PRIVACY AND GENOMIC RESEARCH

William W. Lowrance, Ph.D.

(lowrance@orange.fr)

SHIP Conference,

Exploiting Existing Data

St. Andrews, 10 September 2011

Objectives of genomic research

- Continue to explore genomic factors of health and disease
- Probe tumorigenesis and cancers in more depth (cancer being basically a genomic malfunction)
- Use as an R&D tool in developing preventive, diagnostic, and therapeutic techniques
- Characterize the human microbiome
- Over time, develop genomically personalized medicine and public health genomics.

A resource: Eric Green and NHGRI, "Charting a course for genomic medicine from base pairs to bedside," *Nature* 470, 204-213 (2011).

The challenges posed by the facts

The human genome:

- is extensive and breathtakingly fine-grained: 3,000,000,000 code-bits (the "bases" T, A, G, C)
- is intrinsic to the body and hardly changes during the lifetime, except in cancer cells
- is identically present in every cell of the body, except in red blood cells
- influences most personal attributes
- is unique to the individual, except for identical twins.

What genomic data look like

...TTTCCGTATGCGTAGCCACTTACCCTCCTAGTAGTAGTAG...

through 3,000,000,000 of what statisticians might call data-cells, each carrying T, A, G, or C

Alteration, insertion, or deletion of just a few T, A, G, C can make a big difference, whether the genome is being considered as:

- a dynamic program-tape (of myriad "apps"), or
- a potentially identifying hyper-barcode.

What genomic data look like, *cont*.

at sequence scale: ...ATGTTCGAAATCCGPCTCCCA...

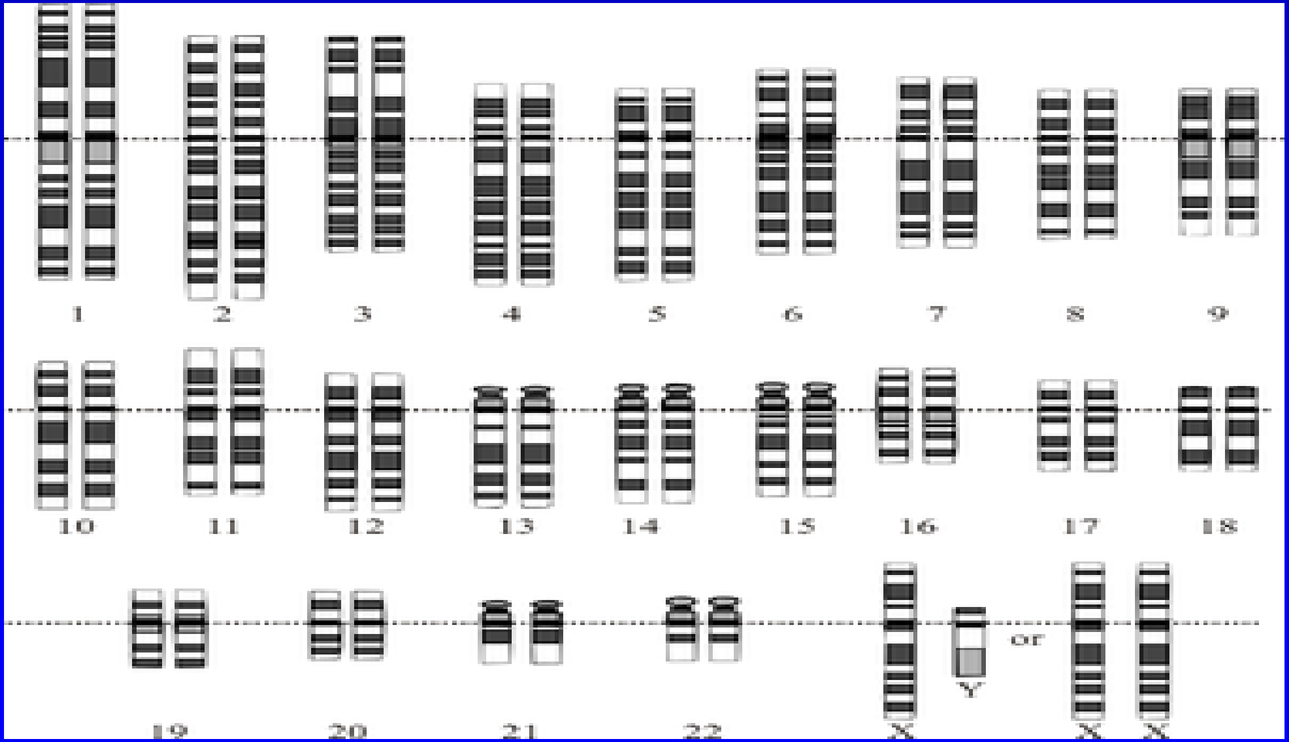
at gene scale: insulin-like growth factor gene IGF2BP2

at body scale: red hair, heritable renal dysplasia

at family scale: ancestry, parentage, family health history, descendants.

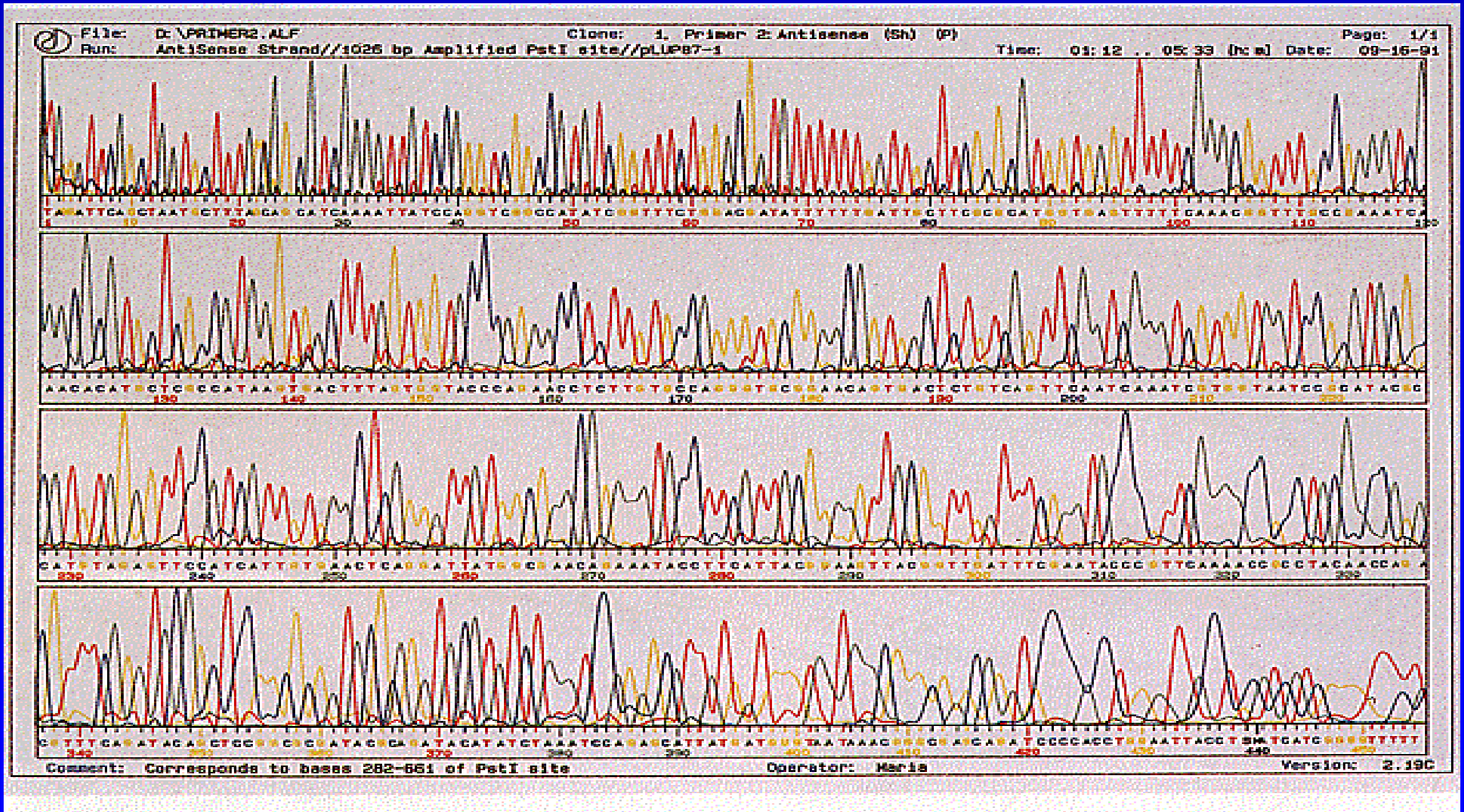
What genomic data look like, *cont.* –

the chromosomes (22 + XX or XY)



What genomic data look like, *cont*. –

a sequence "read"



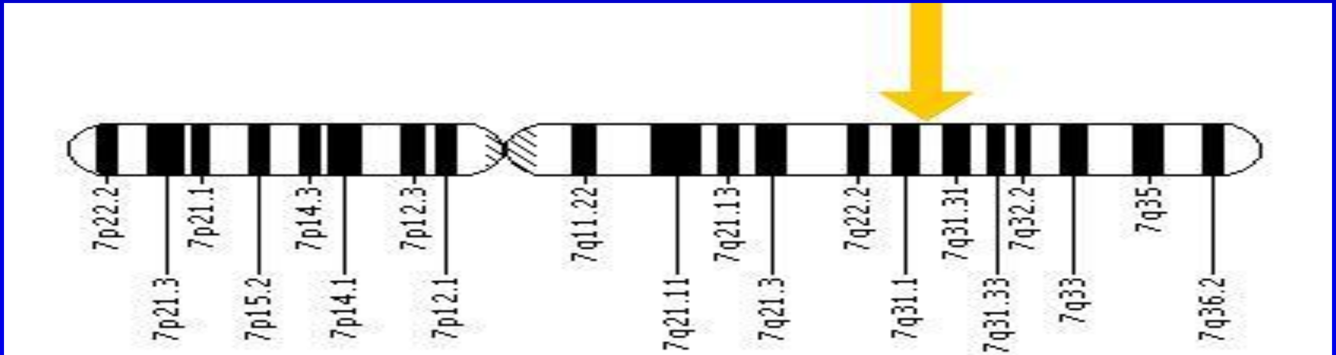
File: D:\PRIMER2.ALF Clone: 1, Primer 2: Antisense (Sh) (P) Page: 1/1
Run: AntiSense Strand//1026 bp Amplified PstI site//pLUP87-1 Time: 01:12 .. 05:33 [h:m] Date: 09-16-91

Comment: Corresponds to bases 282-661 of PstI site Operator: Maria Version: 2.19C

Lowrance, SHIP, 10-9-2011

What genomic data look like, *cont*. −

mutations at site 7q31.2 in the sequence from 17,120,016−17,308,718 that cause cystic fibrosis (autosomal recessive, and several versions)



Sources of DNA and genomic data for research (overlapping)

- small, tightly focused studies
- research cohorts, biobanks, clinical trials
- existing clinical data and/or archived biospecimens
- newborn screening data and/or Guthrie bloodspots
- data-sharing platforms (WT Case Control Consortium, ALSPAC, NIH dbGaP, International Cancer Genomics Consortium, UK Data Archive...)

Example: UK Biobank

- Recruited 500,000 people from the four UK countries, oversampling some minority groups
- Gained broad consent, including permission to link to lifetime NHS Px and Rx data, registries, and other databases, and to genotype as needed
- Collected health and lifestyle data, performed physical exams, collected blood and urine (and from many participants, saliva)
- Conducted full eye exam on 100,000 participants
- Stores specimens in its own high-security −80º robot-retrieval biorepository in Manchester.

UK Biobank, *cont*.

- Operates via an Ethics and Governance Framework
- Is governed by a Board of Directors, and watched over by an independent Ethics and Governance Council (EGC)
- Is a resource to which scientists anywhere and in any sector can apply to use, via restricted access and contractual agreements. A Data Access Committee decides.
- Will be ready for research use soon.

Refs: http://www.ukbiobank.ac.uk; and http://www.egcukbiobank.org.uk

Some challenges with consent in genomic research

- The science is very, very, very hard to comprehend, making "fully informed" consent in the classical sense impossible
- Broad consent can be essential:
  - to make masses of rich resources available for genomic studies (GWAS etc), and
  - to accommodate to the difficulty of precisely defining and limiting research "purposes."

Challenges with consent, *cont*.

- Often it is impossible to predict what findings may emerge, and what their implications may be
- Many findings (or at least the raw data for them) turn up "incidentally"
- Incidental findings can shock: "You know, that nice man you have always called 'Dad'?"

Challenges with consent, *cont*.

- It can be difficult to assess the risks of either identity or trait disclosure in advance, and so is difficult to provide fair notice of the proposition to which people consent
- Possibly, findings can be used against data-subjects' interests
- Findings can have implications (good or bad) for blood relatives, and for other relatives as well − whether or not with their consent
- Genotype data have strong implications for identifiability, a central concern when people are asked to consent.

Challenges with consent, *cont.*
So; there are problems with...

- Comprehensibility
- Breadth of consent
- Purpose specification and limitation
- Privacy risk assessment
- Provision of fair notice
- Incidental findings
- Implications for unaware and unconsenting relatives who may become data-subjects, de facto
- Identifiability.

Identifiability is pivotal!

- If data are identified or identifiable, they may be considered "personal" (or "personally identifiable," etc) data under law

> "Personal data" means data which relate to a living individual who can be identified – (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller... – UK Data Protection Act

- If they are personal data, consent and/or ethics review may be required
- This strongly affects data collecting, transfer, sharing, access, and use

The senses in which genotype data "identify"

- Genotype data don’t "identity" in the name-and-address sense
- If fairly extensive, they are an intrinsic unique tag − which may help match or *single-out*
- And depending, they may allow inferring some descriptive characteristics − and thus *point-to*.

De-identifying genomic data for research

Tactics:

(a) degrading the data before releasing

(b) irreversibly de-identifying the data

(c) separating the identifiers and key-coding.

De-identifying tactic (a): degrading the data before releasing

- can be done, such as by randomly substituting some A for T, or G for C
- almost always degrades usefulness, because most analyses depend on precise fine details.

De-identifying tactic (b): irreversibly de-identifying the data

- is occasionally done, such as:
  - when surveying the background occurrence of some heritable phenomenon
  - when cross-referencing data with corresponding biospecimens, and then destroying the identifiers and all links to the sources
- but obviously has limitations, because can't validate later or recontact.

De-identifying tactic (c): separating the identifiers and key-coding

- is used in all health research, including genomic research
- the equivalent can be performed via a complex data linkage system
- works well − *if* performed carefully, *if* the key is properly safeguarded, and *if* the use of the key to reconnect personal identifiers to the genotype data is strictly controlled
- because de-identification is a matter of degree, restrictions may still be needed, and safeguards are always needed.

The inverse direction:
Identifying non-identified genomic data

Tactics:

(a) matching genotype to identified or identifiable reference genotype data

(b) linking genomic+associated data (clinical, registry, etc) with other data

(c) profiling, i.e. inferring likely appearance, ethnicity, health factors, behavior, or other traits from the genotype.

Identifying tactic (a): matching genotype to identified or identifiable reference genotype data

- may match to hospital, lab, police, military, research, or other collections
- may match to blood-relatives' genotype not in a collection
- is much more certain than matching via other attributes
- proven in criminal forensics and in identifying victims of war, terrorism, and disasters.

Identifying tactic (b): linking genomic+associated data with other data

- is becoming ever easier as databases grow (official ID, demographic, electoral roll, registration, certification, financial, marketing, telecommunication, genealogy, online social broadcasting...)
- often can narrow down to just a few possibilities.

Identifying tactic (c): profiling

- involves inferring traits from genotype
- is only "probabilistic," but is gaining scope and power as genomic science progresses
- now can deduce sex and blood type, likely skin pigmentation, freckling, hair thickness, curl, and color, eye color, basic frame and facial proportions, and some other physical attributes; some health factors; and maybe some behavioral attributes.

Identifiabilty and "human subject" status: the U.S. OHRP policy

"OHRP considers private information or specimens not to be individually identifiable when they cannot be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems." [meaning key-coding]

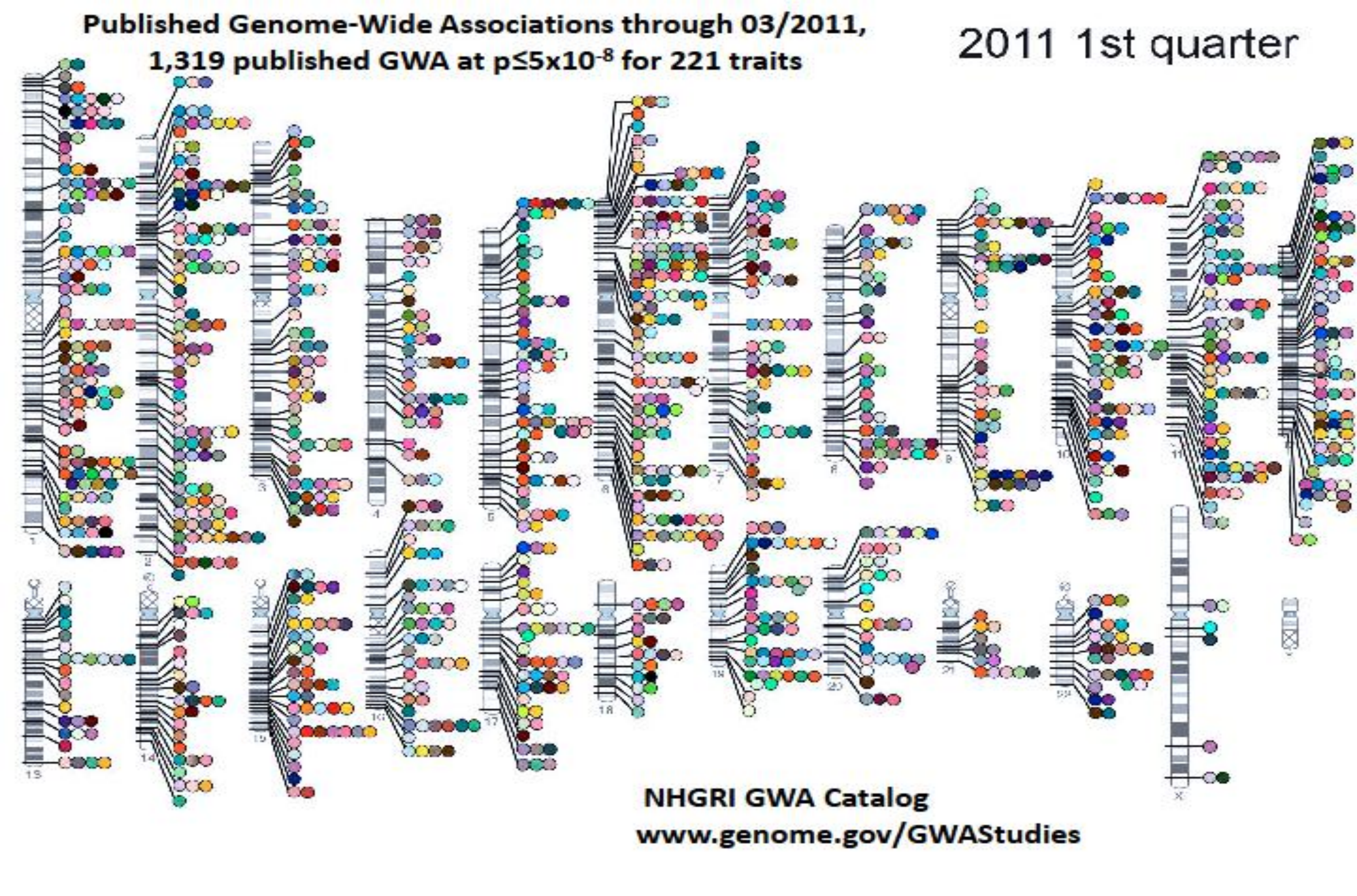
Implication: If data are not identifiable *to the researchers*, there is no "human subject," and so full ethics review and/or consent may be unnecessary.   Important!

Ref: U.S. Office for Human Research Protections (2008); http://www.hhs.gov/ohrp/policy/cdebiol.html

A productive current approach: GWAS

- Genome-wide association studies (GWAS) search for associations between genomic variants and health factors, including drug-response factors
- May survey a million or more markers called SNPs in many thousands of subjects; being facilitated by biobanks and electronic health records
- Scientifically very productive; clinical application is beginning
- Results tend to be shared widely, but identifiability has to be tended to carefully. Most sharing is via restricted access.

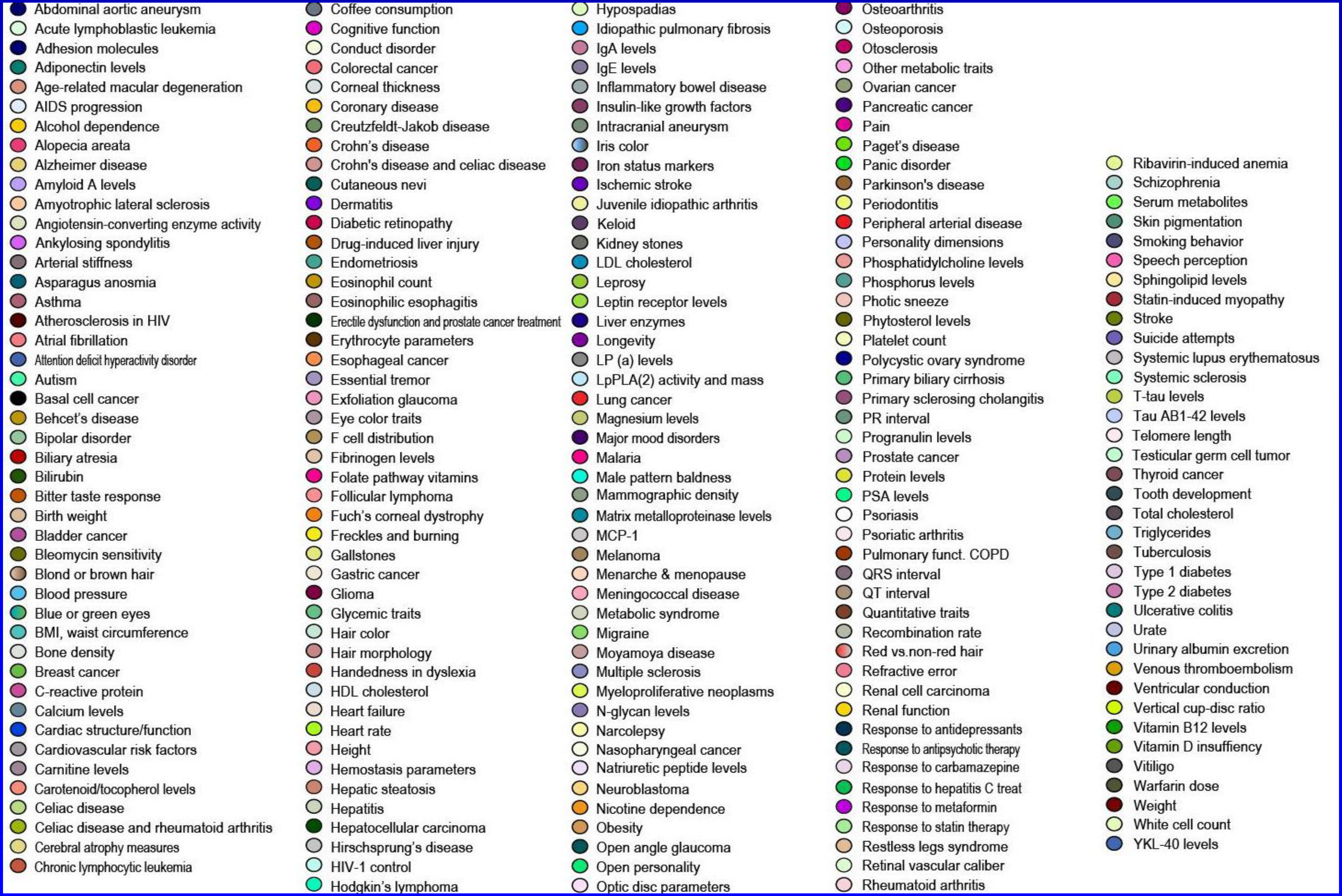
A resource: "Catalog of published genome-wide association studies"; www.genome.gov/26525384.



Published Genome-Wide Associations through 03/2011,
1,319 published GWA at $p \leq 5 \times 10^{-8}$ for 221 traits

2011 1st quarter

NHGRI GWA Catalog
www.genome.gov/GWAStudies



- ● Abdominal aortic aneurysm
- ● Acute lymphoblastic leukemia
- ● Adhesion molecules
- ● Adiponectin levels
- ● Age-related macular degeneration
- ● AIDS progression
- ● Alcohol dependence
- ● Alopecia areata
- ● Alzheimer disease
- ● Amyloid A levels
- ● Amyotrophic lateral sclerosis
- ● Angiotensin-converting enzyme activity
- ● Ankylosing spondylitis
- ● Arterial stiffness
- ● Asparagus anosmia
- ● Asthma
- ● Atherosclerosis in HIV
- ● Atrial fibrillation
- ● Attention deficit hyperactivity disorder
- ● Autism
- ● Basal cell cancer
- ● Behcet's disease
- ● Bipolar disorder
- ● Biliary atresia
- ● Bilirubin
- ● Bitter taste response
- ● Birth weight
- ● Bladder cancer
- ● Bleomycin sensitivity
- ● Blond or brown hair
- ● Blood pressure
- ● Blue or green eyes
- ● BMI, waist circumference
- ● Bone density
- ● Breast cancer
- ● C-reactive protein
- ● Calcium levels
- ● Cardiac structure/function
- ● Cardiovascular risk factors
- ● Carnitine levels
- ● Carotenoid/tocopherol levels
- ● Celiac disease
- ● Celiac disease and rheumatoid arthritis
- ● Cerebral atrophy measures
- ● Chronic lymphocytic leukemia
- ● Coffee consumption
- ● Cognitive function
- ● Conduct disorder
- ● Colorectal cancer
- ● Corneal thickness
- ● Coronary disease
- ● Creutzfeldt-Jakob disease
- ● Crohn's disease
- ● Crohn's disease and celiac disease
- ● Cutaneous nevi
- ● Dermatitis
- ● Diabetic retinopathy
- ● Drug-induced liver injury
- ● Endometriosis
- ● Eosinophil count
- ● Eosinophilic esophagitis
- ● Erectile dysfunction and prostate cancer treatment
- ● Erythrocyte parameters
- ● Esophageal cancer
- ● Essential tremor
- ● Exfoliation glaucoma
- ● Eye color traits
- ● F cell distribution
- ● Fibrinogen levels
- ● Folate pathway vitamins
- ● Follicular lymphoma
- ● Fuch's corneal dystrophy
- ● Freckles and burning
- ● Gallstones
- ● Gastric cancer
- ● Glioma
- ● Glycemic traits
- ● Hair color
- ● Hair morphology
- ● Handedness in dyslexia
- ● HDL cholesterol
- ● Heart failure
- ● Heart rate
- ● Height
- ● Hemostasis parameters
- ● Hepatic steatosis
- ● Hepatitis
- ● Hepatocellular carcinoma
- ● Hirschsprung's disease
- ● HIV-1 control
- ● Hodgkin's lymphoma
- ● Hypospadias
- ● Idiopathic pulmonary fibrosis
- ● IgA levels
- ● IgE levels
- ● Inflammatory bowel disease
- ● Insulin-like growth factors
- ● Intracranial aneurysm
- ● Iris color
- ● Iron status markers
- ● Ischemic stroke
- ● Juvenile idiopathic arthritis
- ● Keloid
- ● Kidney stones
- ● LDL cholesterol
- ● Leprosy
- ● Leptin receptor levels
- ● Liver enzymes
- ● Longevity
- ● LP (a) levels
- ● LpPLA(2) activity and mass
- ● Lung cancer
- ● Magnesium levels
- ● Major mood disorders
- ● Malaria
- ● Male pattern baldness
- ● Mammographic density
- ● Matrix metalloproteinase levels
- ● MCP-1
- ● Melanoma
- ● Menarche & menopause
- ● Meningococcal disease
- ● Metabolic syndrome
- ● Migraine
- ● Moyamoya disease
- ● Multiple sclerosis
- ● Myeloproliferative neoplasms
- ● N-glycan levels
- ● Narcolepsy
- ● Nasopharyngeal cancer
- ● Natriuretic peptide levels
- ● Neuroblastoma
- ● Nicotine dependence
- ● Obesity
- ● Open angle glaucoma
- ● Open personality
- ● Optic disc parameters
- ● Osteoarthritis
- ● Osteoporosis
- ● Otosclerosis
- ● Other metabolic traits
- ● Ovarian cancer
- ● Pancreatic cancer
- ● Pain
- ● Paget's disease
- ● Panic disorder
- ● Parkinson's disease
- ● Periodontitis
- ● Peripheral arterial disease
- ● Personality dimensions
- ● Phosphatidylcholine levels
- ● Phosphorus levels
- ● Photic sneeze
- ● Phytosterol levels
- ● Platelet count
- ● Polycystic ovary syndrome
- ● Primary biliary cirrhosis
- ● Primary sclerosing cholangitis
- ● PR interval
- ● Progranulin levels
- ● Prostate cancer
- ● Protein levels
- ● PSA levels
- ● Psoriasis
- ● Psoriatic arthritis
- ● Pulmonary funct. COPD
- ● QRS interval
- ● QT interval
- ● Quantitative traits
- ● Recombination rate
- ● Red vs.non-red hair
- ● Refractive error
- ● Renal cell carcinoma
- ● Renal function
- ● Response to antidepressants
- ● Response to antipsychotic therapy
- ● Response to carbamazepine
- ● Response to hepatitis C treat
- ● Response to metaformin
- ● Response to statin therapy
- ● Restless legs syndrome
- ● Retinal vascular caliber
- ● Rheumatoid arthritis
- ● Ribavirin-induced anemia
- ● Schizophrenia
- ● Serum metabolites
- ● Skin pigmentation
- ● Smoking behavior
- ● Speech perception
- ● Sphingolipid levels
- ● Statin-induced myopathy
- ● Stroke
- ● Suicide attempts
- ● Systemic lupus erythematosus
- ● Systemic sclerosis
- ● T-tau levels
- ● Tau AB1-42 levels
- ● Telomere length
- ● Testicular germ cell tumor
- ● Thyroid cancer
- ● Tooth development
- ● Total cholesterol
- ● Triglycerides
- ● Tuberculosis
- ● Type 1 diabetes
- ● Type 2 diabetes
- ● Ulcerative colitis
- ● Urate
- ● Urinary albumin excretion
- ● Venous thromboembolism
- ● Ventricular conduction
- ● Vertical cup-disc ratio
- ● Vitamin B12 levels
- ● Vitamin D insuffiency
- ● Vitiligo
- ● Warfarin dose
- ● Weight
- ● White cell count
- ● YKL-40 levels

Lowrance, SHIP, 10-9-2011

Ex: Vanderbilt University BioVU

- Performs GWAS using:
  - clinical samples scheduled to be destroyed
  - a "synthetic derivative" of the data in electronic health records
- Informs patients and the public via notices, brochures, and newspaper articles; provides for easy opting-out at any time
- Randomly excludes 2% of samples so it isn’t possible to know whether a patient is represented in the database
- Manages research access via a tight data use agreement.

(BioVU is part of a 7-center consortium called eMERGE.)

Aggregating isn't what it used to be

- Until 2008, in order to protect the identities of the data-subjects, high-level data from GWAS and other data-sets were posted on the web in aggregated, i.e., pooled, form.
- Then in late 2008 it was shown that individual genotypes can be distinguished in mixtures of DNA from 100 or more people, and so can detect whether a query genotype is present
- Now most extensive genome data are shared via restricted access (under permissions, commitments, and safeguards...), or via highly restricted access (data enclaves).

Ref: Nils Homer et al., *PLoS Genetics* 4(8): e1000167 (2008).

The protections relied on

- De-identification − a long story, but nonetheless...
- Restriction of access, internally and when transferring or sharing data, limiting to certified external researchers, etc.
- Perhaps managing data via resource platforms, elaborate linkage systems, or data enclaves
- Safeguarding, safeguarding, safeguarding
- Penalizing inappropriate releases and uses.
- [Or, releasing data openly but with full and understanding assumption of risk by the data-subjects, as the Personal Genome Project does.]

The Personal Genome Project: a foreshadowing of the future?

- Led by George Church at Harvard, who believes that current consent is illegitimate and that safeguards can't be relied on
- Hopes to recruit 100,000 volunteers (has 1,100 now)
- Puts candidates through an extensive educational process, and then documents consent to having extensive health information and full genome published openly on the web.

See handout...

Refs: http://www.personalgenomes.org; and

John Conley et al., "Enabling responsible public genomics," *Health Matrix* 20, 325-385 (2010); http://www.genomicslawreport.com/wp-content/uploads/ 2011/02/Health_Matrix_-_Journal_of_Law-Medicine_Vol_20_2010.pdf

Opposing extreme models of data access

- At one extreme: Sequester data tightly and manage access, mainly to at least partially de-identified data, via data enclaves (safe havens, research data centers, etc.)
- At the other extreme: Release identified data openly, with consent.
- In-between...?

Two conjectures, inviting reaction

- As genomic science becomes ever more sophisticated, the cost of sequencing and other genotyping techniques continues to drop, genotyping becomes more routine, genomic databases continue to grow, and pedigree data are released publicly by genealogy databases, the identifiability of genomic data generally will increase.
- As genomic data become integrated with, or linked to, electronic health records, disease and other registries, social databases, and research databases, the identifiability of the data in those collections will also tend to increase.

Some continuing issues

- Acceptability of broad consent, the ethical status of data that are not identifiable to the researcher, and so on
- What consent, assent, authorization, or other permission should consist in for complex research
- How to deal with identifiability issues: in policy, and in practice
- Whether any rights adhere to biospecimens or data after they have been thoroughly dissociated from a person
- Ethical obligations to relatives
- Relation of genomics with notions of ancestry, race, ethnicity
- How to manage genotype data linked to networked electronic health records.