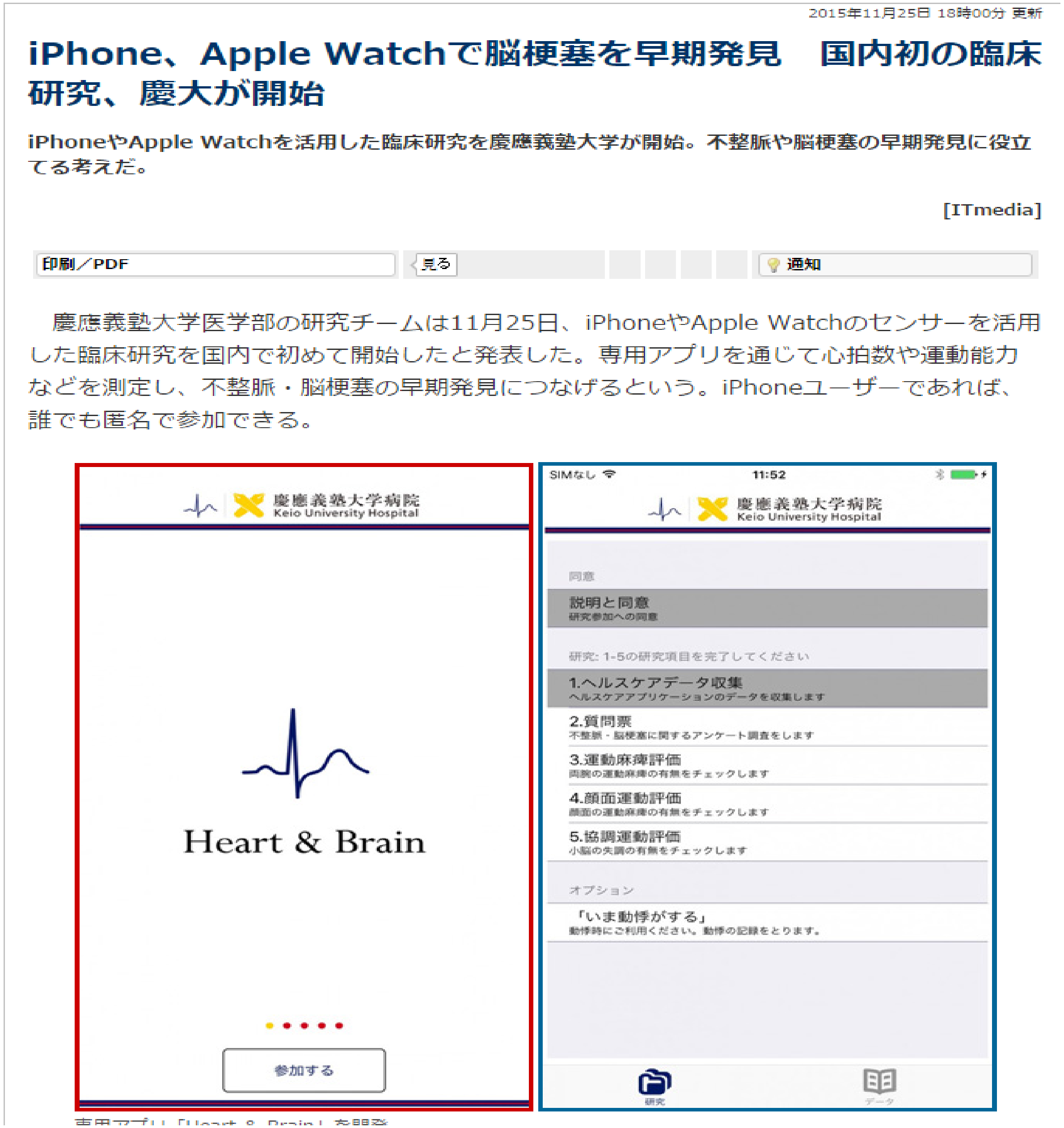
VANTAGE

VANGARD

Apple Announced ResearchKit on 4/14/2015

- **ResearchKit™, an open source software framework designed for medical and health research that helps doctors, scientists and other researchers gather data more frequently and more accurately from participants using mobile devices, is now available to researchers and developers.**

- **The first research apps developed using ResearchKit study asthma, breast cancer, cardiovascular disease, diabetes and Parkinson's disease, and have enrolled over 60,000 iPhone users in just the first few weeks of being available on the App Store.**



2015年11月25日 18時00分 更新

iPhone、Apple Watchで脳梗塞を早期発見　国内初の臨床研究、慶大が開始

iPhoneやApple Watchを活用した臨床研究を慶應義塾大学が開始。不整脈や脳梗塞の早期発見に役立てる考えだ。

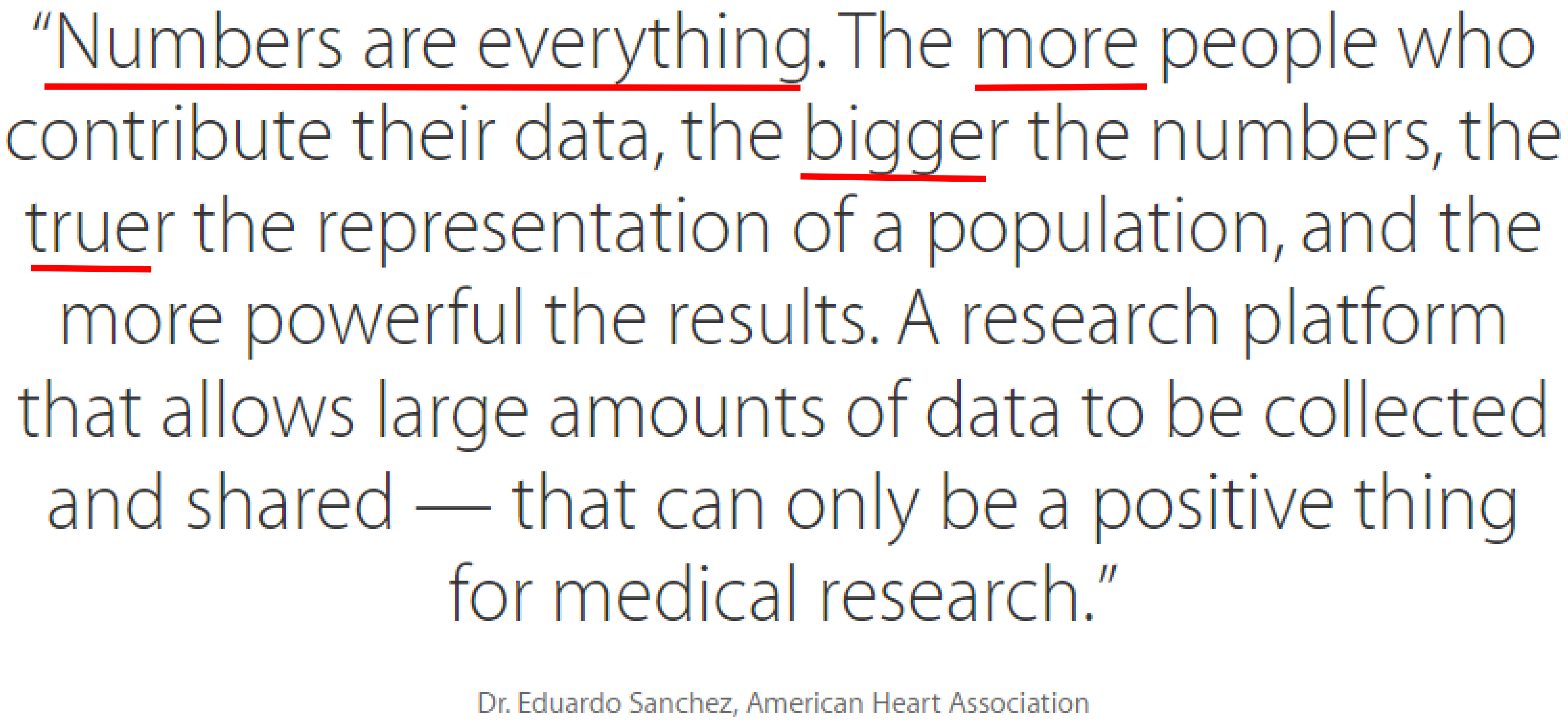
[ITmedia]

　慶應義塾大学医学部の研究チームは11月25日、iPhoneやApple Watchのセンサーを活用した臨床研究を国内で初めて開始したと発表した。専用アプリを通じて心拍数や運動能力などを測定し、不整脈・脳梗塞の早期発見につなげるという。iPhoneユーザーであれば、誰でも匿名で参加できる。

慶應義塾大学病院
Keio University Hospital
Heart & Brain
参加する
SIMなし
11:52
慶應義塾大学病院
Keio University Hospital
同意
説明と同意
研究参加への同意
研究: 1-5の研究項目を完了してください
1.ヘルスケアデータ収集
ヘルスケアアプリケーションのデータを収集します
2.質問票
不整脈・脳梗塞に関するアンケート調査をします
3.運動麻痺評価
両腕の運動麻痺の有無をチェックします
4.顔面運動評価
顔面の運動麻痺の有無をチェックします
5.協調運動評価
小脳の失調の有無をチェックします
オプション
「いま動悸がする」
動悸時にご利用ください。動悸の記録をとります。
研究
データ

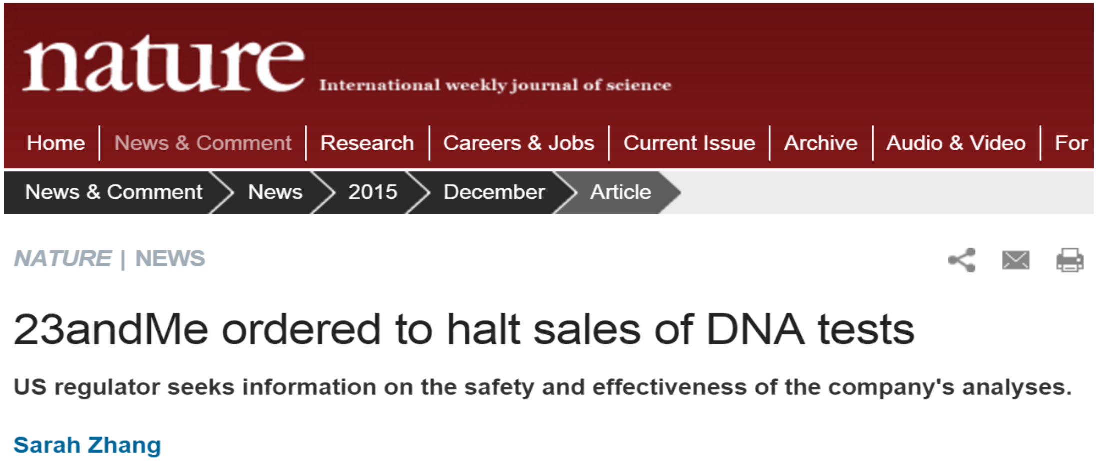
専用アプリ「Heart & Brain」を開発

Apple Announced ResearchKit on 4/14/2015



"Numbers are everything. The more people who contribute their data, the bigger the numbers, the truer the representation of a population, and the more powerful the results. A research platform that allows large amounts of data to be collected and shared — that can only be a positive thing for medical research."

Dr. Eduardo Sanchez, American Heart Association



NATURE | NEWS

23andMe ordered to halt sales of DNA tests

US regulator seeks information on the safety and effectiveness of the company's analyses.

Sarah Zhang

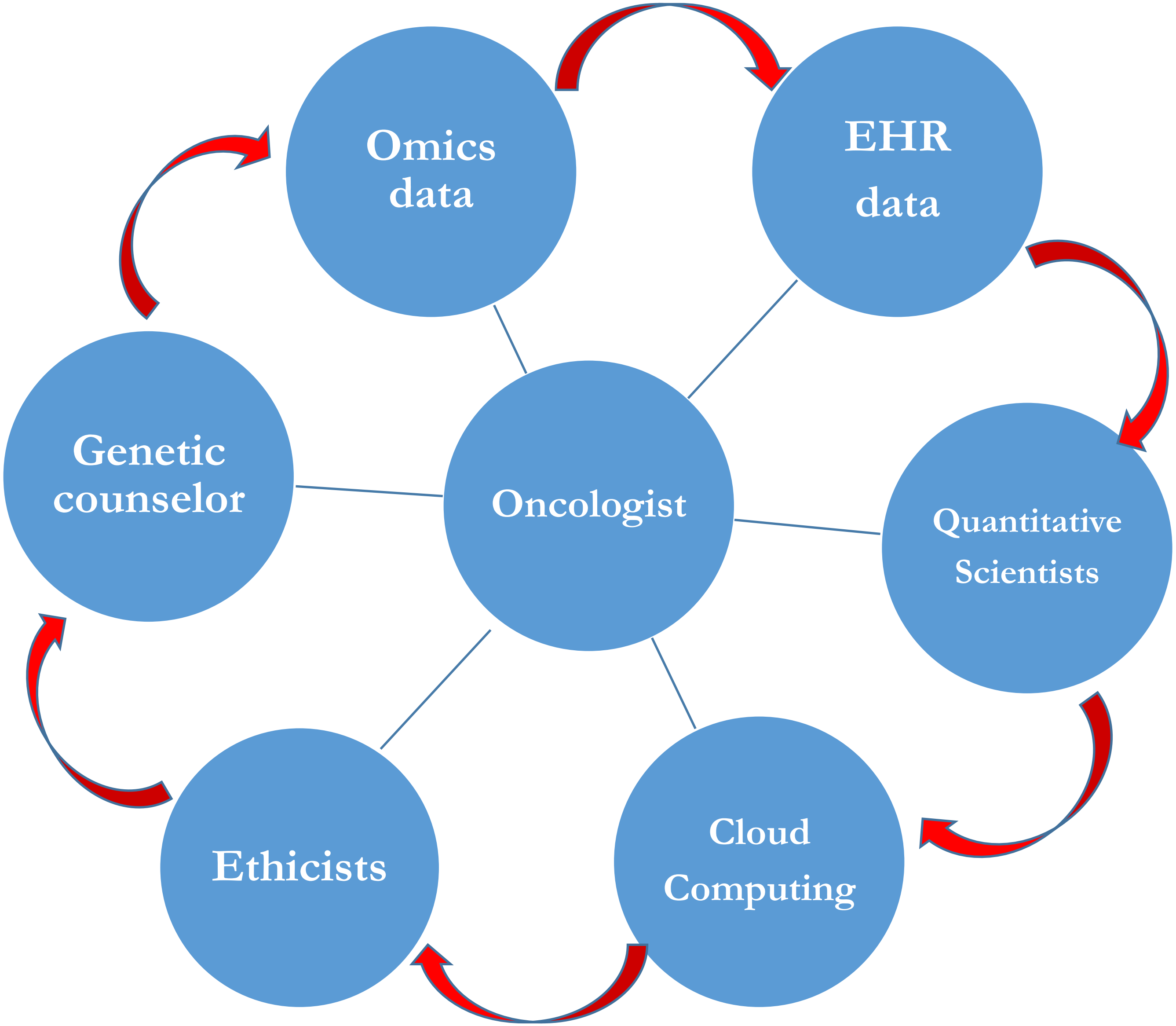
Nov 25, 2013

- After 14 face-to-face and teleconference meetings, hundreds of email exchanges, and dozens of written communications, FDA orders 23andMe to halt sales of DNA test kit.
- 23andMe Uses a sample of a user's saliva
- Claims to identify up to 254 diseases and other medical conditions
- 23andMe has submitted to the FDA twice for review as a medical device but has failed to satisfy the FDA's concerns

23andMe is back in the genetic testing business with FDA approval Oct, 2015

- In 2013: Assessments on 254 diseases and conditions for $99
- In 2015: Assessments on 36 inherited disorders for $199
- More than one million subscribers in the company's database
- The company also just announced a new drug discovery and development venture.

Cancer Precision Medicine



Omics data

EHR data

Quantitative Scientists

Cloud Computing

Ethicists

Genetic counselor

Oncologist

Spotlight

ARTWORK Tamar Cohen, *Andrew J Buboltz*
2011, silk screen on a page from a high school
yearbook, 8.5" x 12"

Data Scientist:

The Sexiest Job of the 21st Century

A new role is fast gaining prominence in organizations: that of the data scientist. Data scientists are the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information. As companies rush to capitalize on the potential of big data, the largest constraint many face is the scarcity of this special talent.

The shortage of data scientists is becoming a serious constraint in some sectors.

Data scientists today are akin to the Wall Street "quants" of the 1980s and 1990s.

Ben Chams - Fotolia

NGS – Data Analysis

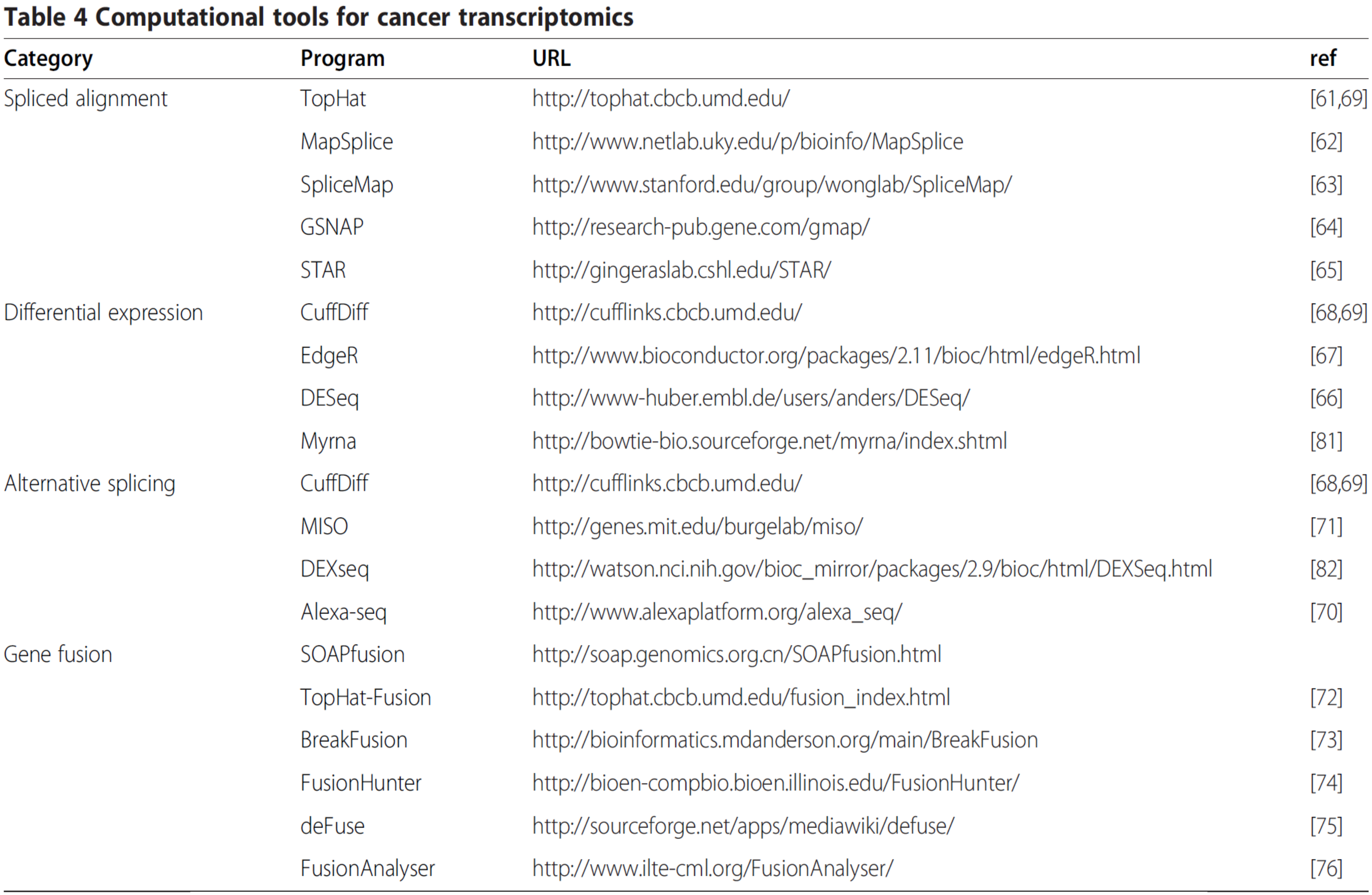
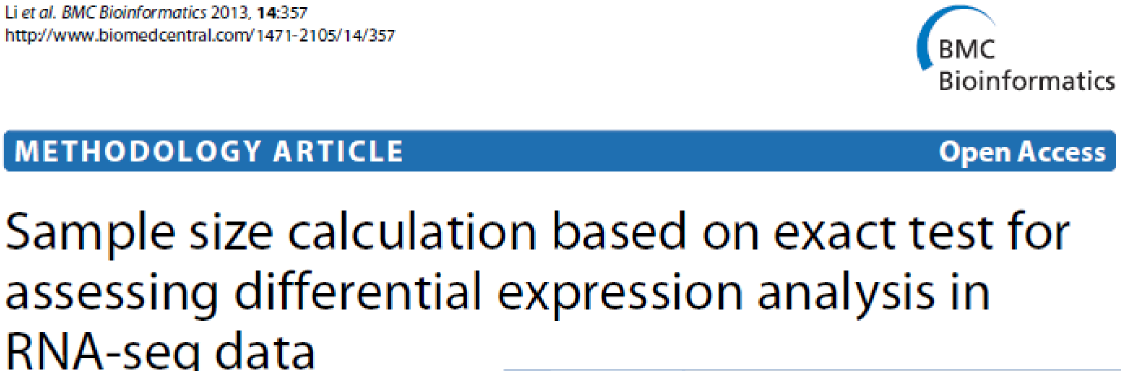


Table 4 Computational tools for cancer transcriptomics

Category	Program	URL	ref
Spliced alignment	TopHat	http://tophat.cbcb.umd.edu/	[61,69]
	MapSplice	http://www.netlab.uky.edu/p/bioinfo/MapSplice	[62]
	SpliceMap	http://www.stanford.edu/group/wonglab/SpliceMap/	[63]
	GSNAP	http://research-pub.gene.com/gmap/	[64]
	STAR	http://gingeraslab.cshl.edu/STAR/	[65]
Differential expression	CuffDiff	http://cufflinks.cbcb.umd.edu/	[68,69]
	EdgeR	http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html	[67]
	DESeq	http://www-huber.embl.de/users/anders/DESeq/	[66]
	Myrna	http://bowtie-bio.sourceforge.net/myrna/index.shtml	[81]
Alternative splicing	CuffDiff	http://cufflinks.cbcb.umd.edu/	[68,69]
	MISO	http://genes.mit.edu/burgelab/miso/	[71]
	DEXseq	http://watson.nci.nih.gov/bioc_mirror/packages/2.9/bioc/html/DEXSeq.html	[82]
	Alexa-seq	http://www.alexaplatform.org/alexa_seq/	[70]
Gene fusion	SOAPfusion	http://soap.genomics.org.cn/SOAPfusion.html	
	TopHat-Fusion	http://tophat.cbcb.umd.edu/fusion_index.html	[72]
	BreakFusion	http://bioinformatics.mdanderson.org/main/BreakFusion	[73]
	FusionHunter	http://bioen-compbio.bioen.illinois.edu/FusionHunter/	[74]
	deFuse	http://sourceforge.net/apps/mediawiki/defuse/	[75]
	FusionAnalyser	http://www.ilte-cml.org/FusionAnalyser/	[76]

Shyr and Liu *Biological Procedures Online* 2013, **15**:4
http://www.biologicalproceduresonline.com/content/15/1/4



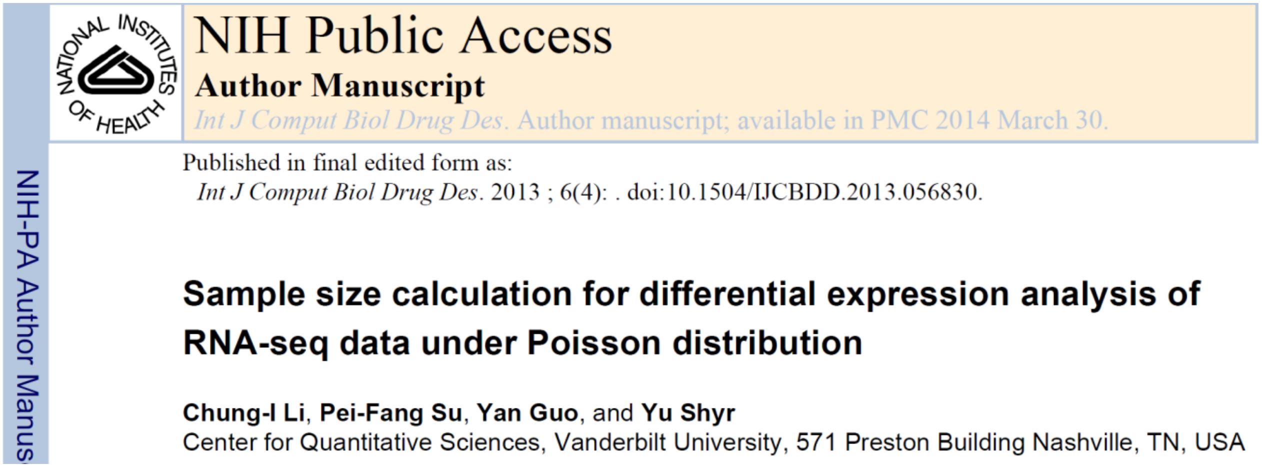
Li *et al. BMC Bioinformatics* 2013, **14**:357
http://www.biomedcentral.com/1471-2105/14/357

BMC Bioinformatics

METHODOLOGY ARTICLE **Open Access**

Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data

Chung-I Li[1,3], Pei-Fang Su[2,3] and Yu Shyr[3*]



NIH-PA Author Manus

NIH Public Access
Author Manuscript
Int J Comput Biol Drug Des. Author manuscript; available in PMC 2014 March 30.

Published in final edited form as:
Int J Comput Biol Drug Des. 2013 ; 6(4): . doi:10.1504/IJCBDD.2013.056830.

Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution

Chung-I Li, **Pei-Fang Su**, **Yan Guo**, and **Yu Shyr**
Center for Quantitative Sciences, Vanderbilt University, 571 Preston Building Nashville, TN, USA

Biometrics & Biostatistics

Shyr and Li, J Biomet Biostat 2014, 5:3
http://dx.doi.org/10.472/2155-6180.1000198

Research Article **Open Access**

Sample Size Calculation of RNA-sequencing Experiment-A Simulation-Based Approach of TCGA Data

Biometrics & Biostatistics

Shyr and Li, J Biomet Biostat 2014, 5:3
http://dx.doi.org/10.472/2155-6180.1000198

Research Article

Open Access

Sample Size Calculation of RNA-sequencing Experiment-A Simulation-Based Approach of TCGA Data

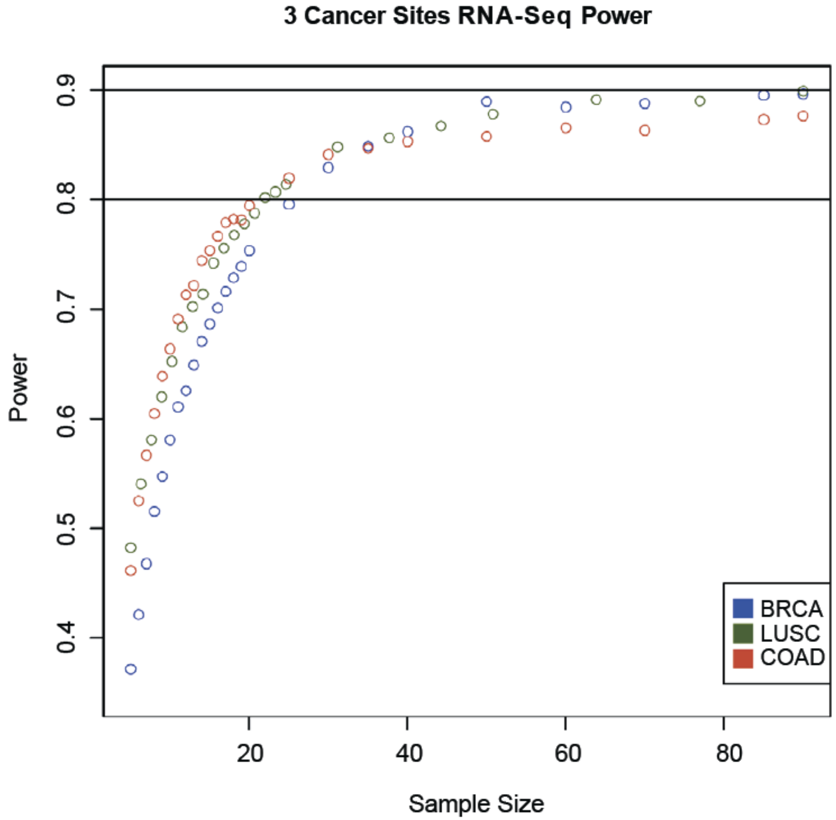
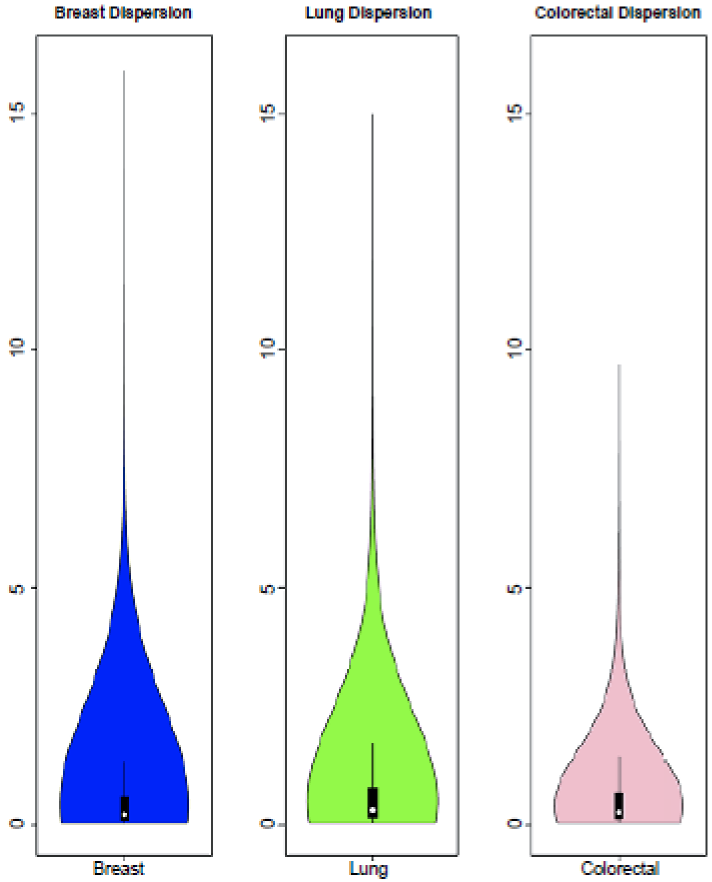
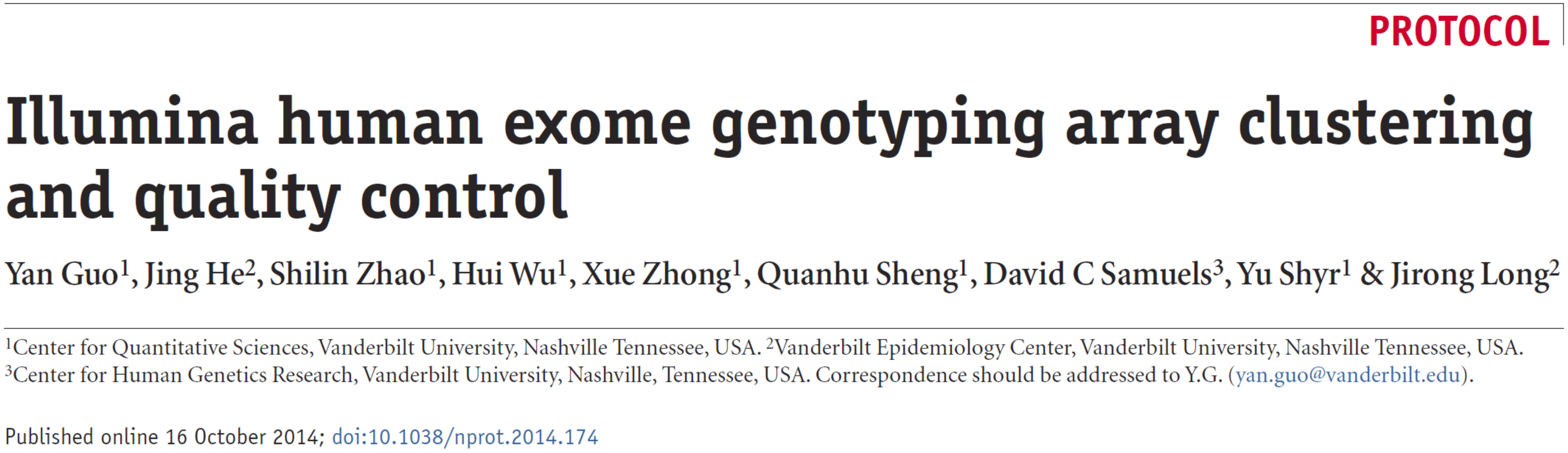


Figure 3: Scatterplot of the samples size and power (minimum reads=5 and FDR=0.05).





PROTOCOL

Illumina human exome genotyping array clustering and quality control

Yan Guo[1], Jing He[2], Shilin Zhao[1], Hui Wu[1], Xue Zhong[1], Quanhu Sheng[1], David C Samuels[3], Yu Shyr[1] & Jirong Long[2]

[1]Center for Quantitative Sciences, Vanderbilt University, Nashville Tennessee, USA. [2]Vanderbilt Epidemiology Center, Vanderbilt University, Nashville Tennessee, USA. [3]Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee, USA. Correspondence should be addressed to Y.G. (yan.guo@vanderbilt.edu).

Published online 16 October 2014; doi:10.1038/nprot.2014.174

GenomeStudio section

- Loading the data into GenomeStudio
- Performing automatic clustering
- QC on SNPs located in a haploid genome
- QC based on GenTrain score
- QC based on cluster separation
- QC based on Mendelian error and replication error
- QC based on other criteria
- Calling rare SNPs
- Final filtering
- Exporting from GenomeStudio

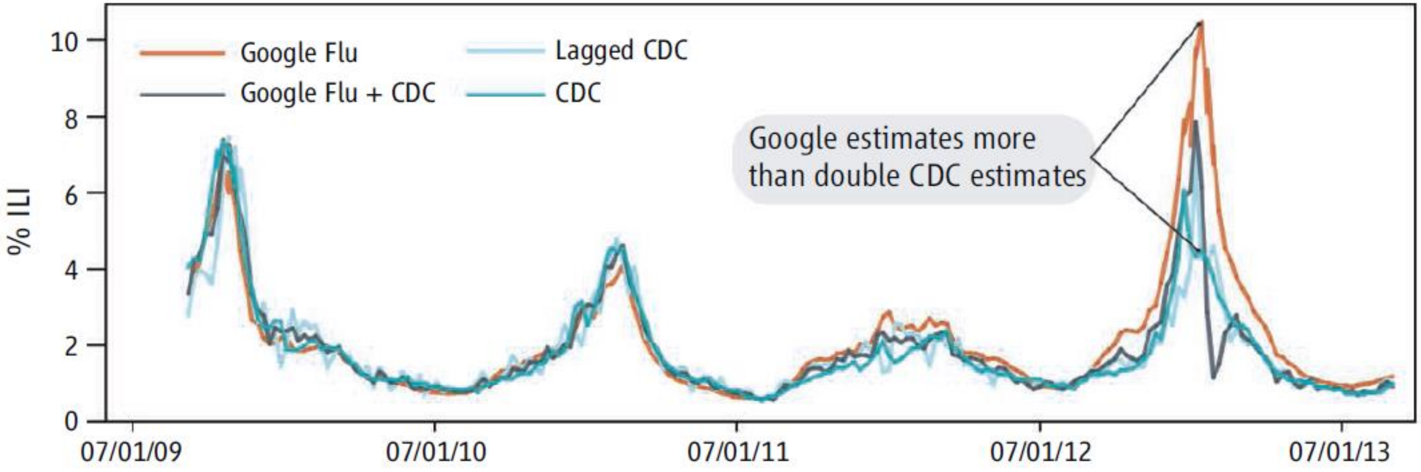
Post-GenomeStudio section

- Converting all SNPs to the forward strand
- Checking for gender mismatch
- Checking for race mismatch
- Checking for relatedness
- Checking for Hardy-Weinberg equilibrium (HWE) outliers
- Checking for heterozygosity outliers
- Checking consistency between exome chip genotype and 1000 Genomes Project[17] or HapMap[18] genotype
- Checking for minor allele frequency (MAF) consistency between exome chip and 1000 Genomes Project genotypes
- Checking for batch effects

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]



Google estimates more than double CDC estimates

- This should have been a warning that the big data were over-fitting the small number of cases—a standard concern in data analysis.

SCIENCE VOL 343 14 MARCH 2014

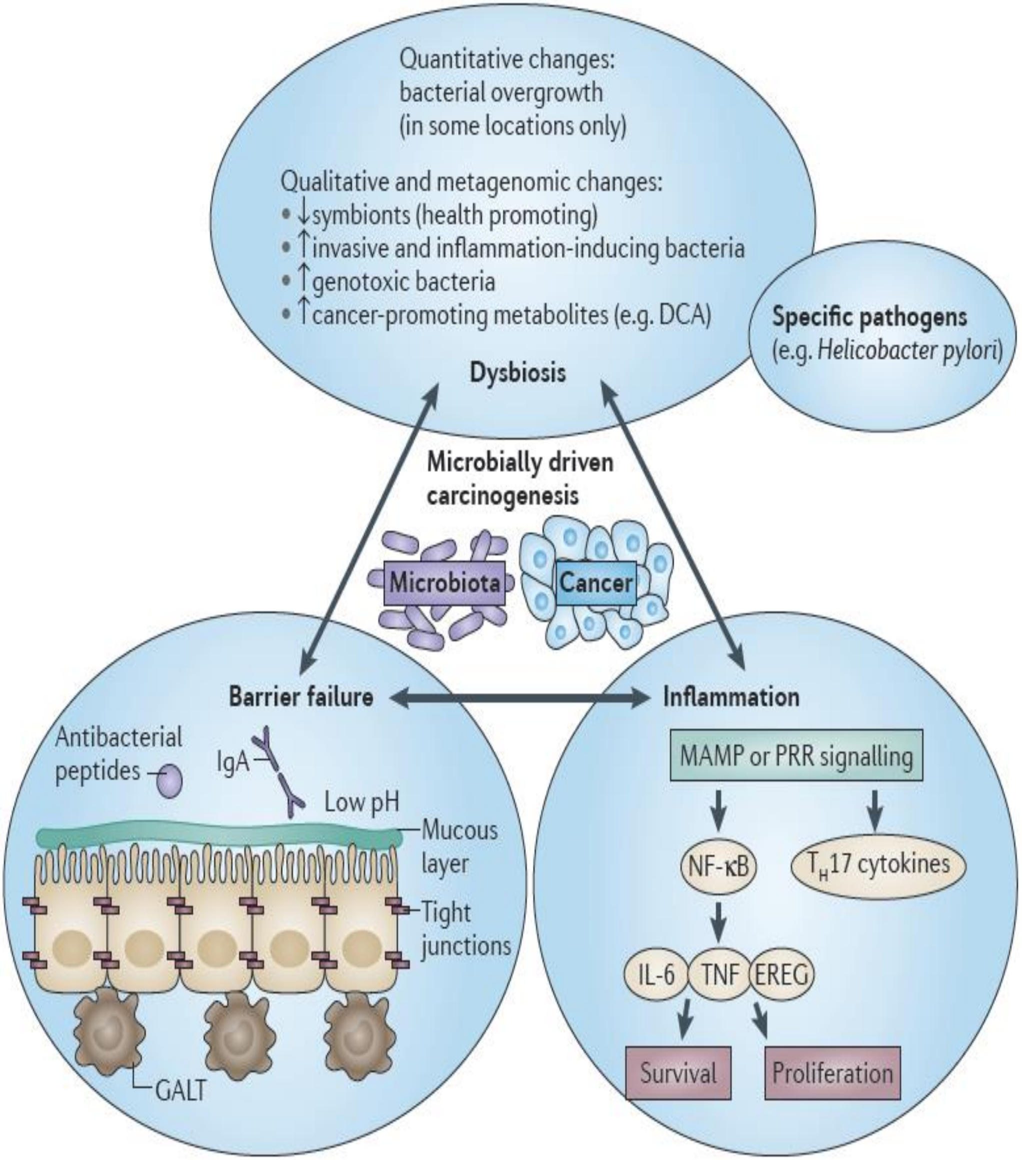
Microbiome and PheWAS

NATURE REVIEWS | **CANCER** VOLUME 13 | NOVEMBER 2013

The m

Robert F. Sc

Abstract | Mi
relationship
defects in th
homeostasis
(infection, di
disease. Incr
carcinogene
microbiota a
genotoxicity
prevention.



ch symbiotic
fe. However,
sensing and
ntal changes
nd promote
icrobiota in
he bacterial
es, dysbiosis,
e for cancer

Emerging roles of the microbiome

Research on the microbiome is an emerging science. Recent work suggests the benefits derived by humans from their microbiotas may have profound consequences for health.

Roles include

- Food digestion and nutrition
- Regulation of metabolism
- Processing and detoxifying environmental chemicals
- Development and regulation of the immune system
- Prevention of invasion and growth of pathogens
- Role in carcinogenesis: cancer susceptibility and progression
- Biological control of certain diseases and deficiencies

MICROBIOME

Microbes aid cancer drugs

Gut bacteria boost immunotherapies

6 NOVEMBER 2015 • VOL 350 ISSUE 6261

By **Mitch Leslie**

sciencemag.org SCIENCE

Sciencexpress Reports

Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota

Marie Vétizou,

5 November 2015 / Page 1 / 10.1126/science.aad1329

Sciencexpress Reports

Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy

Ayelet Sivan

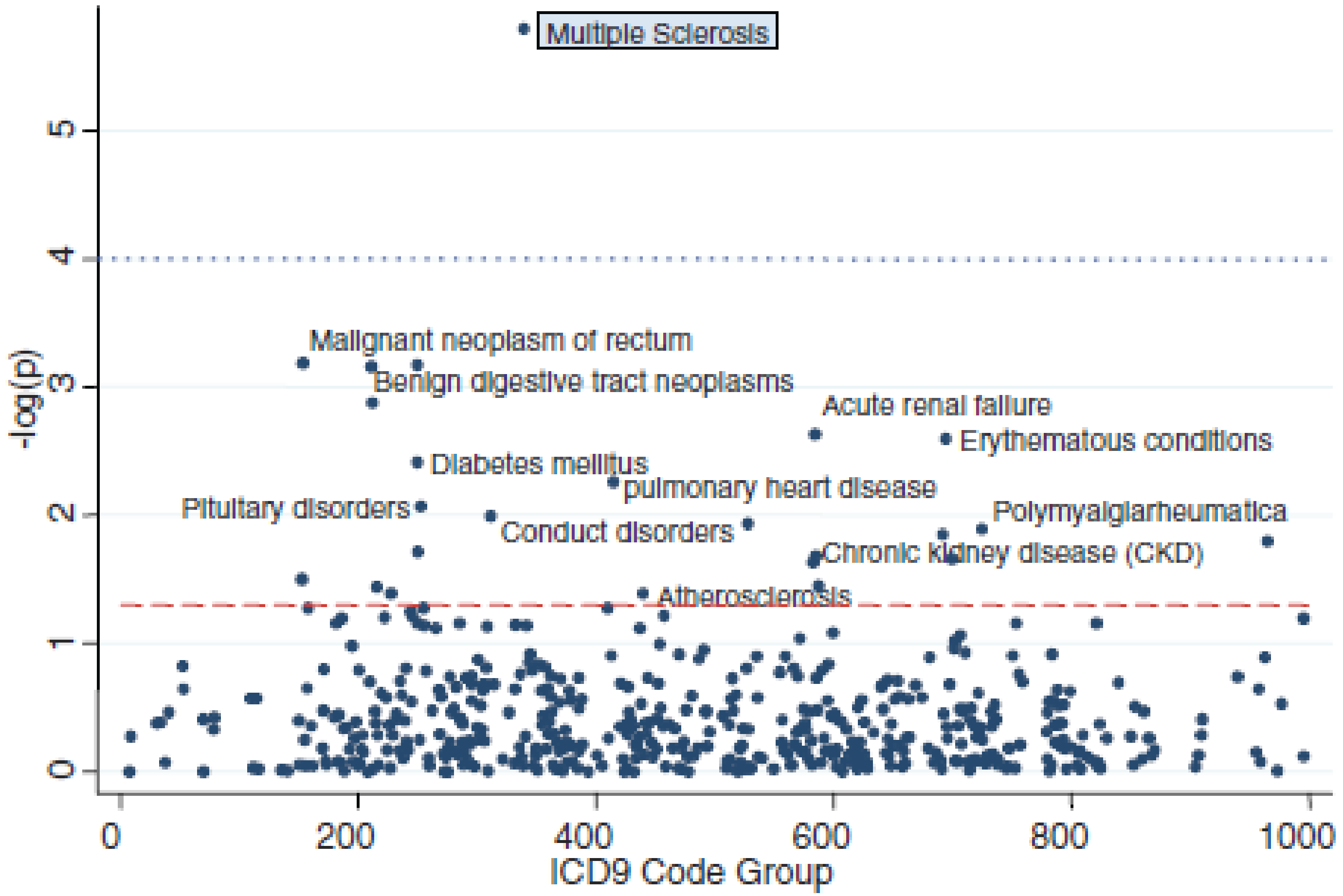
5 November 2015/ Page 1 / 10.1126/science.aac4255

PheWAS

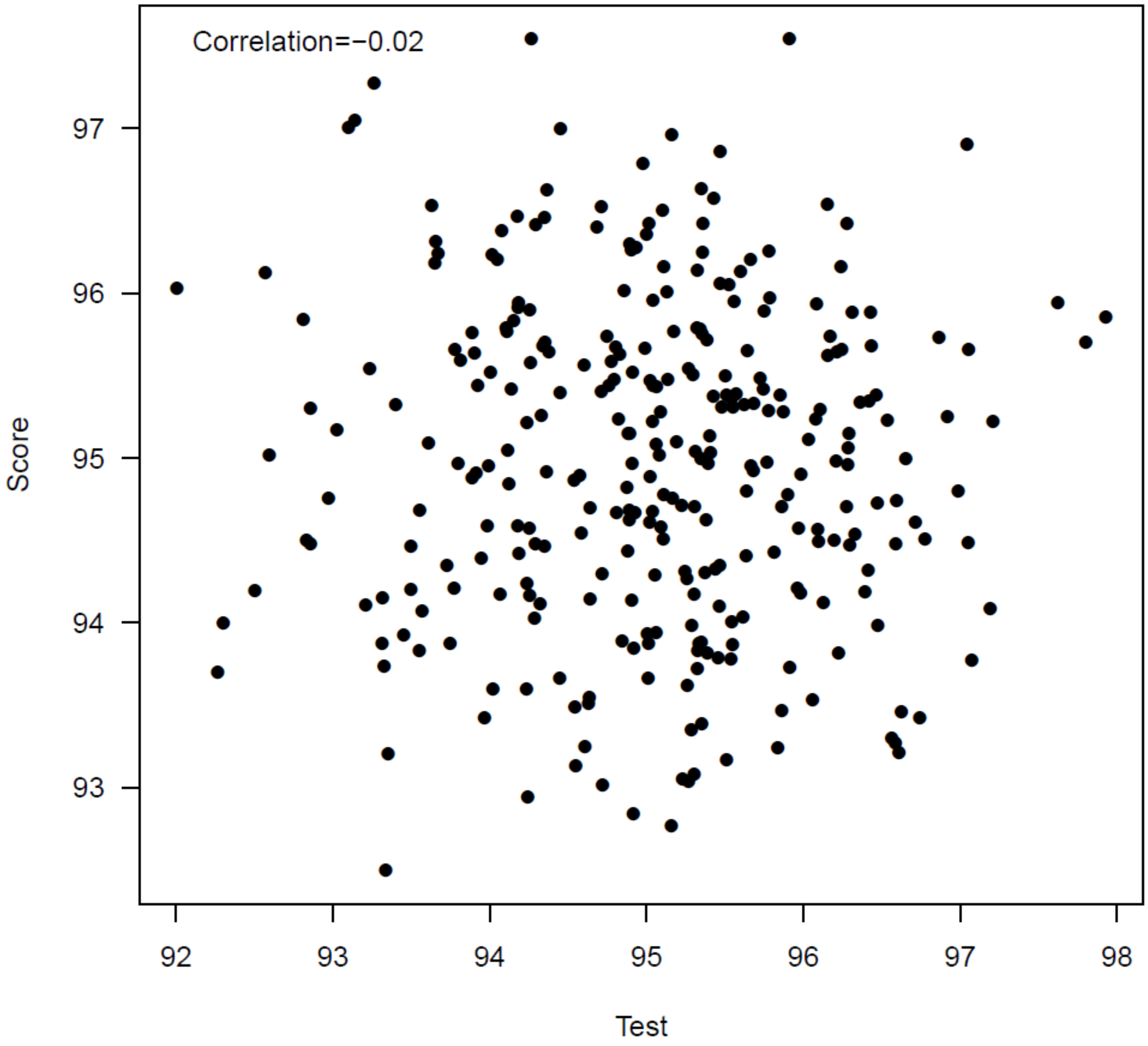
- PheWAS explores the association between a number of common genetic variations and a wide variety of phenotypes
- The combination of the extensive collection of de-identified medical records in the Synthetic Derivative and the genomic information in BioVU is ideal for PheWAS

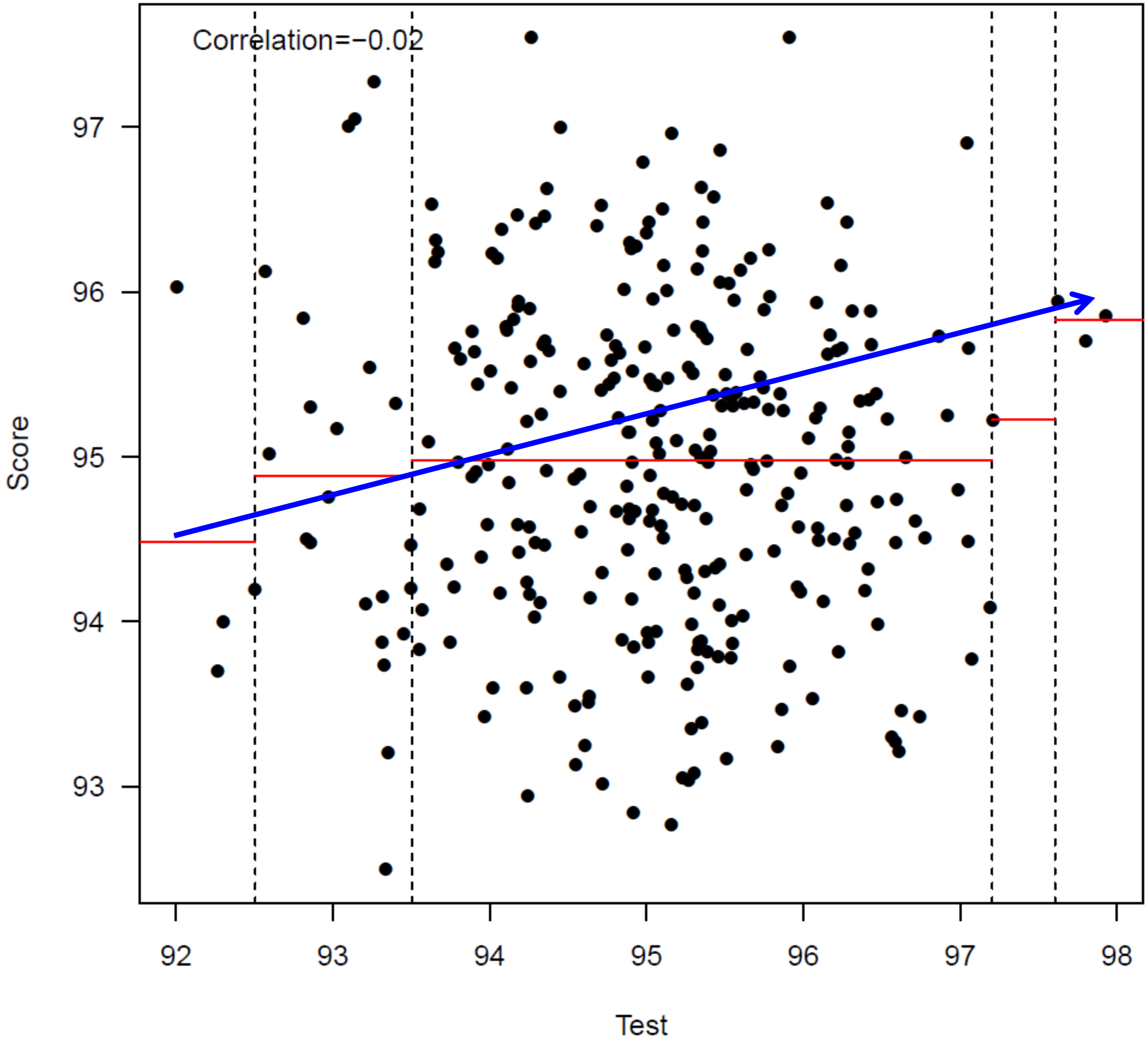
PheWAS

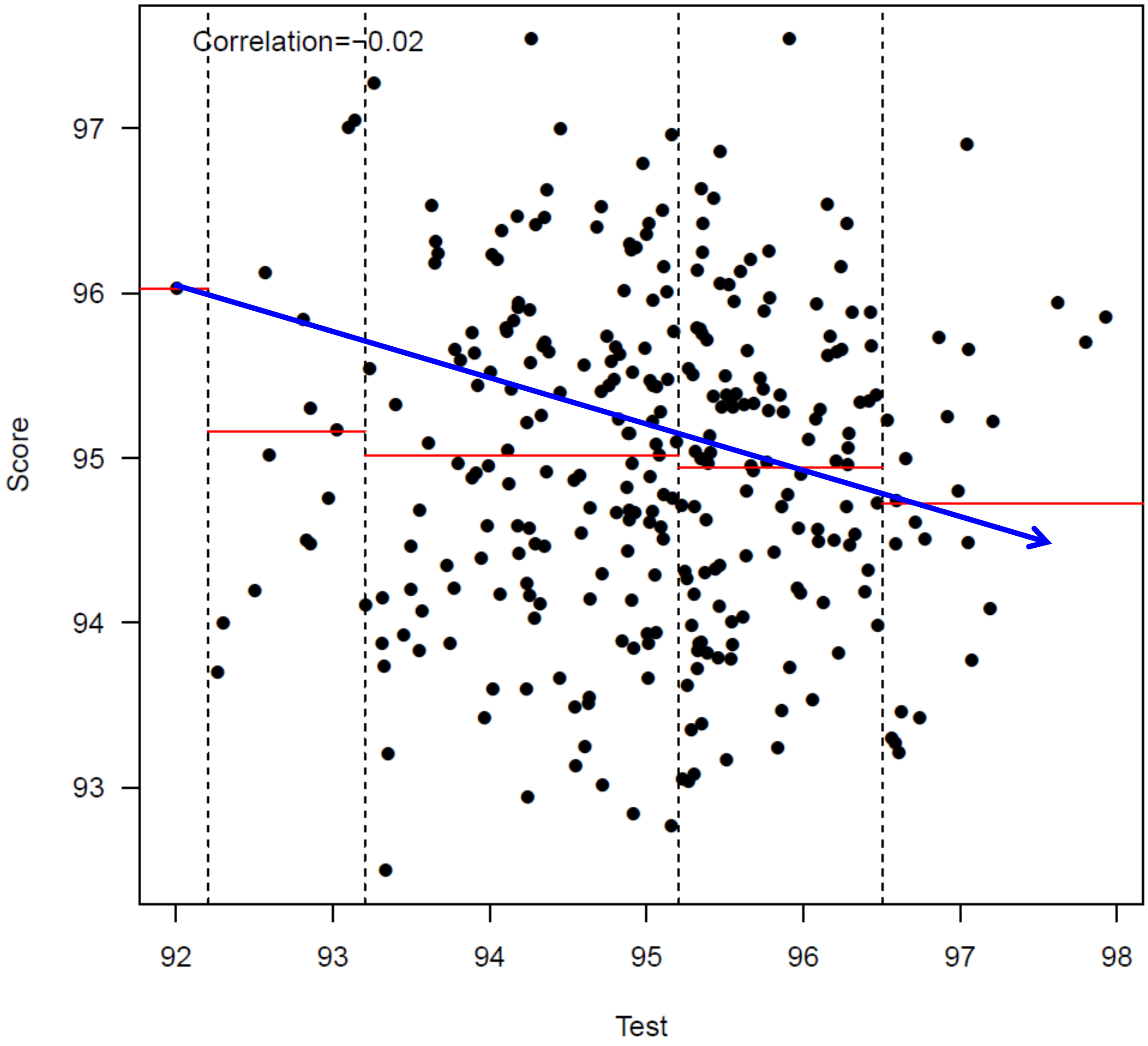
PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations

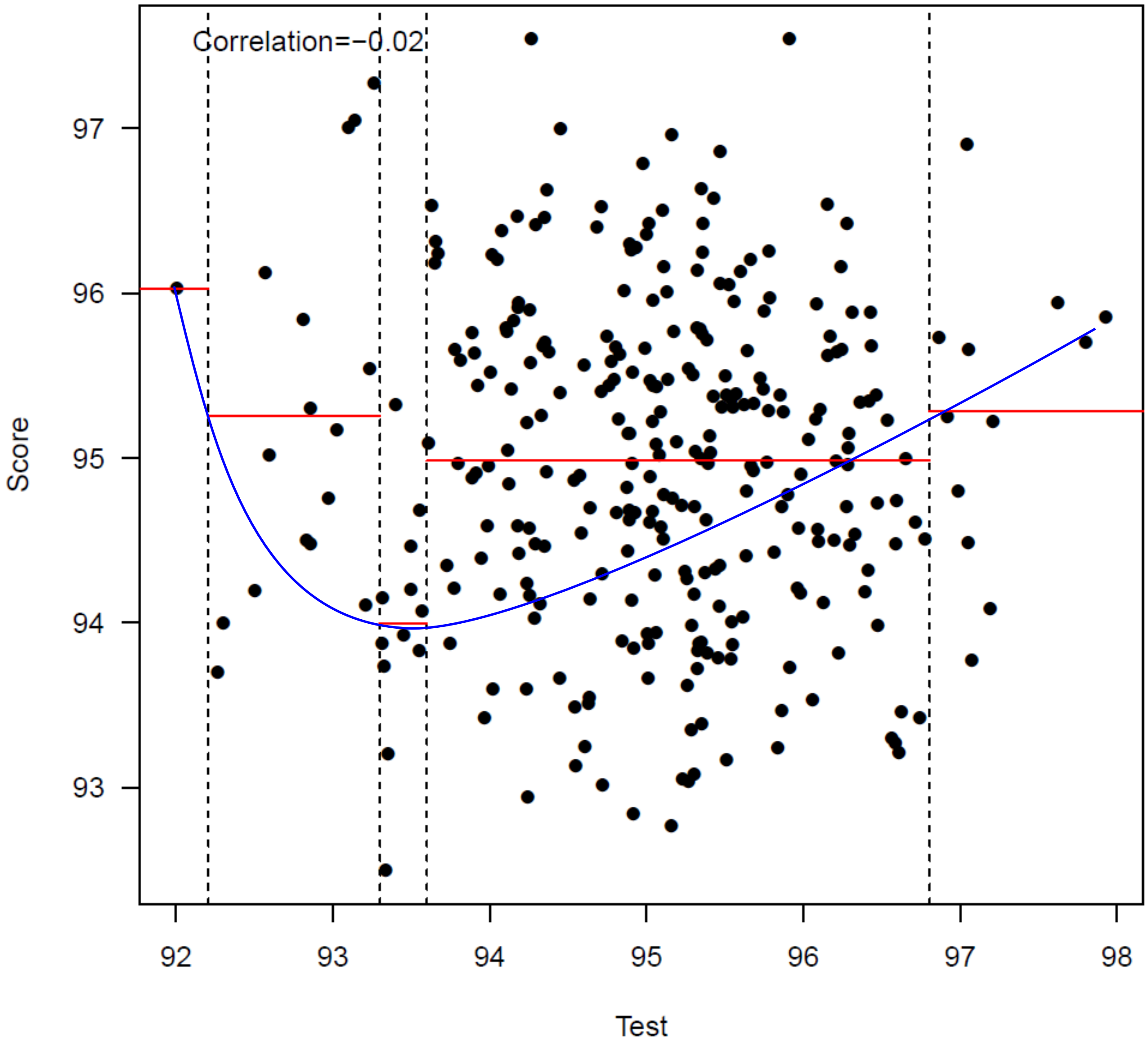


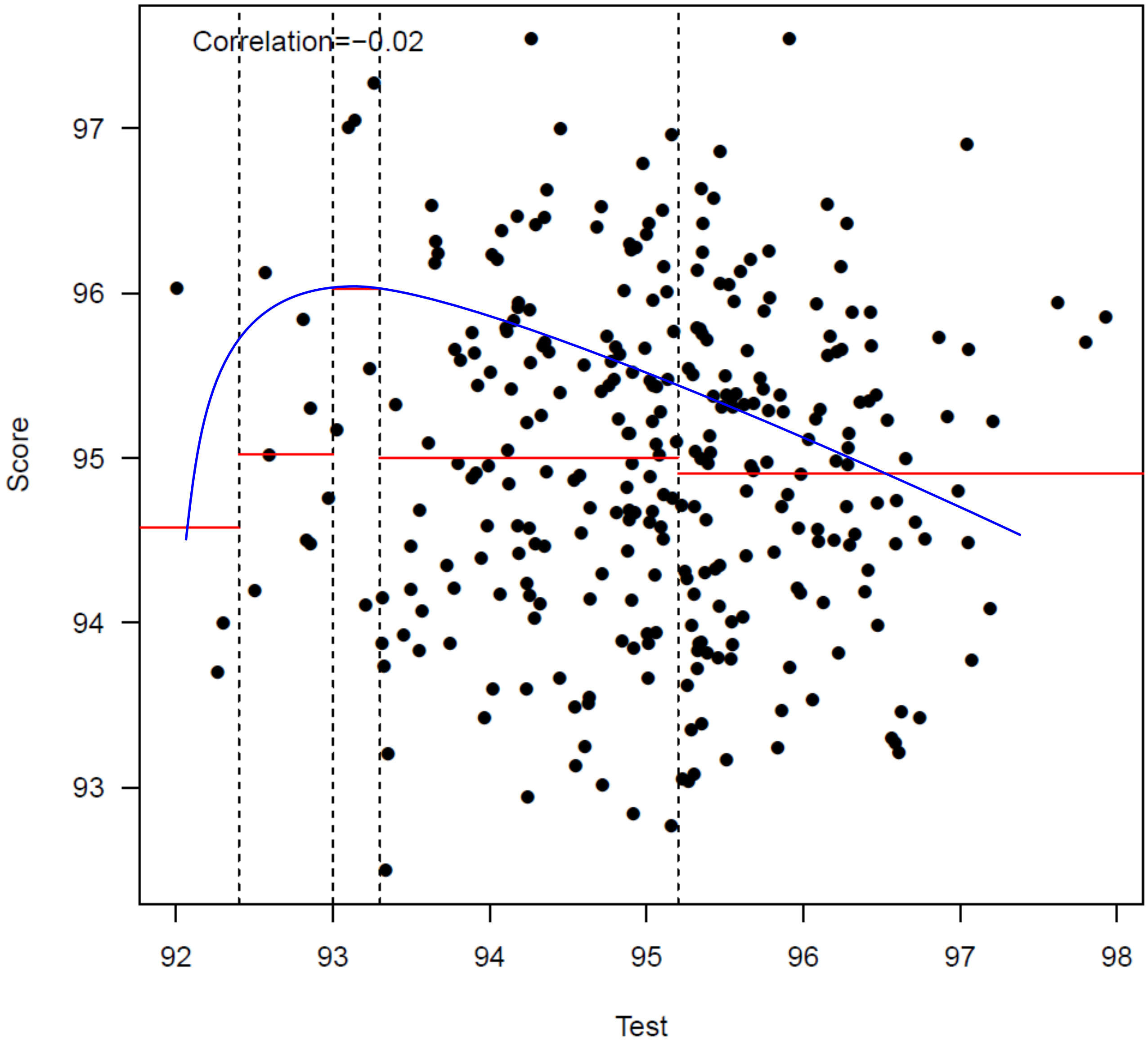
Bioinformatics *Vol. 26 no. 9 2010, pages 1205–1210* ***J.C.Denny et al.***


Correlation=−0.02
Score
97
96
95
94
93
Test
92
93
94
95
96
97
98


Correlation=−0.02
Score
Test
97
96
95
94
93
92
93
94
95
96
97
98


Correlation=−0.02
97
96
95
94
93
Score
92
93
94
95
96
97
98
Test


Correlation=−0.02
Score
97
96
95
94
93
Test
92
93
94
95
96
97
98


Correlation=−0.02
97
96
95
94
93
Score
92
93
94
95
96
97
98
Test

END

Questions