Genomic Data Quality and Advancing Research

Neil Risch, Ph.D.
University of California, San Francisco
Kaiser Permanente Northern California

Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research – a Workshop

Institute of Medicine Roundtable on Translating Genomic-Based Research for Health

December 8, 2014

KAISER PERMANENTE®

UCSF

Sources and type of genetic/genomic data at UCSF/Kaiser

- Metabolic (inborn error) disorder results (Mass Spec)
- Chromosome studies (cytogenetics, fish)
- Array CGH for smaller chromosome anomalies
- Mendelian disorder testing – DNA based (send out)
- Tumor sequencing

- Currently test results (especially externally derived) are in pdf form linked to the EMR. This does not make research simple.

- . Quality of data depends on two things:

  a. Reliability (reproducibility)

  b. Validity (Predictive value for clinical outcomes)

Reliability for most tests is generally high; validity is complex, as there is generally a high rate of results with uncertain implications (e.g. variants of unknown significance)

Reliability for NGS DNA sequencing is not yet that good, especially as compared to Sanger Sequencing

Estimating genotype error rates from high-coverage next-generation sequence data

Jeffrey D. Wall,[1,2] Ling Fung Tang,[3] Brandon Zerbe,[2] Mark N. Kvale,[2] Pui-Yan Kwok,[2,3] Catherine Schaefer,[4] and Neil Risch[1,2,4]

Genome Res. published online October 10, 2014

Exome and whole-genome sequencing studies are becoming increasingly common, but little is known about the accuracy of the genotype calls made by the commonly used platforms. Here we use replicate high-coverage sequencing of blood and saliva DNA samples from four European-American individuals to estimate lower bounds on the error rates of Complete Genomics and Illumina HiSeq whole-genome and whole-exome sequencing. **Error rates for nonreference genotype calls range from 0.1% to 0.6%**, depending on the platform and the depth of coverage. Additionally, we found (1) no difference in the error profiles or rates between blood and saliva samples; (2) Complete Genomics sequences had substantially higher error rates than Illumina sequences had; (3) **error rates were higher (up to 6%) for rare or unique variants**; (4) error rates generally declined with genotype quality (GQ) score, but in a nonlinear fashion for the Illumina data, likely due to loss of specificity of GQ scores greater than 60; and (5) error rates increased with increasing depth of coverage for the Illumina data. These findings suggest that **caution should be taken in interpreting the results of next-generation sequencing-based association studies, and even more so in clinical application of this technology in the absence of validation by other more robust sequencing or genotyping methods**.

Newborn Screening in Genomic Medicine and Public Health (NHGRI/NICHD)

UCSF component: SEQUENCING OF NEWBORN BLOOD SPOT DNA TO IMPROVE AND EXPAND NEWBORN SCREENING (Bob Nussbaum PI)

- Using whole exome sequencing as an adjunct to traditional Mass Spec (which has very high sensitivity and high specificity) to reduce false positives and identify causal mutations

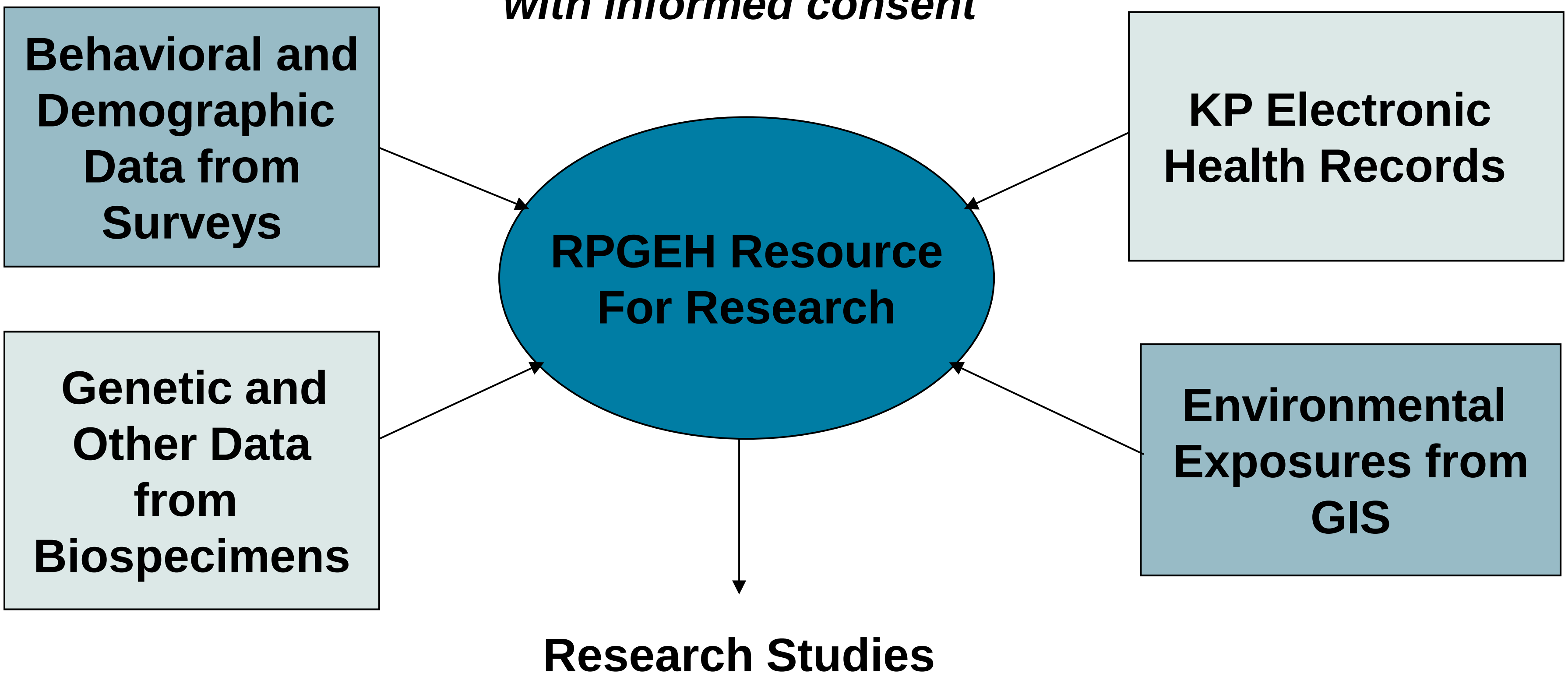
- In terms of return of results (to parents), primary application currently is pharmacogenetics related to disorder in child

- FDA has required us to use an orthogonal technology to create reliable results

Research Program on Genes, Environment, and Health (RPGEH)

A resource for studies on the genetic and environmental influences on health, disease, and aging

500K adult KP members with informed consent



Behavioral and Demographic Data from Surveys

KP Electronic Health Records

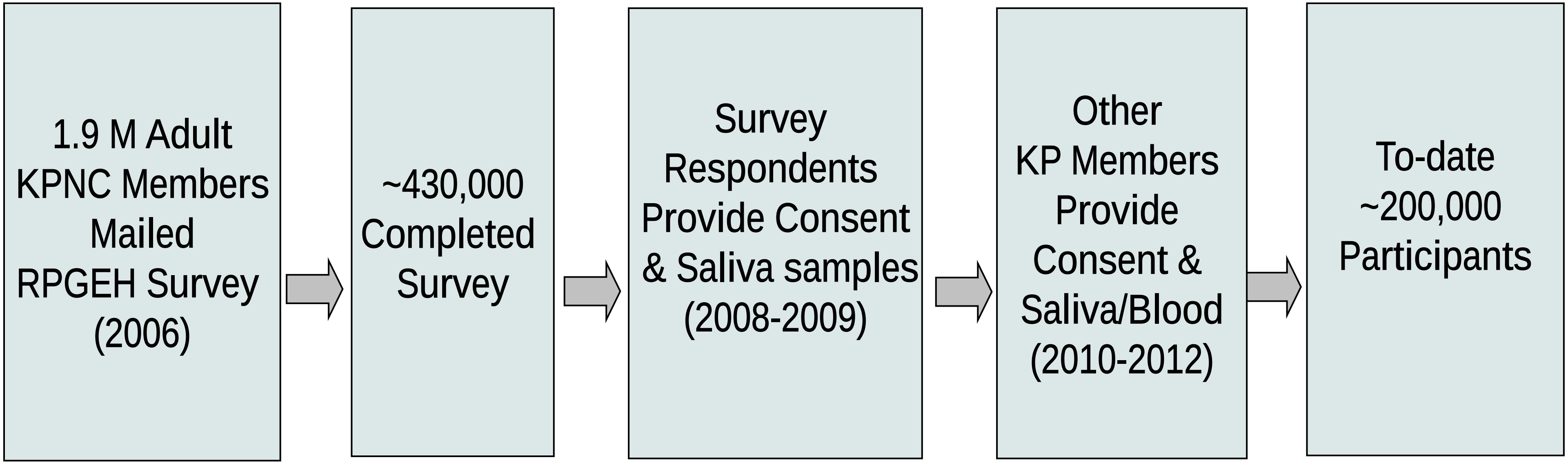
RPGEH Resource For Research

Genetic and Other Data from Biospecimens

Environmental Exposures from GIS

Research Studies

RPGEH Recruitment and Specimen Collection History



1.9 M Adult KPNC Members Mailed RPGEH Survey (2006)

~430,000 Completed Survey

Survey Respondents Provide Consent & Saliva samples (2008-2009)

Other KP Members Provide Consent & Saliva/Blood (2010-2012)

To-date ~200,000 Participants

Survey questions include: demographics, health history, family history, smoking, alcohol, diet, physical activity, and reproductive history

Available in English, Spanish, and Chinese

Broad written informed consent

Genetic Epidemiology Research on Aging (GERA) GO Project

- In 2009, we received a RC2 Grand Opportunity (GO) project funded by NIH (NIA, NIMH, and Common Fund); PI's – Cathy Schaefer, Kaiser Division of Research; Neil Risch, UCSF

- GOAL: Create a resource for research into the genetic and environmental basis for common age-related diseases and their treatment, and factors influencing healthy aging and longevity

- Specific Aims:

1. Conduct genome-wide genotyping at 675,000+ markers on 100,000 participants in RPGEH
2. Assay telomere lengths for the same100,000 samples
3. Develop customized genome-wide SNP arrays, one for each of four major race-ethnicity groups, and use these arrays for genotyping
4. Merge the GWAS and telomere data with the electronic medical record, survey and environmental data in a research database
5. Deposit data in dbGaP and provide collaborative access

Characteristics of the GERA Cohort (N= 110,266)

Characteristics	Numbers	% of Cohort
Gender		
Females	63,883	57.9%
Males	46,383	42.1%
Age at Specimen Donation		
18-39	7,025	6.3%
40-59	33,344	30.2%
60-79	57,856	52.4%
80+	12,041	10.9%

Characteristics of the GERA Cohort (Continued)

Characteristics	Numbers	% of Cohort
Race – Ethnicity		
African American	3,552	3.5%
Asian	9,190	8.3%
Latino / Mixed	11,976	10.9%
Non-Hispanic White	85,548	77.6%

Length of KP Membership in GERA Cohort

Length of KPNC Membership	%
Less than 2 years	2.1
2 to 4.9 years	3.3
5 to 9.9 years	18.9
10 to 19.9 years	24.7
20 or more years	51.0

Developing Phenotypes from the KP Electronic Medical Record

- Comprehensive electronic records beginning in 1995
  - Epic-based EMR including physician notes beginning in 2006
  - Complex and multidimensional – treated vs. untreated measures; timing of comorbidities and treatment
  - Longitudinal – many measures per person
- Diagnoses / Health Conditions
  - Validated registries for a number of conditions
  - Use relatively simple ICD-9 based algorithm; validate against registries
  - Repeated observation increases reliability and validity
- Pharmacy and labs analyzed separately and linked to diagnoses and conditions
  - Highly accurate

Selected Health Conditions in the GERA Cohort

Disease Category	Condition	Numbers of Cases (2011)
CVD	Acute Coronary Syndrome Stroke Peripheral Vascular Disease	27,296 7,740 4,741
Psychiatric	Major Depression Panic Disorder	21,483 2,046
Respiratory	Asthma Chronic bronchitis	17,345 2,746
Cancer	Breast (female) Prostate Melanoma of skin Colon	4,700 4,364 1,657 1,161
Diabetes	Type 2 Diabetes	14,734

Other Types of Data - Multidimensionality

- EKGs (N≈60,000)
- MRI/CT Scans (Brain: N≈30,000)
- Mammographic Density (N≈45,000)
- Ophthalmologic Exams (Nearly all)
- Audiograms (N≈25,000)
- Lipid Panels, Fasting Glucose, CBC & serum chemistries
- Blood Pressures
- Body Mass Index

Genotyping on the Affymetrix Axiom System

- Conducted at UCSF Institute for Human Genetics under the direction of Pui-Yan Kwok, M.D., Ph.D.

- Genotyping completed in 14 months (70 billion genotypes)
- Success rate = 104,000/110,266 = 94.3%
- Axiom system robust, throughput sustainable
- Average SNP call rate: 99.7%
- Average SNP reproducibility 99.9%
- Package based genotype calling superior to individual plate based genotyping

Summary: GWAS Results

- Genome-wide association analyses of a variety of traits and diseases, ranging from blood pressure and cholesterol and QT interval to prostate cancer and diabetes extracted from the electronic health records has led to the identification of over 600 contributing genetic variants, approximately one-third of which are novel.
- The genetic data are housed in a separate data base from the EHR data, and only available for research purposes.
- However, we are also evaluating the possible return of genetic results to study participants according to the language in the consent form.

Data Access – Two Ways

- **Via Kaiser Permanente Research Program on Genes, Environment and Health**
  - Application via a Web portal
  - Review and approval by Access Review Committee
  - Assistance with preliminary data and other content for proposals
  - Resources include data, specimens, lab services, programming

- **Via NIH's dbGaP**
  - Data now available through dbGaP
  - Reconsented subjects (about 78% of entire cohort)
  - Application via the dbGaP process
  - Resources include data only

Funding

RPGEH:

- 2005-2006 Wayne and Gladys Valley Foundation
- 2005-2006 Ellison Medical Foundation
- 2009-2010 Robert Wood Johnson Foundation
- 2005-present Kaiser Permanente

GERA Cohort:

- 2009 – 2012 RC2 Grand Opportunity Award from NIA, NIMH and NIH Director’s Office

Acknowledgements

Kaiser Permanente

Catherine Schaefer
Carol Somkin
Carlos Iribarren
Stephen Van Den Eeden
Rachel Whitmer
Charles Quesenberry
Mary Henderson
Larry Walter
Sarah Rowell
David Smethurst
Judith Millar
Sunita Miles
Sheryl Connell
Reid Wearley
Christine Aquino
Maria Miranda
Elaine Chung
Ivo Violich
Sharon Matthews
Julie Harris
Andrea Altschuler
Lori Sakoda

Eric Jorgenson
Marianne Sadler
Dana Ludwig
Stan Sciortino
Ling Shen
Dilrini Ranatunga
Petra Liljestrand
Bernie McGuire
Chia Zau
Julia Kay
Marcia Ewing
Paul Young
Kathleen Sampel-Morris
Gustavo Parra
Eboni Stephens
Deborah Burman
Lynn Simonson
Inga Wagar
Marvella Villasenior
Terrance Chinn
Jun Shan
Diane Olberg

UCSF

Neil Risch
Elizabeth Blackburn
Pui Yan Kwok
Thomas Hoffmann
Stephanie Hesselson
Mark Kvale
Yambazi Banda
Kyle Lapham
Jue Lin
Eunice Wan
Philippe Jolivalt
Jasmin Eshragh
Jeanette Atilano
Yang Cao
Simon Wong
Brad Dispensa
Tanu Shenoy
Richard Lao
Simi Mathauda
Chris Wen
Cory Fergus
Shigeshi Yamomoto

Stanford

Hua Tang
Chiara Sabatti
Sophie Candille
Nicholas Johnson
Zhongyang (Thomas) Zhang

Affymetrix

Andrea Finn
Mike Shapero
Yiping Zhan
Jeremy Golub
Teresa Webster
Yontau Yu
Gangmu Mei
David Chan
Mohini Patel

KAISER PERMANENTE®

UCSF

Use of Genomic Data in Clinical Research

- To date, use of genetic data in clinical research, aside from the very limited Mendelian carrier screening and cytogenetic studies, requires participant consent. This has presented the largest limitation on the scope in terms of subject recruitment.

- At the point at which genetic data are considered standard of care and performed routinely as other clinical tests are (results from which are broadly available for research), the situation may change, although the ethical concerns may still engender discussion, for example regarding return of results and their implications and the creation of genetic data for research purposes only and without current clinical relevance.

- Another limitation for research is that data analysis only research proposals are not well received at NIH.