National Cancer Institute

DCEG Core Genotyping Facility

Stephen Chanock, M.D.

Chief, Laboratory of Translational Genomics
Director, Core Genotyping Facility

January 25, 2012

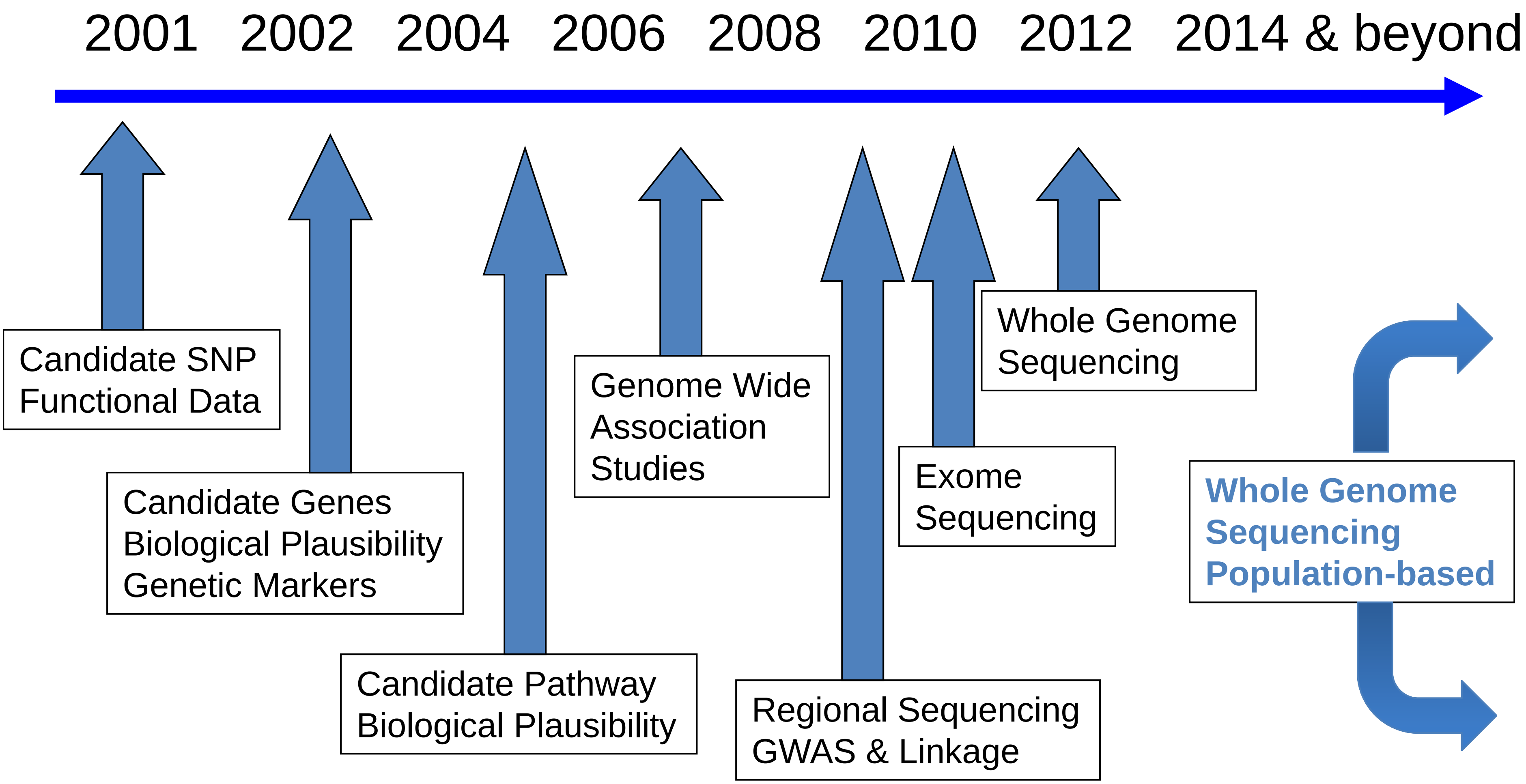
U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health

Mission of Core Genotyping Facility (CGF)

- Conduct of high quality molecular epidemiology studies
  - Emphasis on:
    - Germline contribution to risk
    - Gene-environment interactions
  - Transition to:
    - Germline/somatic interactions
    - Interaction of somatic alterations with environmental risk factors
- Education
    - Genetics analysis courses & seminars

Milestones at the Core Genotyping Facility



2001 2002 2004 2006 2008 2010 2012 2014 & beyond

Candidate SNP
Functional Data

Candidate Genes
Biological Plausibility
Genetic Markers

Candidate Pathway
Biological Plausibility

Genome Wide
Association
Studies

Regional Sequencing
GWAS & Linkage

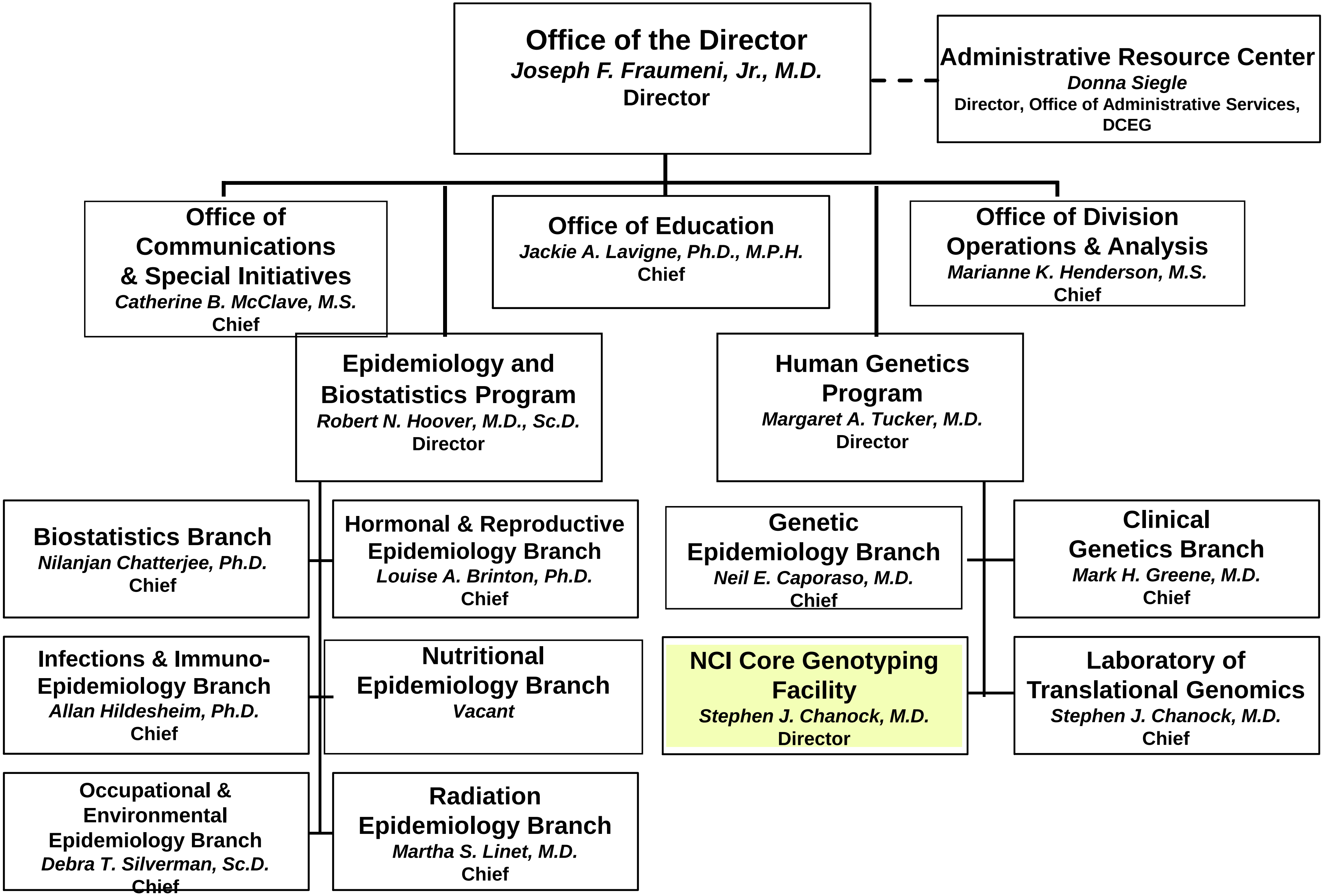
Exome
Sequencing

Whole Genome
Sequencing

**Whole Genome
Sequencing
Population-based**

NATIONAL CANCER INSTITUTE

Division of Cancer Epidemiology and Genetics



Office of the Director
Joseph F. Fraumeni, Jr., M.D.
Director

Administrative Resource Center
Donna Siegle
Director, Office of Administrative Services, DCEG

Office of Communications & Special Initiatives
Catherine B. McClave, M.S.
Chief

Office of Education
Jackie A. Lavigne, Ph.D., M.P.H.
Chief

Office of Division Operations & Analysis
Marianne K. Henderson, M.S.
Chief

Epidemiology and Biostatistics Program
Robert N. Hoover, M.D., Sc.D.
Director

Human Genetics Program
Margaret A. Tucker, M.D.
Director

Biostatistics Branch
Nilanjan Chatterjee, Ph.D.
Chief

Hormonal & Reproductive Epidemiology Branch
Louise A. Brinton, Ph.D.
Chief

Infections & Immuno-Epidemiology Branch
Allan Hildesheim, Ph.D.
Chief

Nutritional Epidemiology Branch
Vacant

Occupational & Environmental Epidemiology Branch
Debra T. Silverman, Sc.D.
Chief

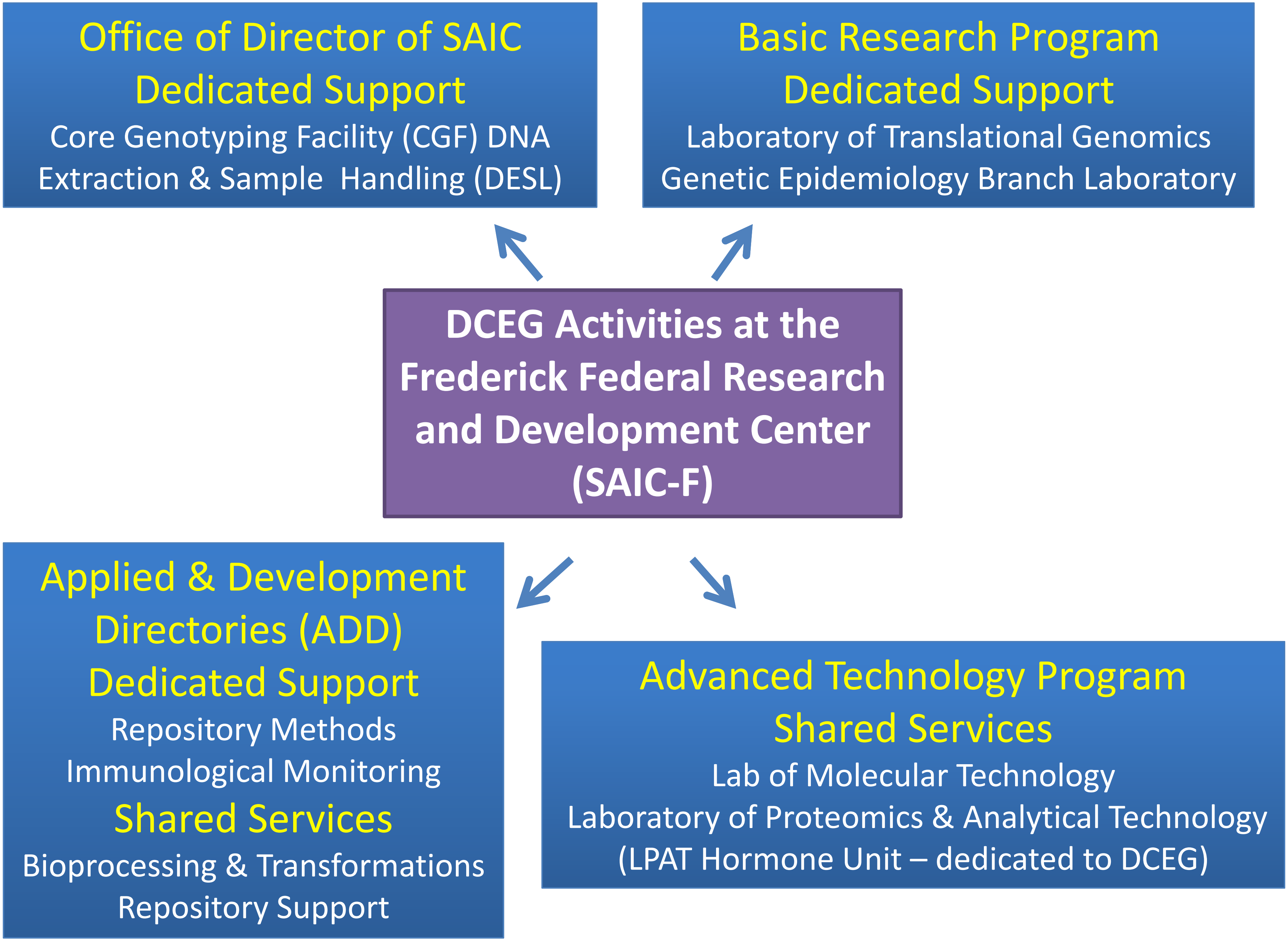
Radiation Epidemiology Branch
Martha S. Linet, M.D.
Chief

Genetic Epidemiology Branch
Neil E. Caporaso, M.D.
Chief

Clinical Genetics Branch
Mark H. Greene, M.D.
Chief

NCI Core Genotyping Facility
Stephen J. Chanock, M.D.
Director

Laboratory of Translational Genomics
Stephen J. Chanock, M.D.
Chief



Office of Director of SAIC
Dedicated Support
Core Genotyping Facility (CGF) DNA Extraction & Sample Handling (DESL)

Basic Research Program
Dedicated Support
Laboratory of Translational Genomics
Genetic Epidemiology Branch Laboratory

DCEG Activities at the Frederick Federal Research and Development Center (SAIC-F)

Applied & Development Directories (ADD)
Dedicated Support
Repository Methods
Immunological Monitoring
Shared Services
Bioprocessing & Transformations
Repository Support

Advanced Technology Program
Shared Services
Lab of Molecular Technology
Laboratory of Proteomics & Analytical Technology
(LPAT Hormone Unit – dedicated to DCEG)

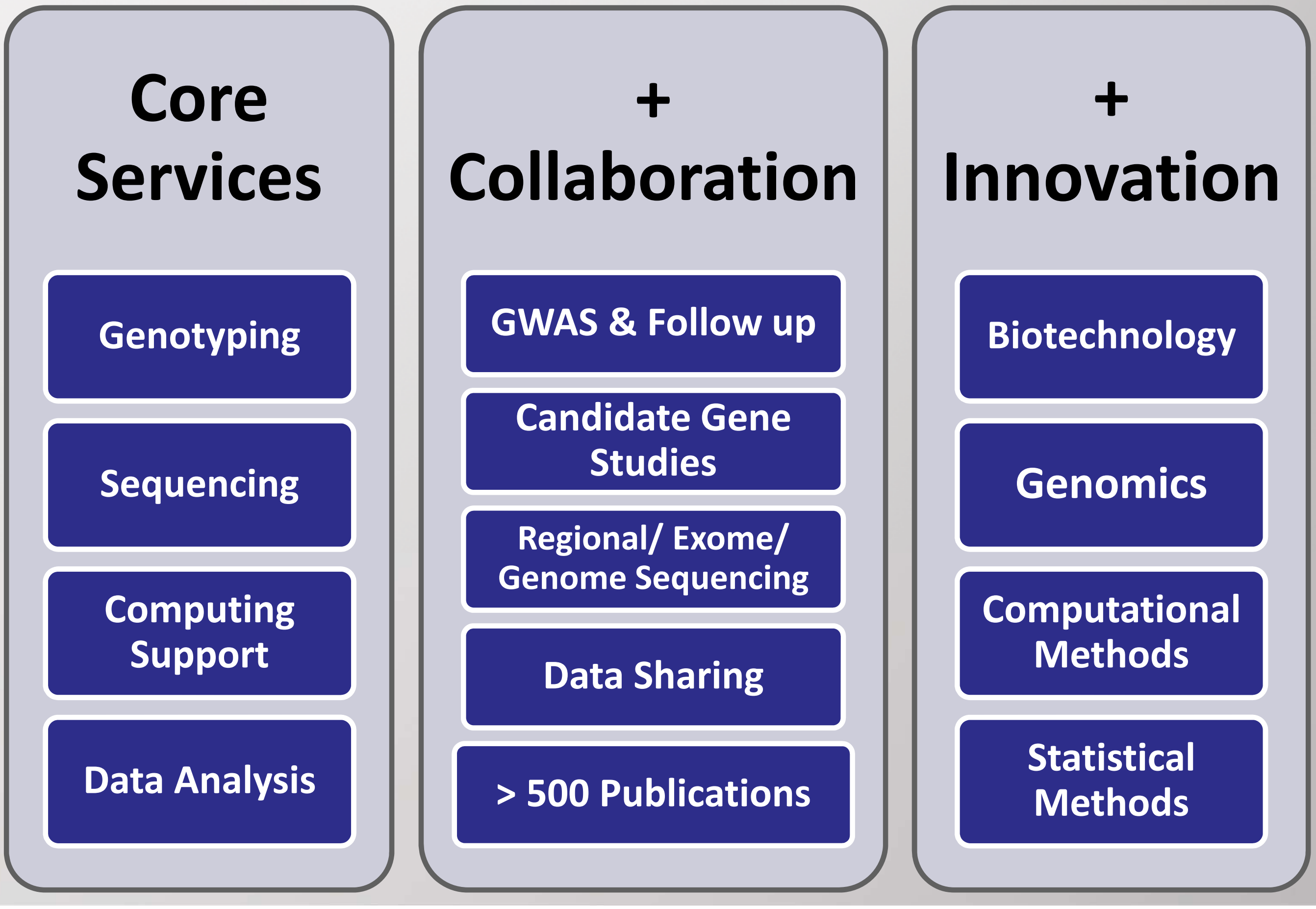
CGF Facilities Footprint

Advanced Technology Center: Gaithersburg

CORE GENOTYPING FACILITY

The Core Genotyping Facility
Dedicated DCEG Facility

What's in a name? Core Plus Plus



Core Services

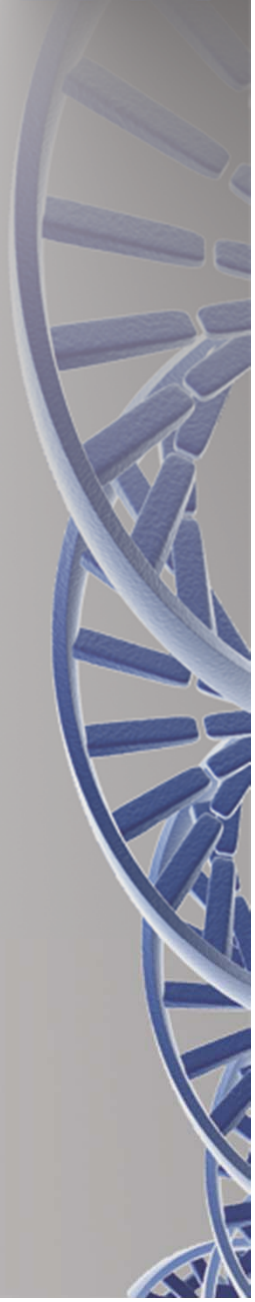
- Genotyping
- Sequencing
- Computing Support
- Data Analysis

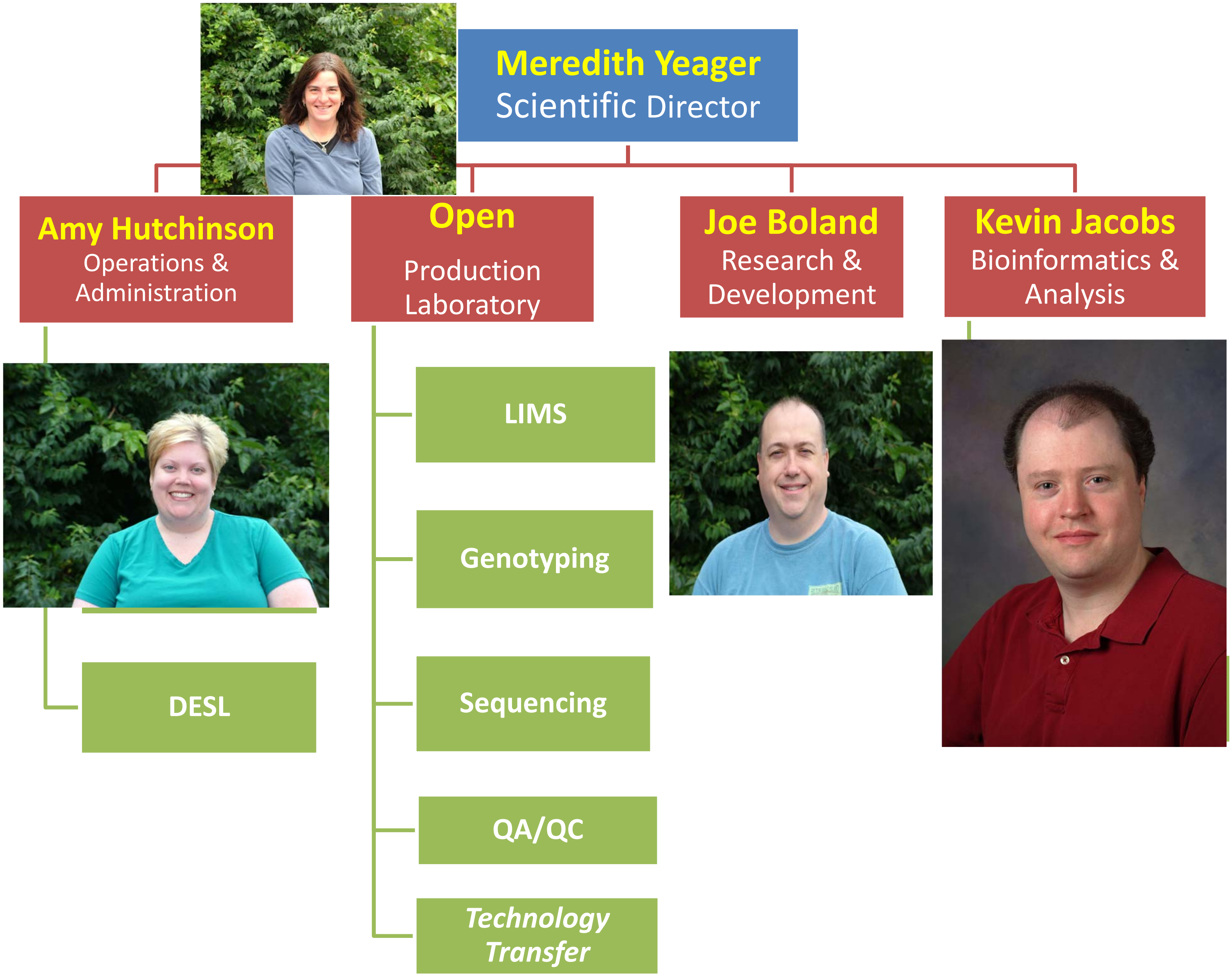
+ Collaboration

- GWAS & Follow up
- Candidate Gene Studies
- Regional/ Exome/ Genome Sequencing
- Data Sharing
- > 500 Publications

+ Innovation

- Biotechnology
- Genomics
- Computational Methods
- Statistical Methods





Meredith Yeager
Scientific Director

Amy Hutchinson
Operations &
Administration

- DESL

Open
Production
Laboratory

- LIMS
- Genotyping
- Sequencing
- QA/QC
- *Technology Transfer*

Joe Boland
Research &
Development

Kevin Jacobs
Bioinformatics &
Analysis

Investigation of Alternatives

- DCEG Conducted Molecular Epidemiology Pilot Study 2001-2003
  - 5 Companies asked to produce defined data sets
  - Common issues
    - Slow
    - Costly
    - Poor performance with QC
- Periodic reassessment of contract work
  - Loss of scientific ownership
  - Variability in deliverables

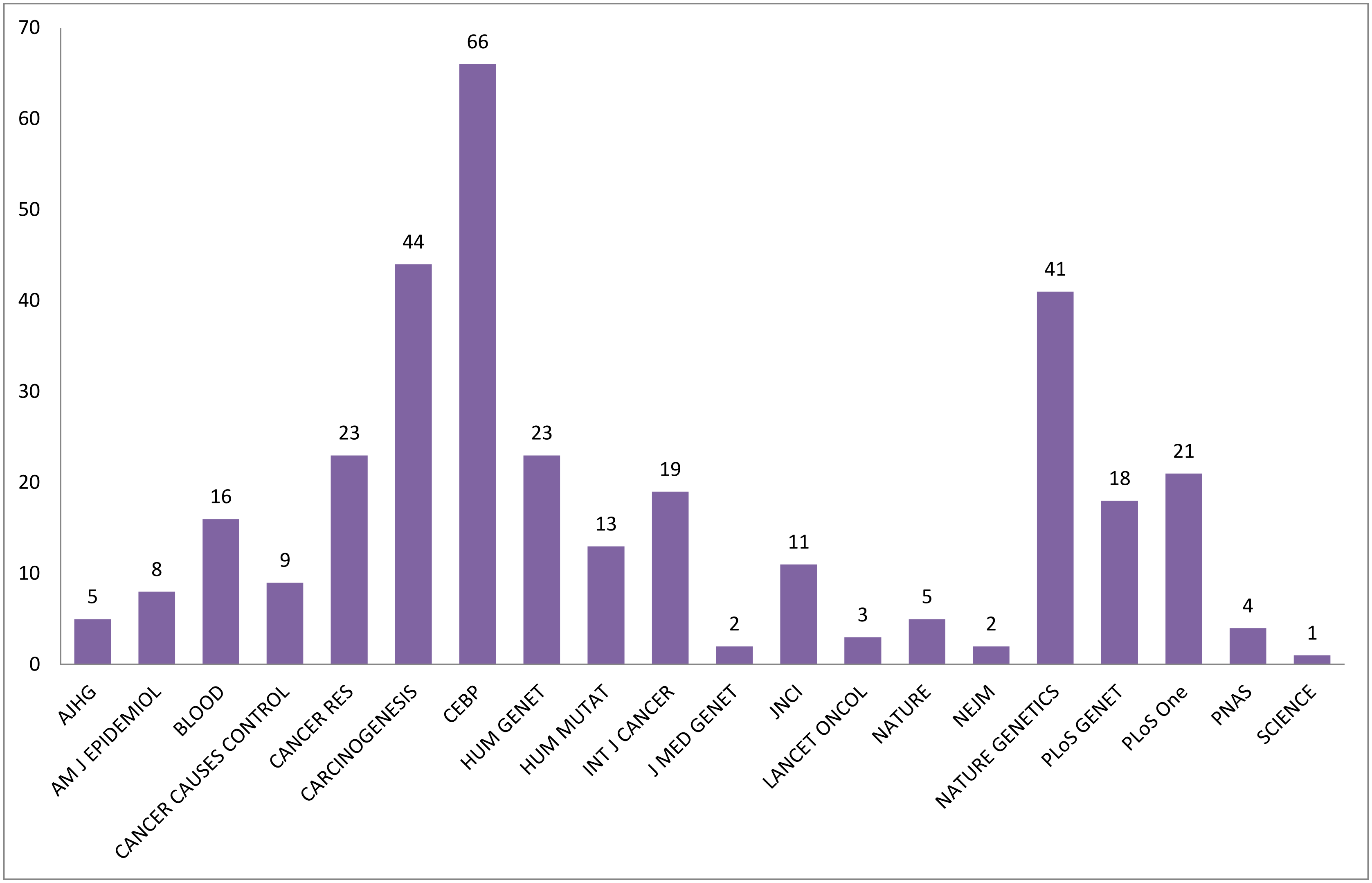
Value of creating CGF within FFRDC

- Close collaboration between NCI investigators and SAIC-F experts
- NCI can monitor every step and assess capacity to meet milestones
- Opportunity to drive scientific challenges in partnership
  - *Bridging Epidemiology and Genetics*

Nimble Personnel Structure

- Reorganization began with 9 SAIC FTEs
  - Reorganization and expansion 2002-2006
  - CGEMS funding for 5 additional analysts
- Current FTEs: 42
  - Shift from wet to dry positions in last 3 years
- Establish expertise for genetic analysis
  - Avoid “blackbox/blackhole” of contract
- Embed NCI oversight within SAIC work flow
  - Daily- no…… hourly discussions

536 CGF Publications for 2002-2011



Review of DCEG Projects for CGF

- Proposals discussed and approved by Branch Chiefs prior to submission
- Varies by scope & cost
  - Senior Leadership for Genomics Committee (SLGC) provides concept review for
    - GWAS chips
    - Sequencing of Exome/Whole Genome
  - Genotype Review Committee (GRC)
    - All projects greater than $25,000

Senior Leadership for Genomics Committee (SLGC)

Mission

Review & Approval of

- GWAS chips
- Exome/WGS

Determines priority for Illumina Infinium

Data Sharing and Access Issues

Membership

J Fraumeni

P Tucker

R Hoover

P Hartge

S Chanock

M Henderson

Monthly Meetings with Minutes

Genotyping Review Committee (GRC)

Mission

Critique of Science

Statistical Review

Approval letter required to proceed to CGF queue

Minutes

Chair can approve small projects & revisions

Membership

Chair:

P Tucker, Director, HGP

PIs from each Branch

rotate every 2 years

S Chanock

K Pitt

CGF Review Processes

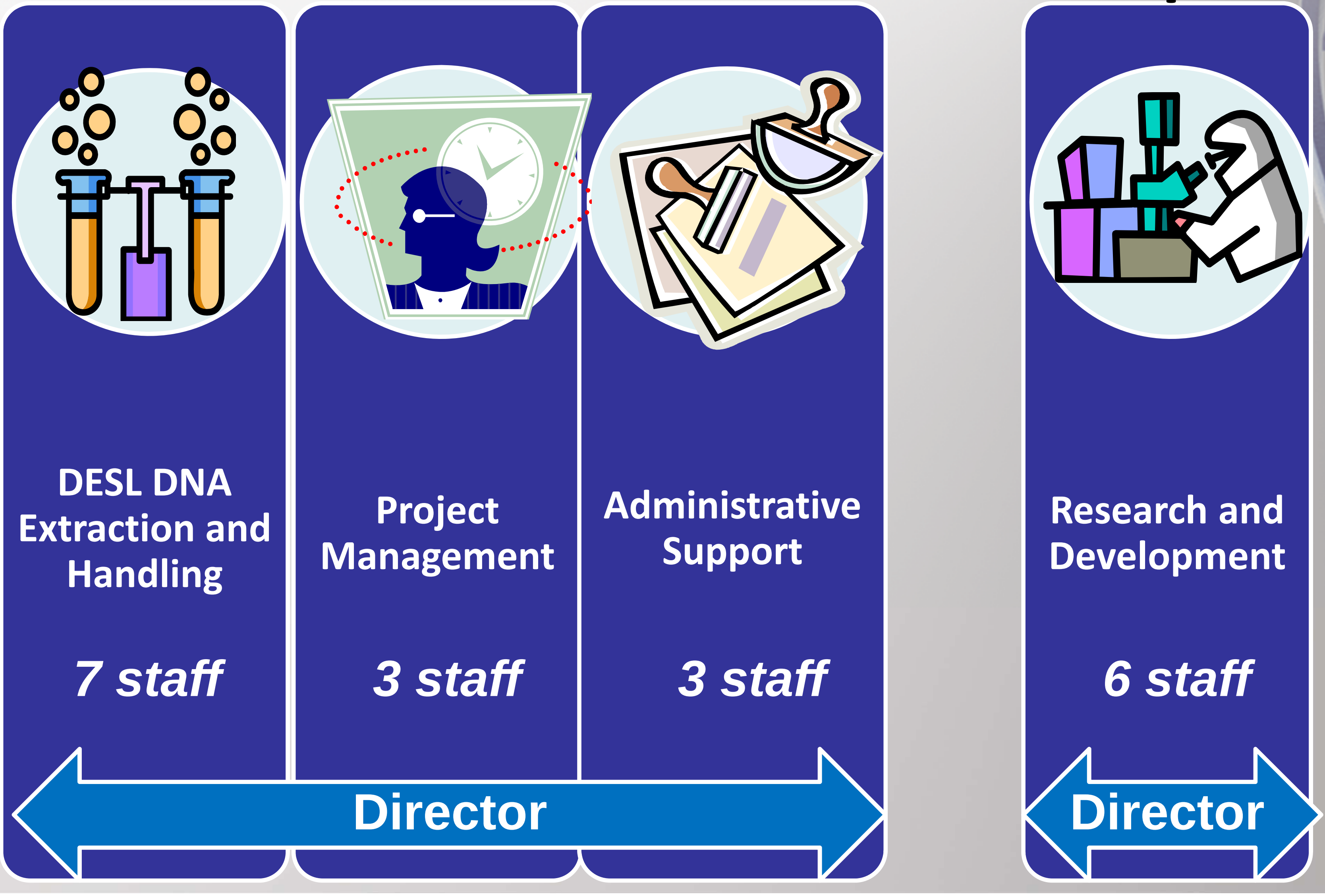
- Weekly conference
- Monthly SLGC meeting
- Quarterly SAIC report
- Biannual review of budget by OD DCEG
- Quadrennial Site Visit
  - May 2012 for CGF

Dedicated Facility Support

- DCEG directly supports
  - Personnel
  - Equipment
  - Maintenance
- Each project competes for DCEG resources

Operations

Research & Development



DESL DNA Extraction and Handling

7 staff

Project Management

3 staff

Administrative Support

3 staff

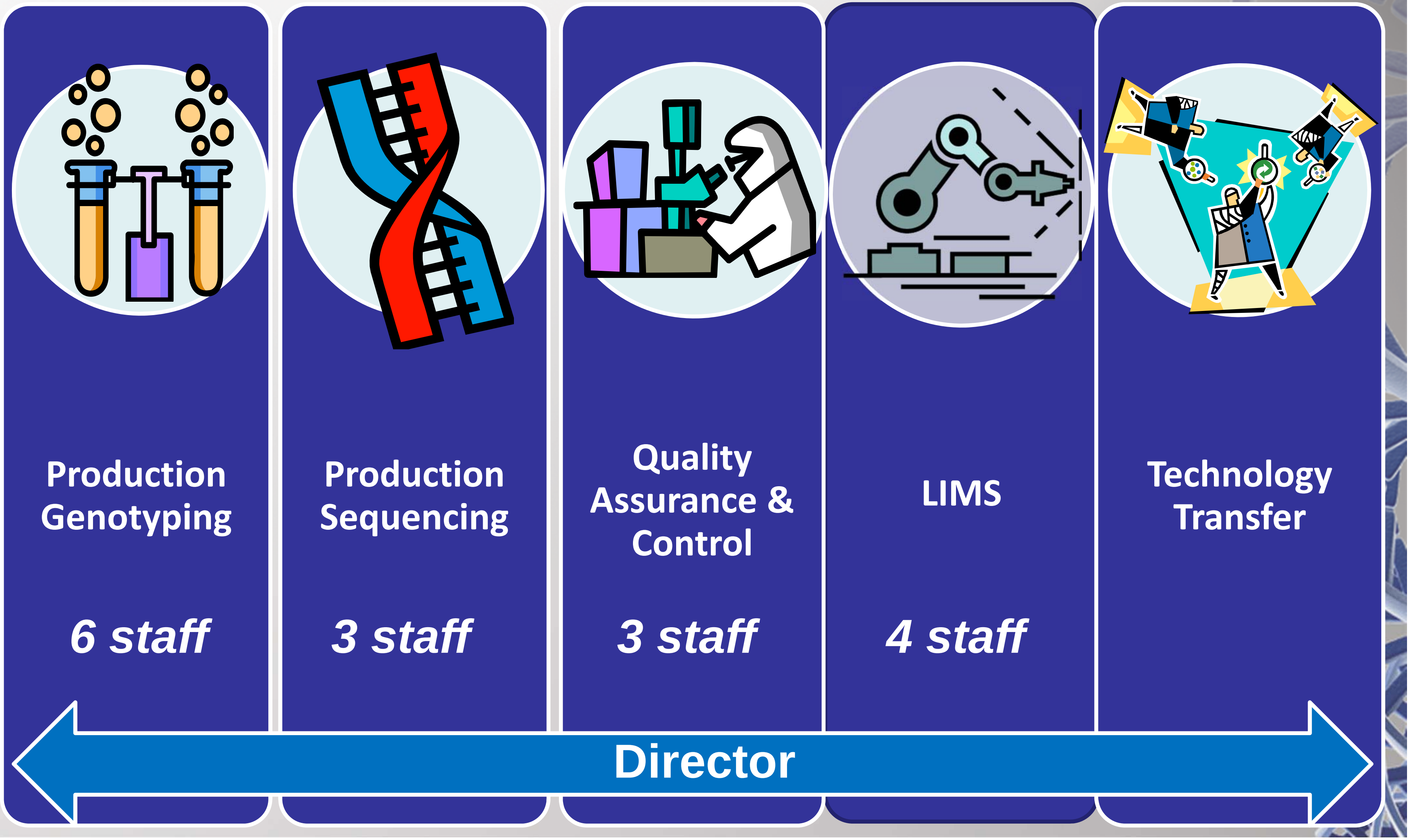
Director

Research and Development

6 staff

Director

Production Laboratory



Production Genotyping	Production Sequencing	Quality Assurance & Control	LIMS	Technology Transfer
6 staff	*3 staff*	*3 staff*	*4 staff*	

Director

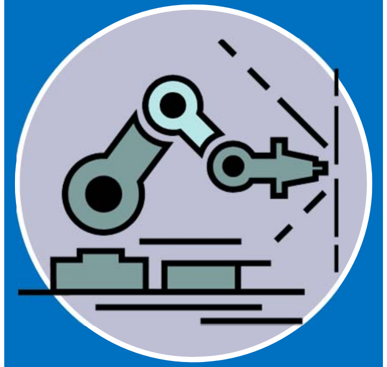
Critical CGF Laboratory Team

Quality Assurance & Control

3 staff

- Review technology performance metrics
- Generate and update:
  - SOPs
  - Staff training
- Equipment maintenance
- Follow-up on laboratory problems
- Cost savings measures

CGF Bioinformatics & Scientific Operations



LIMS, Database & Web

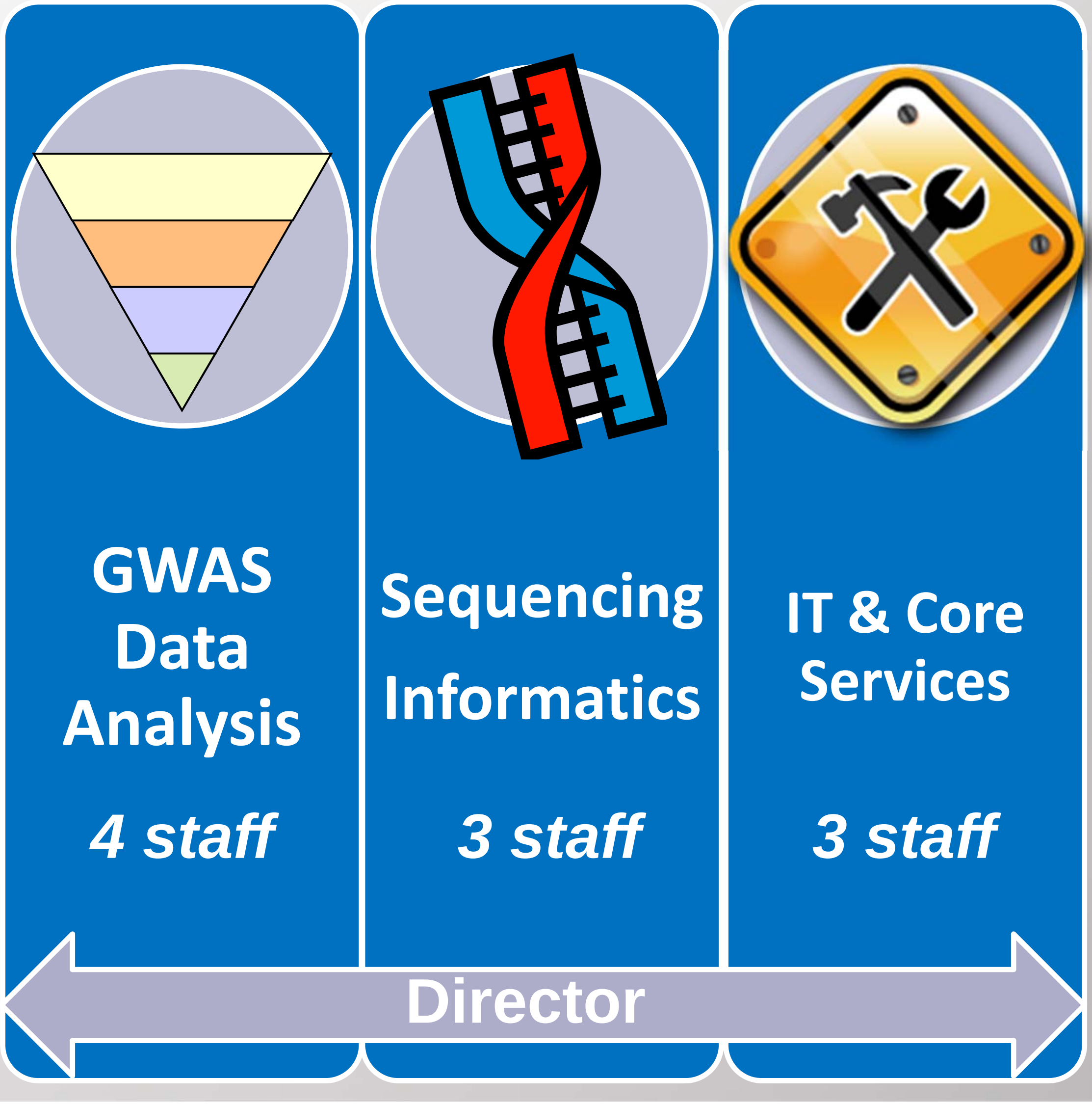
4 staff

- Maintains Commercial LIMS
  - LabVantage 2004
- Customize content for CGF workflow
- Oversees archiving of data
  - Virtual lab note books only
- Oversee security/permissions

- Maintains websites
  - Public CGF
    - http://cgf.nci.nih.gov/
  - VariantGPS (replaces SNP500)
    - http://variantgps.nci.nih.gov

Bioinformatics & Analysis Version 3.0

Science and informatics at warp speed



GWAS Data Analysis

4 staff

Sequencing Informatics

3 staff

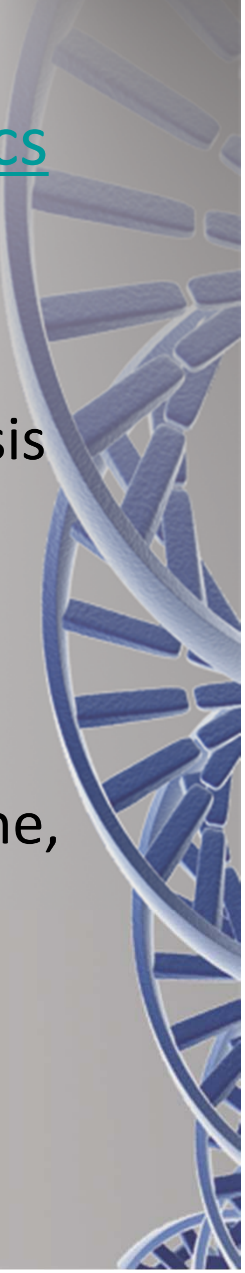
IT & Core Services

3 staff

Director

Open Source Tools

- GLU software: http://code.google.com/p/glu-genetics
- Genotype data
  - SNP array data management
  - Quality control, population structure, & association analysis
- Next-generation sequencing (NGS)
  - Infrastructure to produce and manage alignments
  - Parse and manipulate variants
    - Conversions to/from VCF, GFF, PLINK, BEAGLE, Germline, GLU
    - Annotation of known/novel, function, frequency
  - Efficient *in silico* exome/regional pull-down
  - Visualization tools: Coverage, ploidy, CNV, SV, allelic ratio



Onsite CGF IT Infrastructure

IT & Core Services

3 staff

- High-performance computing clusters

  Over 640 CPU cores, >2 TB RAM

  Supporting CGF

  + DCEG (LTG, BB, REB, GEB)

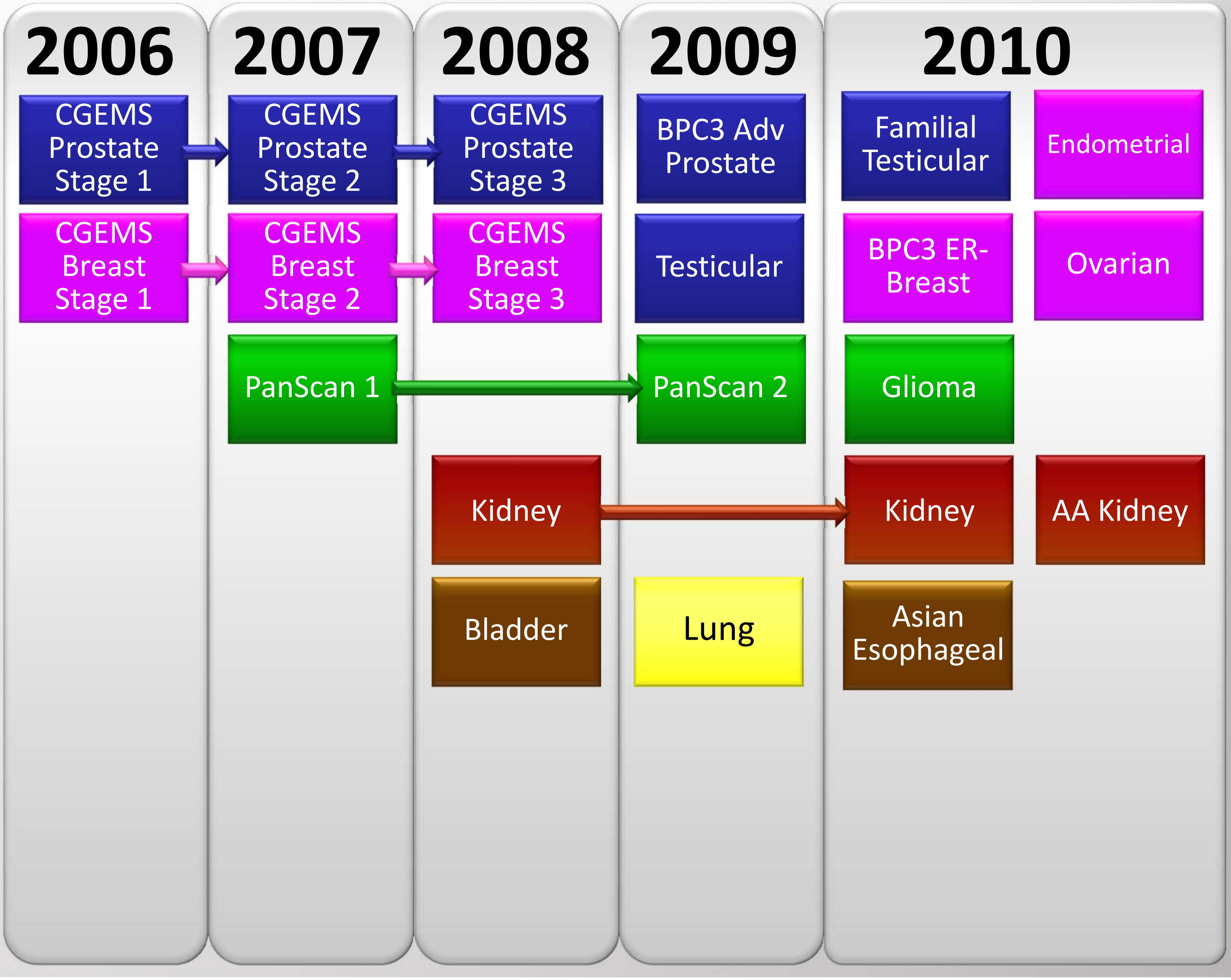
  + CCR/SAIC-F Sequencing Facility
- Laboratory instrument support
  - Integrated high performance computing
- Large-scale data storage subsystems
  - Over 300 TB tier 1 storage
- Local and wide-area networking
- Battery and generator backup of computing and HVAC
- Systems administration and security
  - Interface with CBIIT and CIT

CGF Data Output since 2002

Analyzed & Delivered Data

SNP/CNV Genotypes:	76 x 10^{12}
Regional Sequences:	100 Gbps
High-coverage exomes:	231, 2 Tbps aligned sequence, 200x avg coverage for llumina HiSeq + Nimblegen 10-12x for Roche/454
Whole-genomes:	78, 15 Tbps aligned sequence, 60x avg coverage, Complete Genomics

GWAS Timeline



2006

CGEMS Prostate Stage 1

CGEMS Breast Stage 1

2007

CGEMS Prostate Stage 2

CGEMS Breast Stage 2

PanScan 1

2008

CGEMS Prostate Stage 3

CGEMS Breast Stage 3

Kidney

Bladder

2009

BPC3 Adv Prostate

Testicular

PanScan 2

Lung

2010

Familial Testicular

Endometrial

BPC3 ER- Breast

Ovarian

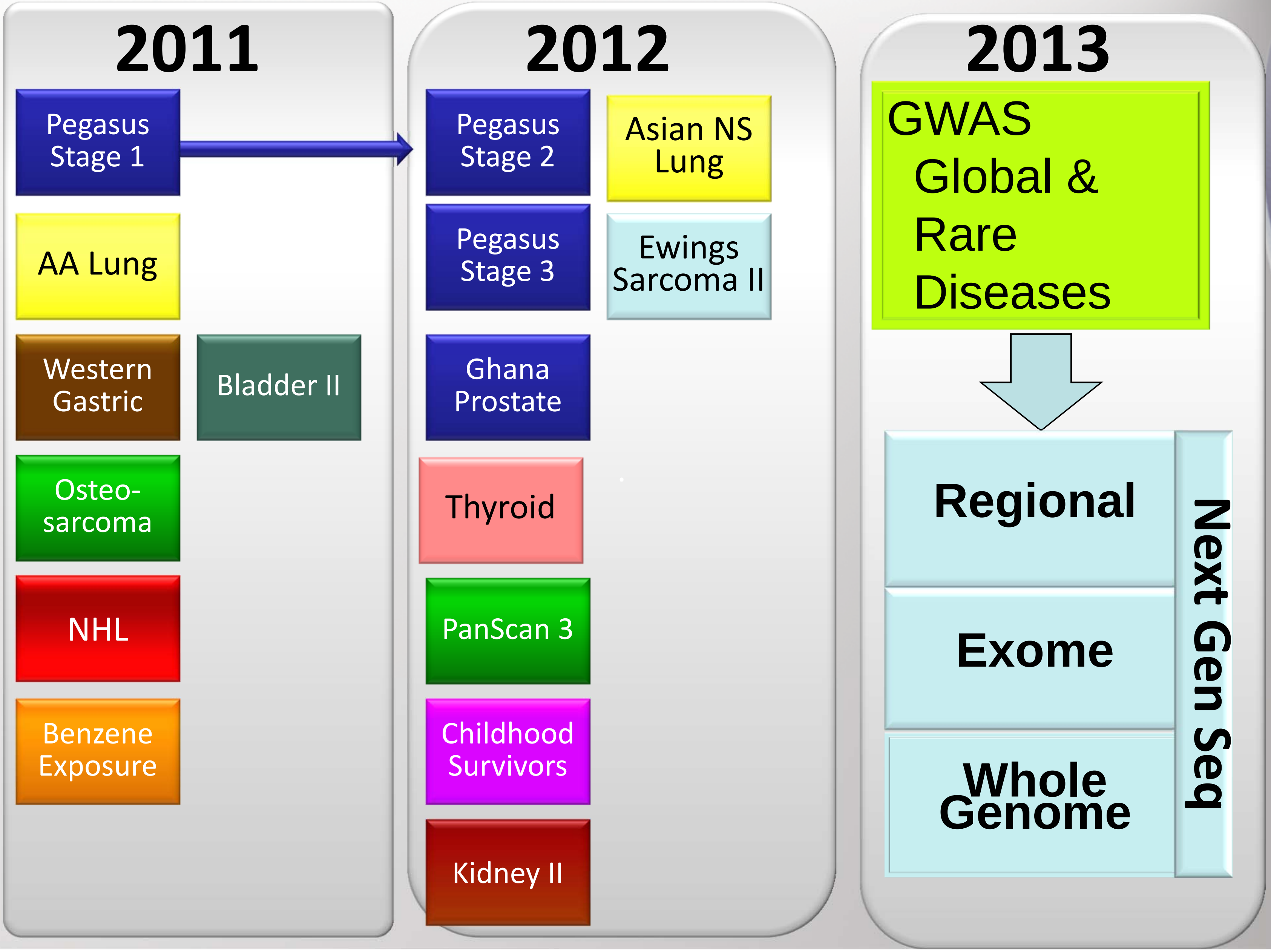
Glioma

Kidney

AA Kidney

Asian Esophageal

GWAS->Sequencing Timeline



2011

- Pegasus Stage 1
- AA Lung
- Western Gastric
- Bladder II
- Osteo-sarcoma
- NHL
- Benzene Exposure

2012

- Pegasus Stage 2
- Asian NS Lung
- Pegasus Stage 3
- Ewings Sarcoma II
- Ghana Prostate
- Thyroid
- PanScan 3
- Childhood Survivors
- Kidney II

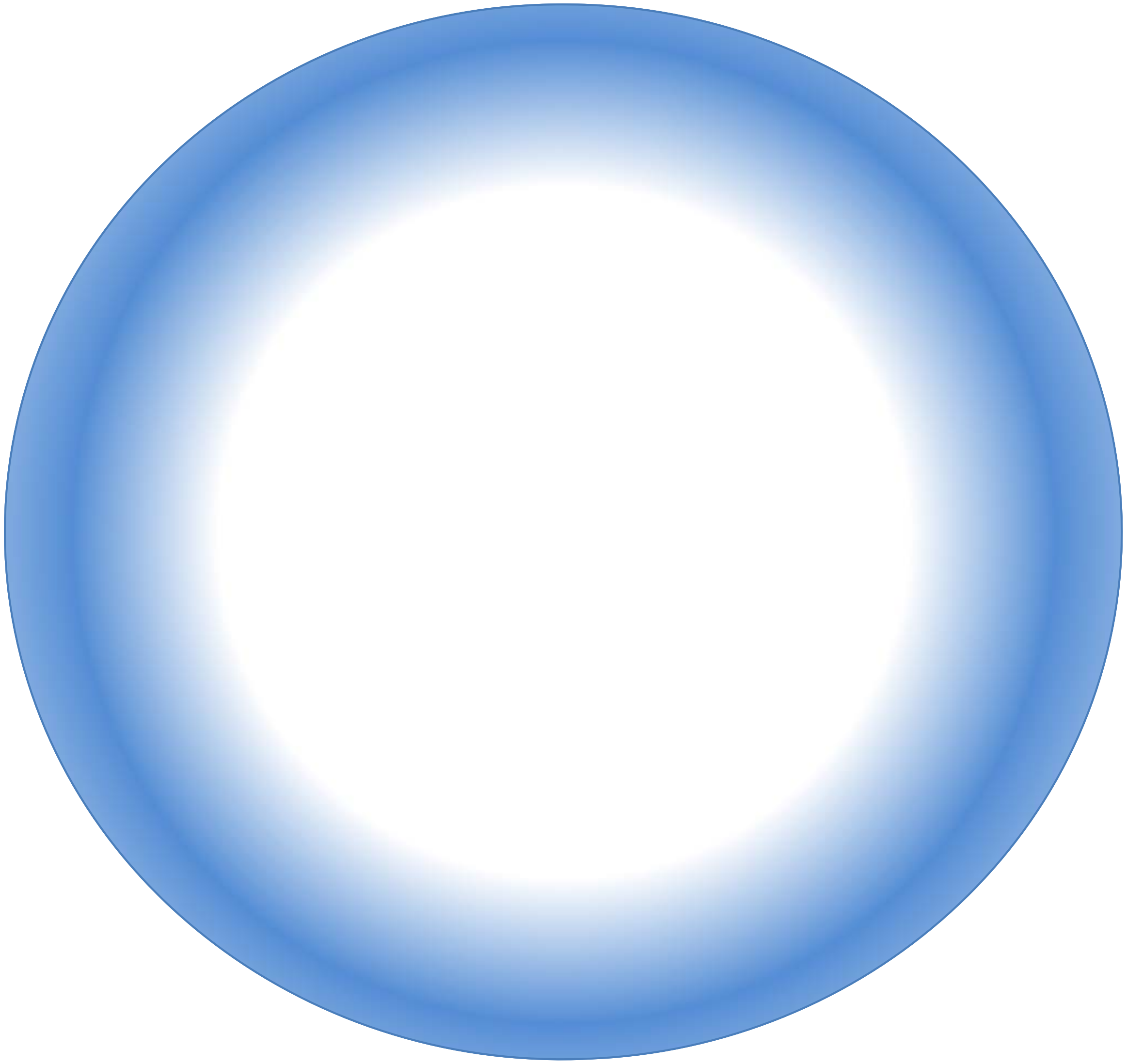
2013

GWAS
Global &
Rare
Diseases

Next Gen Seq

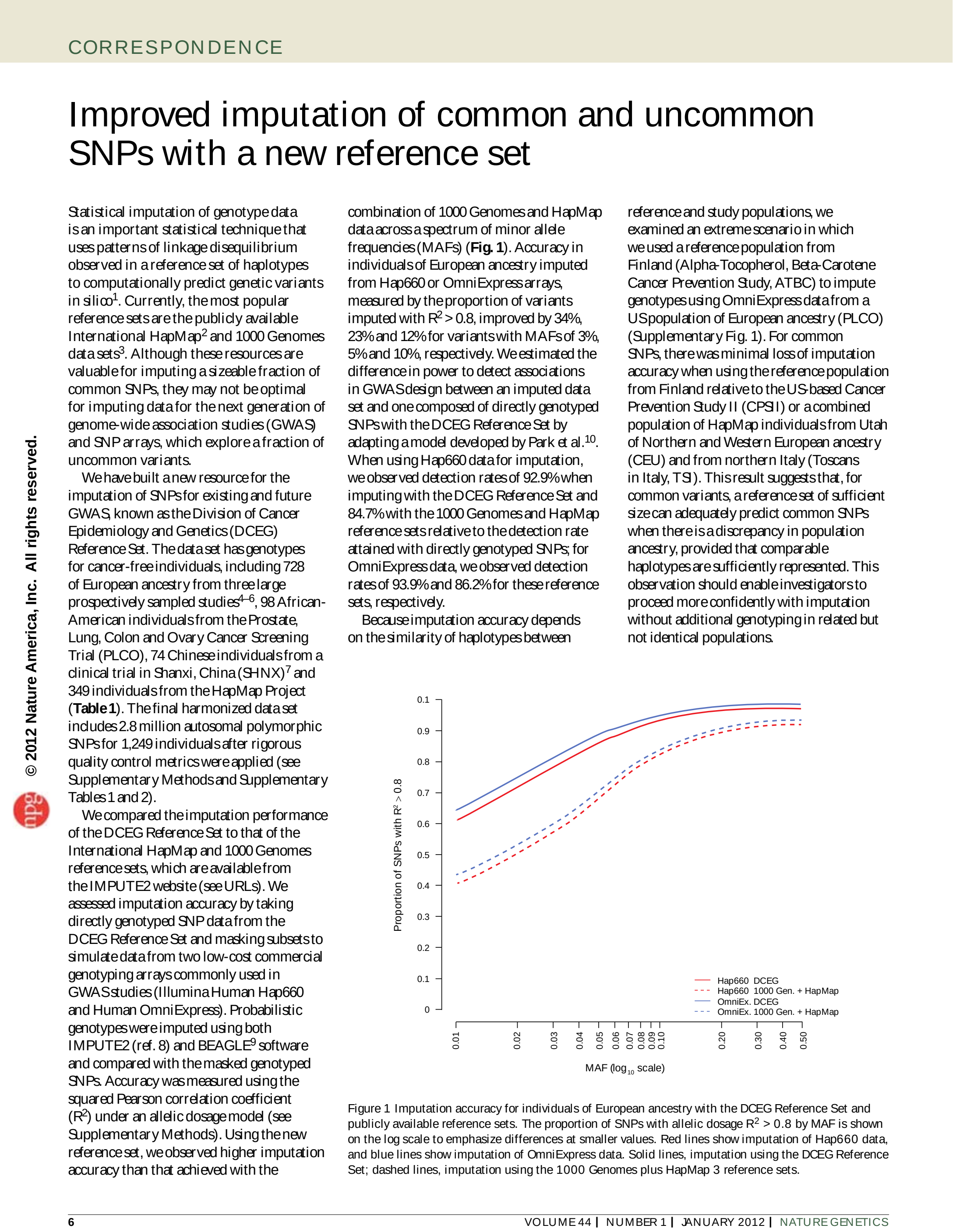
- Regional
- Exome
- Whole Genome

DCEG Total GWAS Set (TGS)



Resource based on DCEG 'TGS'

Zhaoming Wang, Kevin B Jacobs
Meredith Yeager, Amy Hutchinson
Joshua Sampson, Nilanjan Chatterjee,
Demetrius Albanes, Sonja I Berndt
Charles C Chung, W Ryan Diver
Susan M Gapstur, Lauren R Teras
Christopher A Haiman, Brian E Henderson,
Daniel Stram, Xiang Deng, Ann W Hsing,
Jarmo Virtamo, Michael A Eberle,
Jennifer L Stone, Mark P Purdue,
Phil Taylor, Margaret Tucker,
Stephen J Chanock



Improved imputation of common and uncommon SNPs with a new reference set

Statistical imputation of genotype data is an important statistical technique that uses patterns of linkage disequilibrium observed in a reference set of haplotypes to computationally predict genetic variants in silico[1]. Currently, the most popular reference sets are the publicly available International HapMap[2] and 1000 Genomes data sets[3]. Although these resources are valuable for imputing a sizeable fraction of common SNPs, they may not be optimal for imputing data for the next generation of genome-wide association studies (GWAS) and SNP arrays, which explore a fraction of uncommon variants.

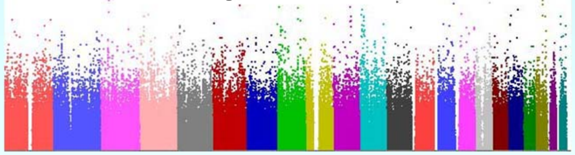
We have built a new resource for the imputation of SNPs for existing and future GWAS, known as the Division of Cancer Epidemiology and Genetics (DCEG) Reference Set. The data set has genotypes for cancer-free individuals, including 728 of European ancestry from three large prospectively sampled studies[4–6], 98 African-American individuals from the Prostate, Lung, Colon and Ovary Cancer Screening Trial (PLCO), 74 Chinese individuals from a clinical trial in Shanxi, China (SHNX)[7] and 349 individuals from the HapMap Project (**Table 1**). The final harmonized data set includes 2.8 million autosomal polymorphic SNPs for 1,249 individuals after rigorous quality control metrics were applied (see Supplementary Methods and Supplementary Tables 1 and 2).

We compared the imputation performance of the DCEG Reference Set to that of the International HapMap and 1000 Genomes reference sets, which are available from the IMPUTE2 website (see URLs). We assessed imputation accuracy by taking directly genotyped SNP data from the DCEG Reference Set and masking subsets to simulate data from two low-cost commercial genotyping arrays commonly used in GWAS studies (Illumina Human Hap660 and Human OmniExpress). Probabilistic genotypes were imputed using both IMPUTE2 (ref. 8) and BEAGLE[9] software and compared with the masked genotyped SNPs. Accuracy was measured using the squared Pearson correlation coefficient (R^2) under an allelic dosage model (see Supplementary Methods). Using the new reference set, we observed higher imputation accuracy than that achieved with the combination of 1000 Genomes and HapMap data across a spectrum of minor allele frequencies (MAFs) (**Fig. 1**). Accuracy in individuals of European ancestry imputed from Hap660 or OmniExpress arrays, measured by the proportion of variants imputed with $R^2 > 0.8$, improved by 34%, 23% and 12% for variants with MAFs of 3%, 5% and 10%, respectively. We estimated the difference in power to detect associations in GWAS design between an imputed data set and one composed of directly genotyped SNPs with the DCEG Reference Set by adapting a model developed by Park et al.[10]. When using Hap660 data for imputation, we observed detection rates of 92.9% when imputing with the DCEG Reference Set and 84.7% with the 1000 Genomes and HapMap reference sets relative to the detection rate attained with directly genotyped SNPs; for OmniExpress data, we observed detection rates of 93.9% and 86.2% for these reference sets, respectively.

Because imputation accuracy depends on the similarity of haplotypes between reference and study populations, we examined an extreme scenario in which we used a reference population from Finland (Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study, ATBC) to impute genotypes using OmniExpress data from a US population of European ancestry (PLCO) (Supplementary Fig. 1). For common SNPs, there was minimal loss of imputation accuracy when using the reference population from Finland relative to the US-based Cancer Prevention Study II (CPSII) or a combined population of HapMap individuals from Utah of Northern and Western European ancestry (CEU) and from northern Italy (Toscans in Italy, TSI). This result suggests that, for common variants, a reference set of sufficient size can adequately predict common SNPs when there is a discrepancy in population ancestry, provided that comparable haplotypes are sufficiently represented. This observation should enable investigators to proceed more confidently with imputation without additional genotyping in related but not identical populations.

Figure 1 Imputation accuracy for individuals of European ancestry with the DCEG Reference Set and publicly available reference sets. The proportion of SNPs with allelic dosage $R^2 > 0.8$ by MAF is shown on the log scale to emphasize differences at smaller values. Red lines show imputation of Hap660 data, and blue lines show imputation of OmniExpress data. Solid lines, imputation using the DCEG Reference Set; dashed lines, imputation using the 1000 Genomes plus HapMap 3 reference sets.

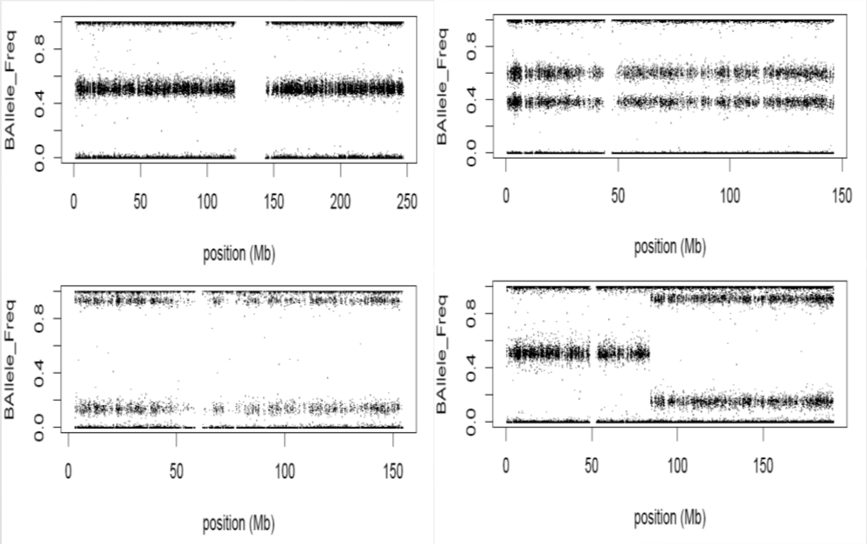
Unanticipated Directions

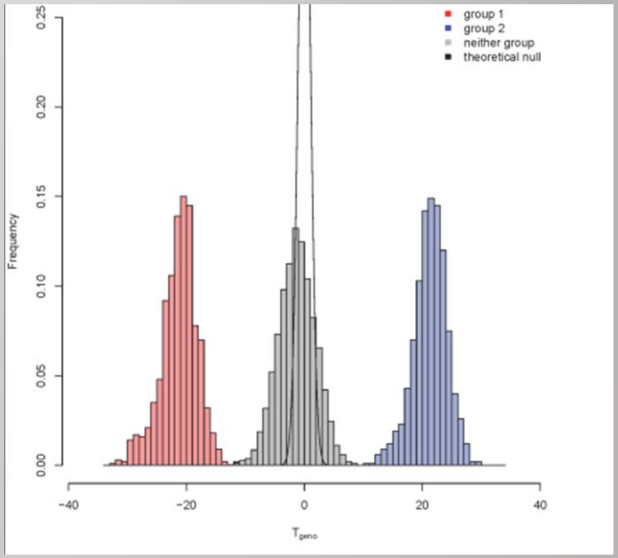


Genome-wide association studies

Large chromosomal abnormalities, structural variation, aneuploidy in Germ-line DNA

Privacy & Confidentiality GWAS membership

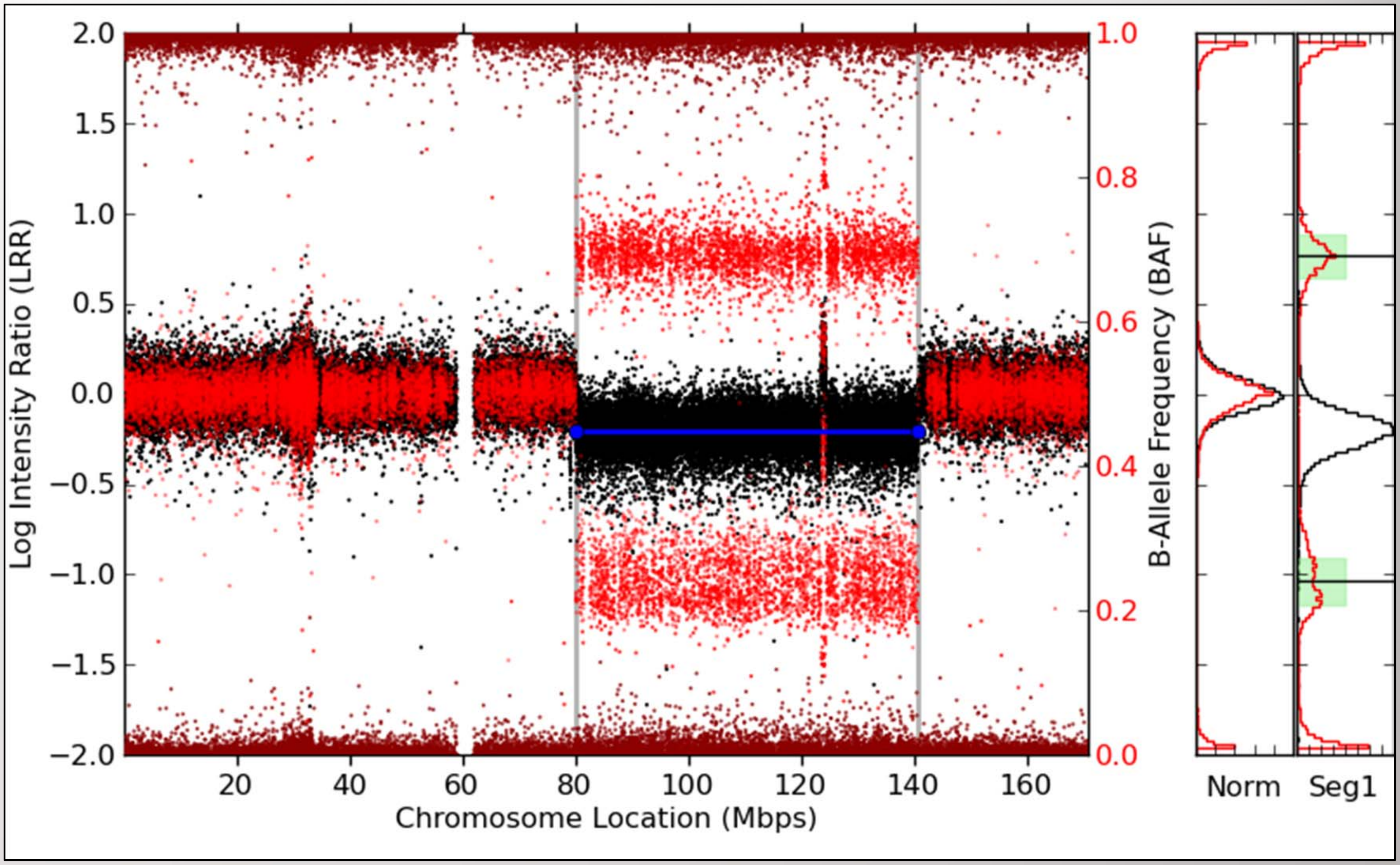




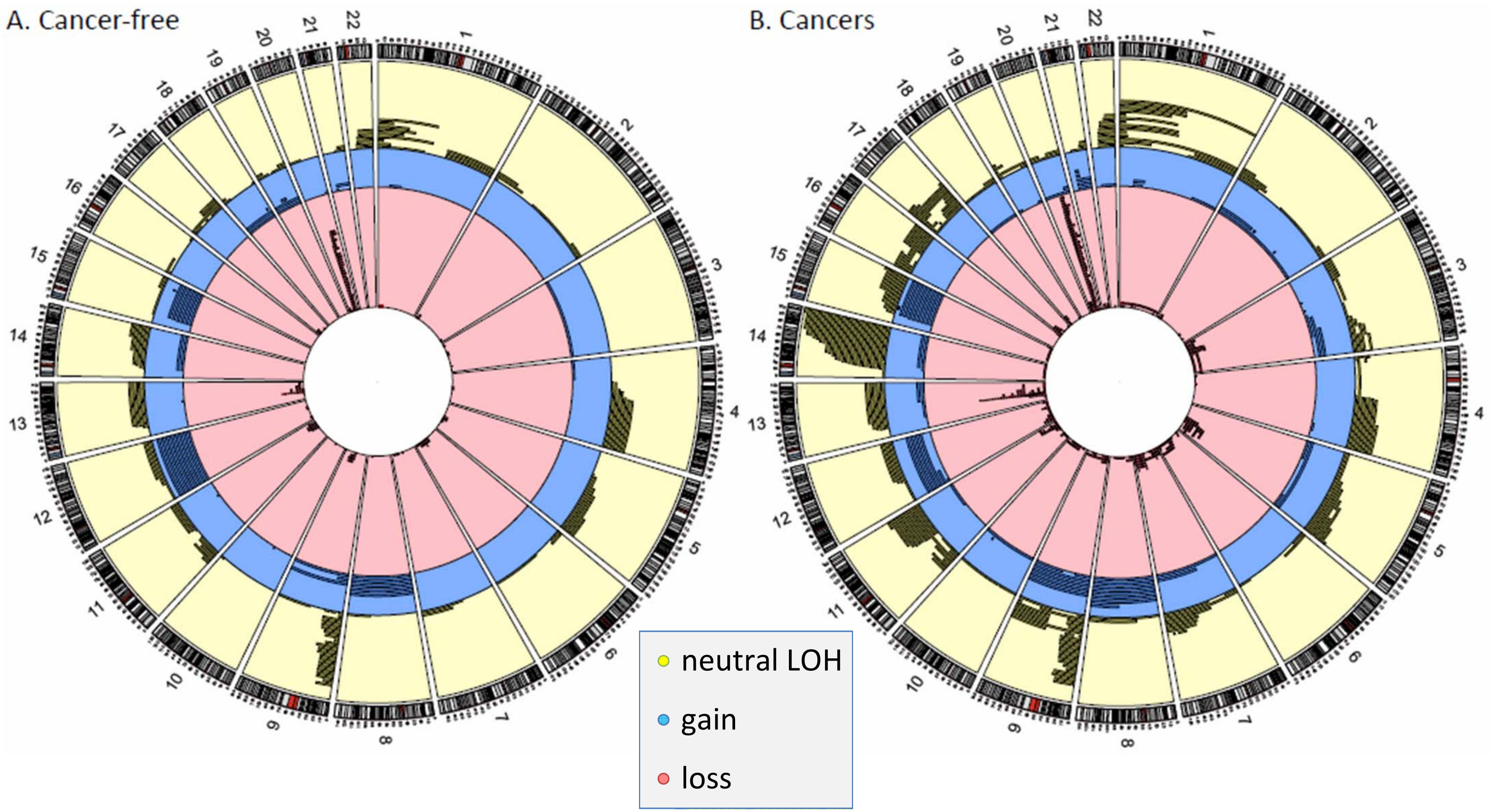
Rodriguez-Santiago AJHG 2010

Jacobs Nature Genetics 2009

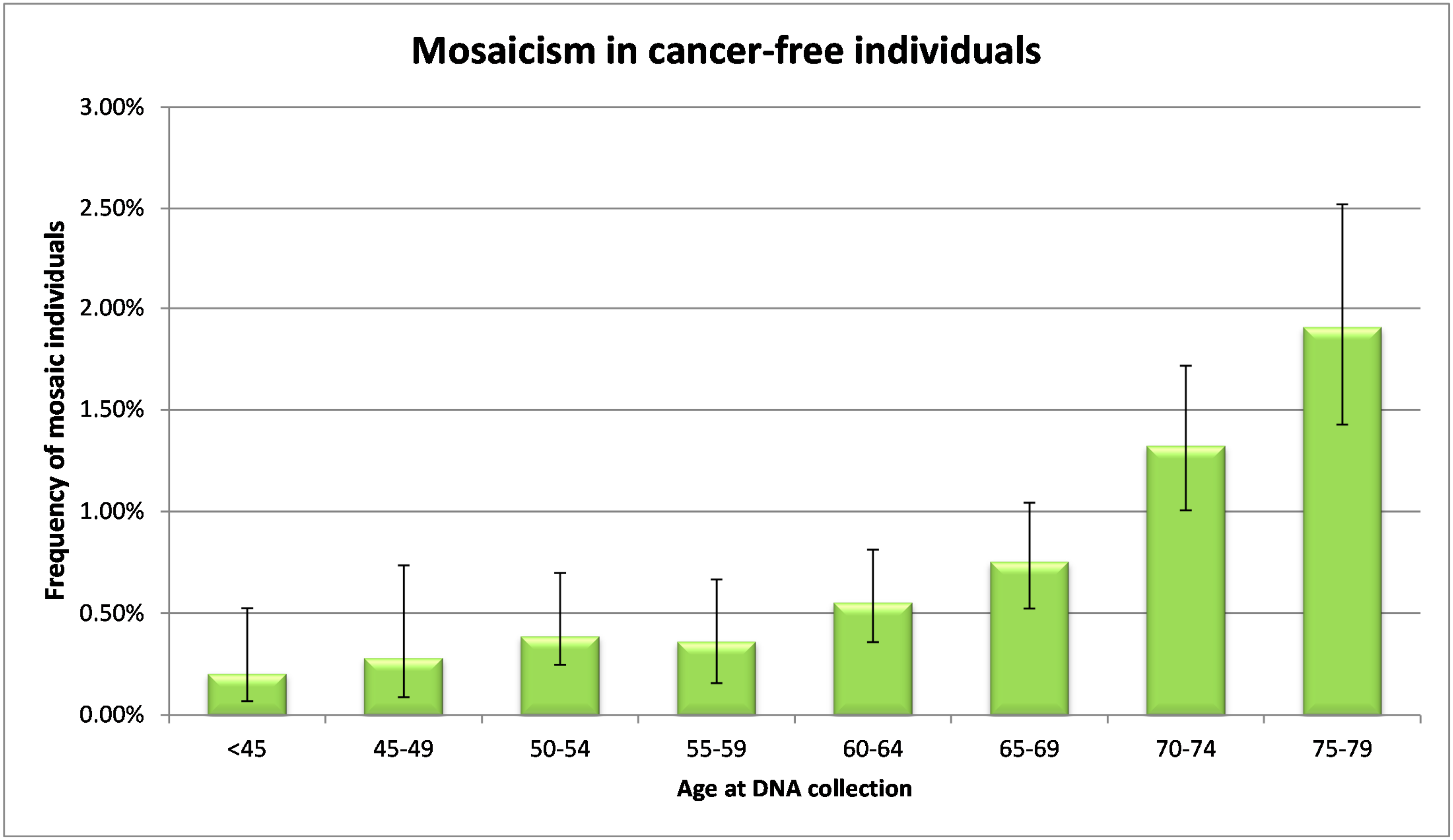
Mosaic Deletion



Circos Plot of large mosaic events (> 2 Mb) in 57,583 individuals



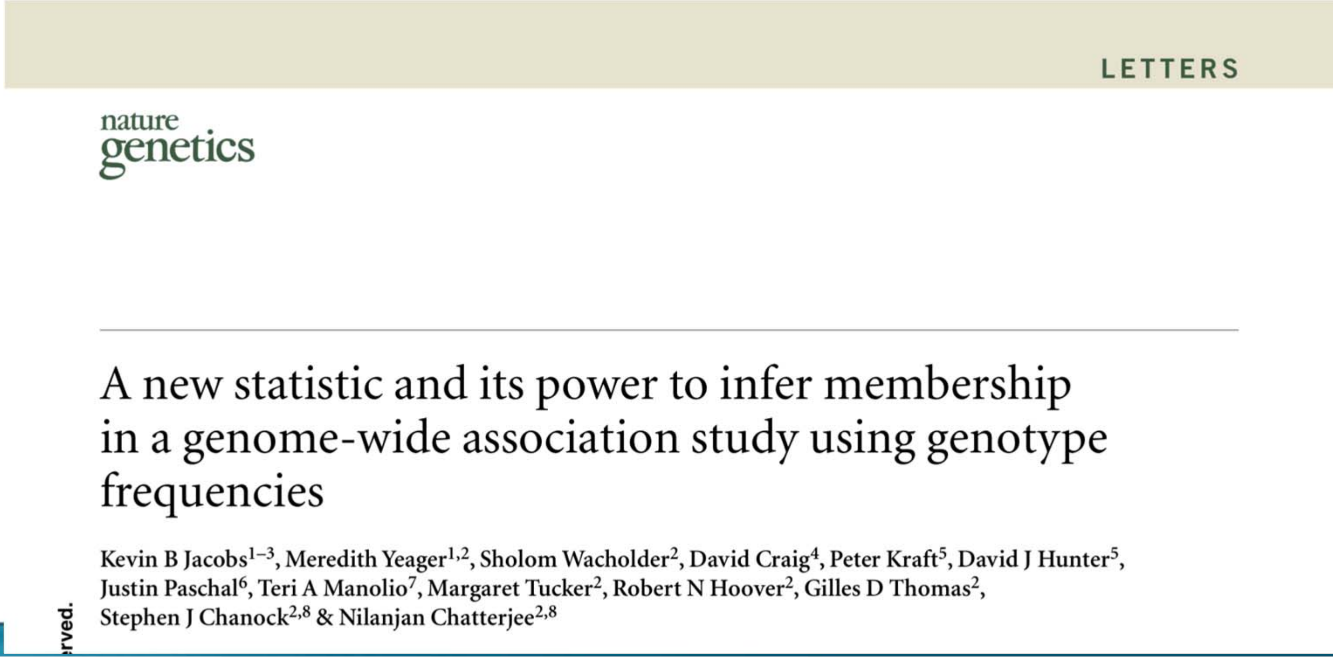
Age at DNA Collection is the Strongest Predictor of Genetic Mosaicism



CGF & Data Sharing

CGEMS
Cancer Genetic Markers of Susceptibility

- Posted first public GWAS datasets for breast & prostate cancer
  - Aggregate data removed in 2008 in response to NIH policy change
- Led development of standards for GWAS posting with dbGaP
- Contributed all DCEG GWAS datasets to dbGaP
- CGF was instrumental in addressing privacy issues with GWAS and other high-dimensional aggregate genomics data



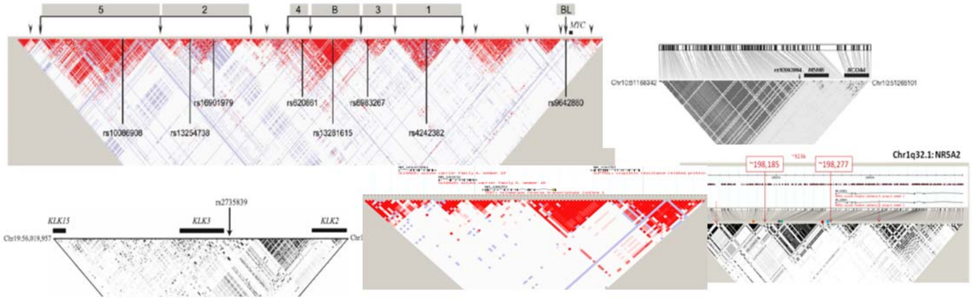
LETTERS

nature genetics

A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies

Kevin B Jacobs[1–3], Meredith Yeager[1,2], Sholom Wacholder[2], David Craig[4], Peter Kraft[5], David J Hunter[5], Justin Paschal[6], Teri A Manolio[7], Margaret Tucker[2], Robert N Hoover[2], Gilles D Thomas[2], Stephen J Chanock[2,8] & Nilanjan Chatterjee[2,8]

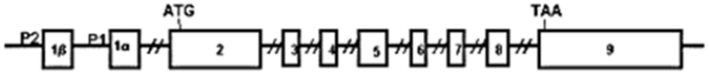
rved.

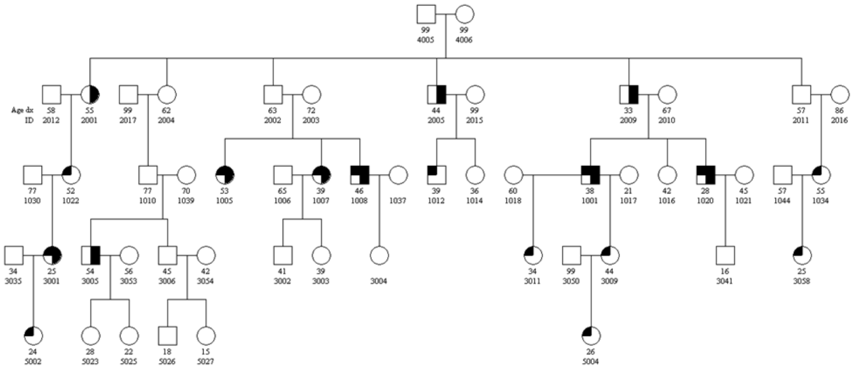


Regional GWAS and linkage follow-up

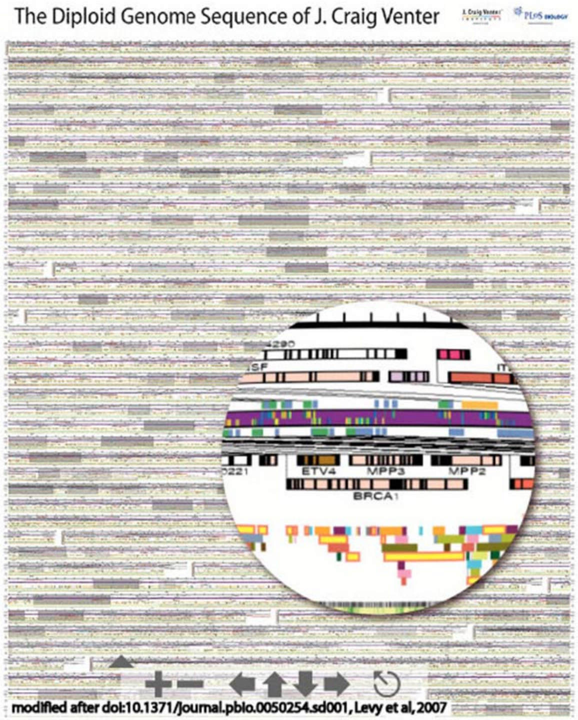
Sequencing

Candidate gene/exon





Whole exome



modified after doi:10.1371/journal.pbio.0050254.sd001, Levy et al, 2007

Whole genome

NGS Capabilities

Roche 454 GS FLX (2)

- Installed 2008
- Chosen for:
  - Read length
  - Multiplexing capability
- Current Output:
  - Multiplexing up to 264 samples
  - Average of 350-400bp/read

Life Technology/Ion Torrent PGM

- First Installed Jan 2011
- 6 machines as of Jan 2012
- Chosen for:
  - Cost
  - Reliability
  - Flexibility

~~Illumina HiScan SQ~~

- ~~Installed August 2010~~

Illumina HiSeq 2000

- Installed April 2011
- Chosen for:
  - Throughput sufficient for exome/whole genome sequencing
- Current Output:
  - 300 Gbps/week (16 exomes)
  - 76-100 bp PE reads
- Expanded sequencing applications
  - CHiPseq
  - RNAseq

Bumps along the way….

2007: Movement into ATP-SAIC

- Expectation of better alignment with program resources

2009: Movement out

- ATP Leadership sought to interrupt close collaboration and direct towards other business opportunities
- Placed under SAIC Research Administration OD

Recent Bump

- Sample handling bottleneck

  CGF processes used for setting up DNA Extraction & Sample Handling Lab (DESL) in 2006

  Increased demands stressed DESL

  Stand alone service lab was realigned with CGF in 2011 due to

  - Quality Control Issues
  - Production Delays

Current Focus of Activities

Role of GWAS for:

1. Less common diseases w/ limited biospecimens
2. Complete our understanding of the contribution of common variant to cancer risk
   - Overall and population specific
3. Denser arrays for less common variation

Family & Special Population Analysis

- Exome & whole-genome sequencing
- Follow-up in families and unrelated subjects

Challenges Ahead

- Transition from GWAS to sequencing for investigation of germ-line susceptibility
- Further integration of environmental exposures
- Optimal storage, processing, and mining of whole-genome sequence data

Critical Mass

Analytical and Bioinformatic Expertise

- Close collaboration from inception to publication
  - Studies
  - Methodology
- Software development & dissemination
- Systematic data sharing
- Integrative analysis across studies & data types

Success of DCEG Core Genotyping Facility

- DCEG's decades of investment in epidemiology & genetics
- Close collaborations between DCEG & FFRDC (CGF) epidemiologists, biostatisticians, geneticists, bioinformaticians and laboratory experts

➢ Dedicated facility framework