DCEG Core Genotyping Facility

Stephen Chanock, M.D.

Chief, Laboratory of Translational Genomics
Director, Core Genotyping Facility

January 25, 2012
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


- Candidate SNP
- Functional Data
- Candidate Genes
- Biological Plausibility
- Genetic Markers
- Candidate Pathway
- Biological Plausibility
- Genome Wide Association Studies
- Exome Sequencing
- Regional Sequencing
- GWAS & Linkage
- Whole Genome Sequencing
- Whole Genome Sequencing
- Population-based
CGF Facilities Footprint

Advanced Technology Center: Gaithersburg

- **Web-lab space:** 4,618 ft²
  - Optimized for genomics workflows
  - Lab staff desks
  - Sample-handling, PCR, and post-PCR areas
  - Dry-lab offices: 1,615 ft²
    - 11 offices/spaces for 24 staff

- **On-site data center:** 105 ft²
- **Additional cryogenic storage:** 145 ft²
The Core Genotyping Facility
Dedicated DCEG Facility

What’s in a name? Core Plus Plus

Core Services
- Genotyping
- Sequencing
- Computing Support
- Data Analysis

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

Core Services + Innovation
- Biotechnology
- Genomics
- Computational Methods
- Statistical Methods

500 Publications
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)

<table>
<thead>
<tr>
<th>Mission</th>
<th>Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critique of Science</td>
<td>Chair:</td>
</tr>
<tr>
<td>Statistical Review</td>
<td>P Tucker, Director, HGP</td>
</tr>
<tr>
<td>Approval letter required to proceed to CGF queue</td>
<td>PIs from each Branch rotate every 2 years</td>
</tr>
<tr>
<td>Minutes</td>
<td>S Chanock</td>
</tr>
<tr>
<td>Chair can approve small projects &amp; revisions</td>
<td>K Pitt</td>
</tr>
</tbody>
</table>
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

- DESL DNA Extraction and Handling: 7 staff
- Project Management: 3 staff
- Administrative Support: 3 staff

Director

Research & Development

- Research and Development: 6 staff

Director
Production Laboratory

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

Director
Critical CGF Laboratory Team

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

Quality Assurance & Control
3 staff
CGF Bioinformatics & Scientific Operations

- 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)
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](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

- 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

IT & Core Services
3 staff
CGF Data Output since 2002

Analyzed & Delivered Data

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Quantity/Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP/CNV Genotypes:</td>
<td>76 x 10^{12}</td>
</tr>
<tr>
<td>Regional Sequences:</td>
<td>100 Gbps</td>
</tr>
<tr>
<td>High-coverage exomes:</td>
<td>231, 2 Tbps aligned sequence, 200x avg coverage for Illumina HiSeq + Nimblegen</td>
</tr>
<tr>
<td></td>
<td>10-12x for Roche/454</td>
</tr>
<tr>
<td>Whole-genomes:</td>
<td>78, 15 Tbps aligned sequence, 60x avg coverage, Complete Genomics</td>
</tr>
</tbody>
</table>
GWAS->Sequencing Timeline

2011
- Pegasus Stage 1
- AA Lung
- Western Gastric
- Osteosarcoma
- NHL
- Benzene Exposure

2012
- Pegasus Stage 2
- Pegasus Stage 3
- Ghana Prostate
- PanScan 3
- Childhood Survivors
- Kidney II
- Asian NS Lung
- Ewings Sarcoma II

2013
- GWAS Global & Rare Diseases
  - Regional
  - Exome
  - Whole Genome
  - Next Gen Seq
DCEG Total GWAS Set (TGS)
Improved imputation of common and uncommon SNPs with a new reference set

Resource based on DCEG ‘TGS’

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. Currently, the most popular reference sets are the publicly available International HapMap² and 1000 Genomes data sets. Although these resources are valuable for imputing a sizable 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 dataset has genotypes for cancer-free individuals, including 728 of European ancestry from three large prospectively sampled studies, 98 African-American individuals from the Proteomic Lung, Colon and Ovary Cancer Screening Trial (PLCO), 74 Chinese individuals from a clinical trial in Shandong, China (SHINX)² and 349 individuals from the HapMap Project (Table 1). The final harmonized dataset 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 HumanHap660 and HumanOmniExpress). Probabilistic genotypes were imputed using both IMPUTE2 (ref. 8) and BEAGLE非常适合 software and compared with the masked genotyped SNPs. Accuracy was measured using the squared Pearson correlation coefficient (R²) under an additive 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 imputing with R² > 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 HapMap 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 suggested 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.
Large chromosomal abnormalities, structural variation, aneuploidy in Germ-line DNA

Privacy & Confidentiality GWAS membership

Unanticipated Directions

Genome-wide association studies

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

Mosaicism in cancer-free individuals

<table>
<thead>
<tr>
<th>Age at DNA collection</th>
<th>Frequency of mosaic individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;45</td>
<td>0.00%</td>
</tr>
<tr>
<td>45-49</td>
<td>0.00%</td>
</tr>
<tr>
<td>50-54</td>
<td>0.00%</td>
</tr>
<tr>
<td>55-59</td>
<td>0.00%</td>
</tr>
<tr>
<td>60-64</td>
<td>0.00%</td>
</tr>
<tr>
<td>65-69</td>
<td>0.00%</td>
</tr>
<tr>
<td>70-74</td>
<td>2.50%</td>
</tr>
<tr>
<td>75-79</td>
<td>3.00%</td>
</tr>
</tbody>
</table>
CGF & Data Sharing

- 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

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

Kevin B Jacobs¹,², Meredith Yeager¹,², Sholom Wacholder², David Craig⁴, Peter Kraft⁵, David J Hunter⁵, Justin Paschal⁶, Teri A Manolio⁷, Margaret Tucker⁴, Robert N Hoover⁴, Gilles D Thomas⁵, Stephen J Chanock²,⁸ & Nilanjan Chatterjee¹,⁸
Sequencing

Regional GWAS and linkage follow-up

Candidate gene/exon

Sequencing

Whole genome

Whole exome
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