BIOINFORMATICS

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 4: GENOME-WIDE ASSOCIATION STUDIES

1 Setting the pace

1.a A hype about GWA studies

1.b Genetic terminology revisited

1.c Genetic association studies

2 Study Designs

2.a Marker level

2.b Subject level

2.c Gender level

3 Preliminary analyses

3.a Quality Control: Hardy-Weinberg equilibrium and missingness

3.b Linkage disequilibrium, haplotypes and SNP tagging

3.c Confounding: population stratification

4 Tests of association

4.a Single SNP

4.b Repeated single SNP tests: Multiple testing correction

4.c Replication

5 Interpretation and follow-up

1 Setting the pace

1.a A hype about GWA studies

“ ‘May he live in interesting times.’

Like it or not we live in interesting times.”

Robert Kennedy, June 7, 1966

How much (sequence) data are available?

- The complete genome sequence of humans and of many other species provides a new starting point for understanding our basic genetic makeup and how variations in our genetic instructions result in disease.

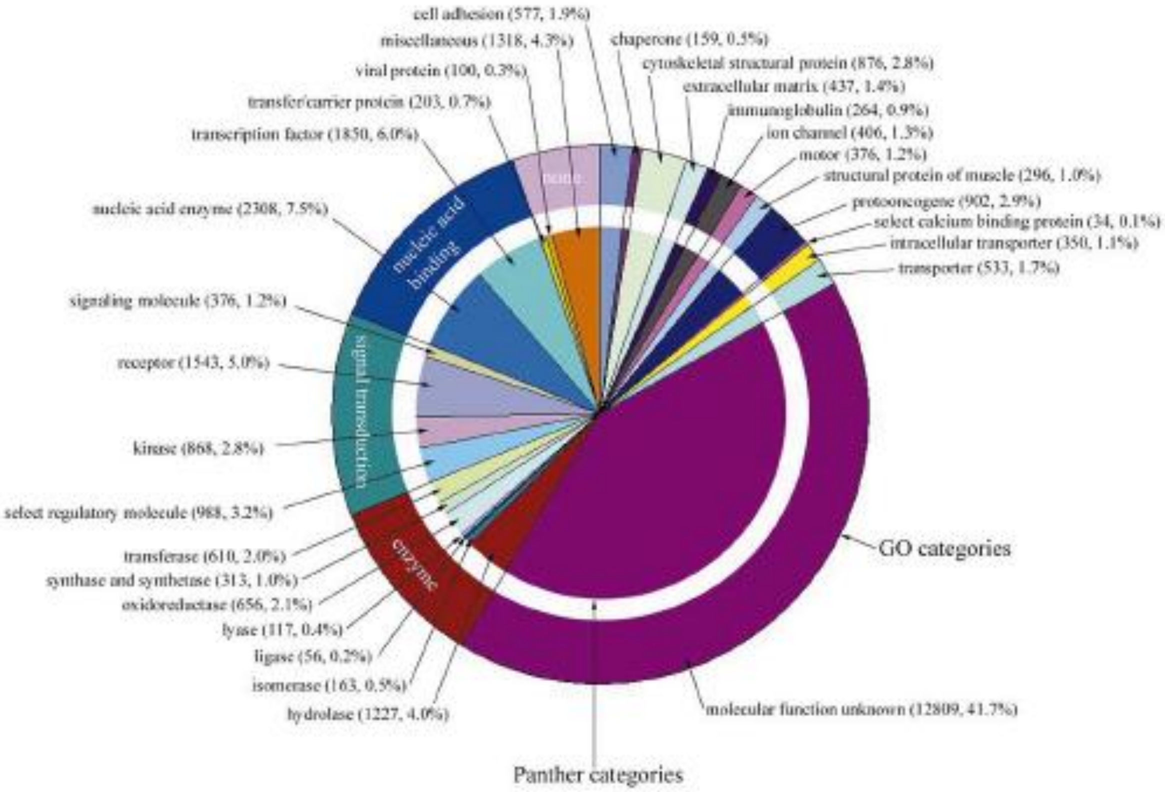
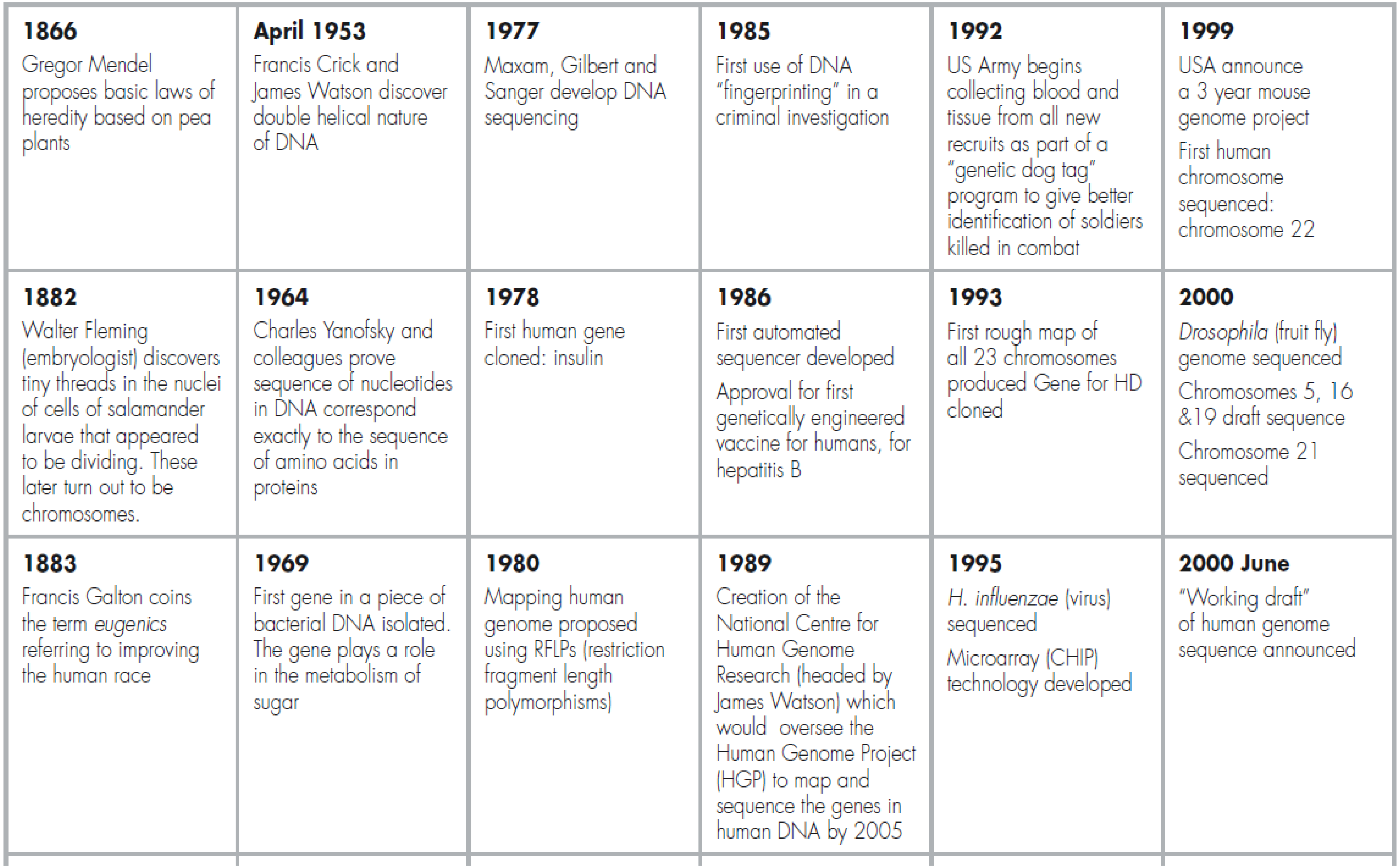
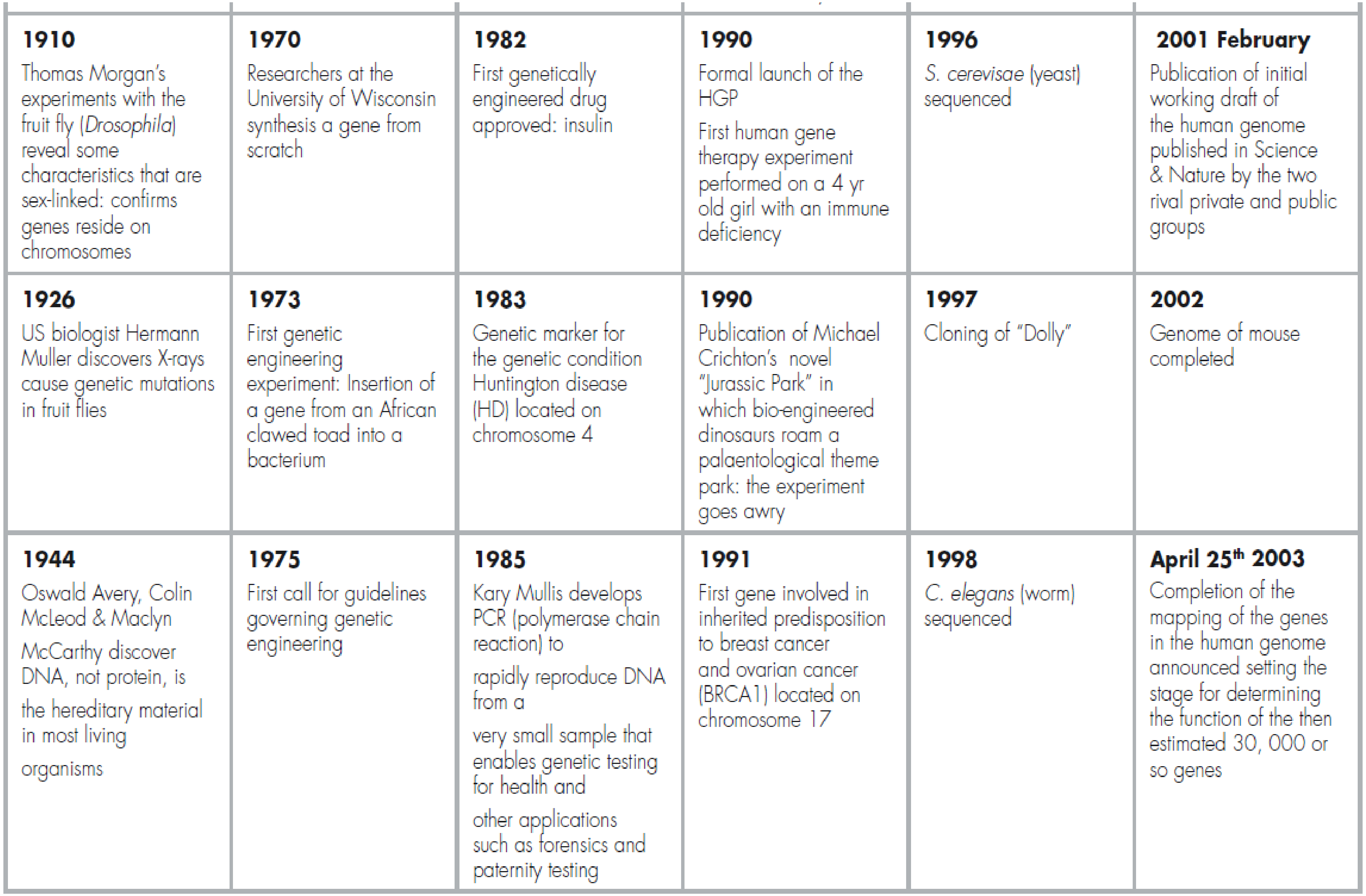


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

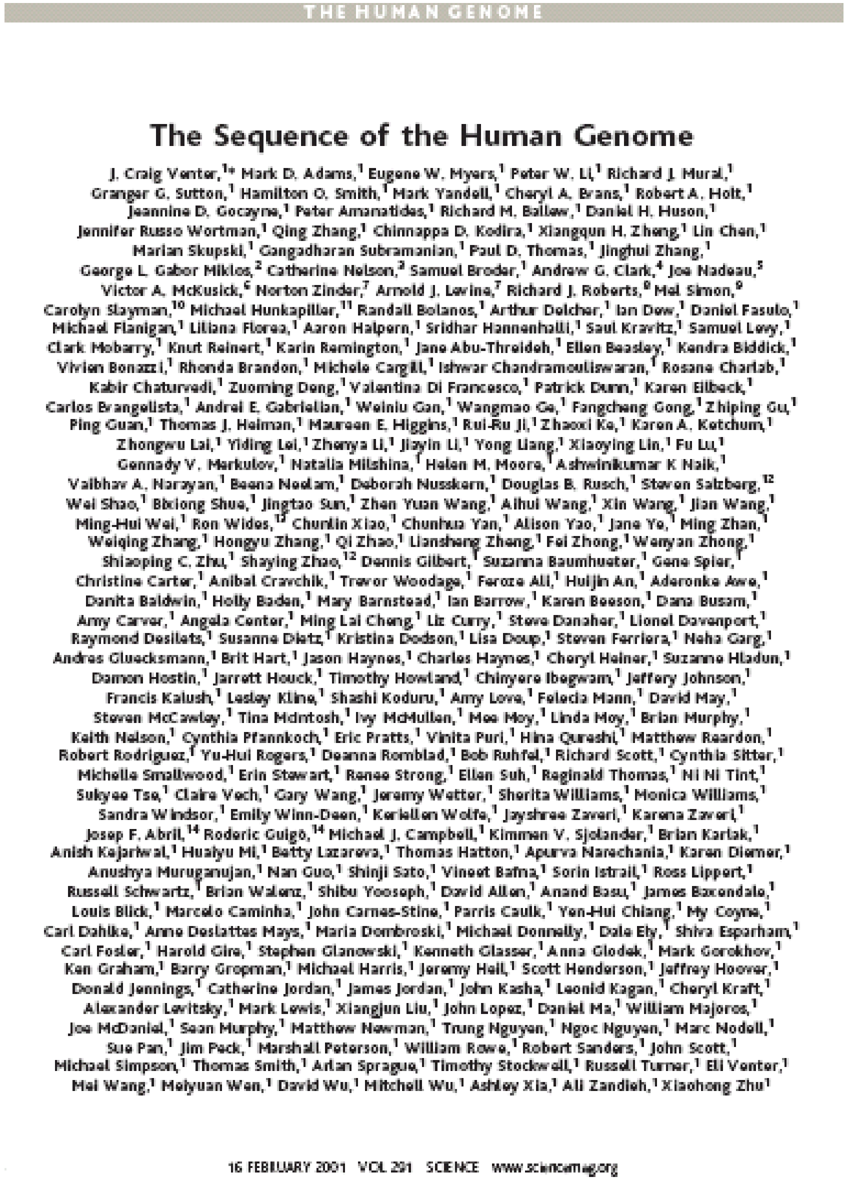
Table 24.1: The history of human genetics discoveries up to the 50th anniversary of the discovery of the DNA helical structure in 1953.



1866 Gregor Mendel proposes basic laws of heredity based on pea plants	**April 1953** Francis Crick and James Watson discover double helical nature of DNA	**1977** Maxam, Gilbert and Sanger develop DNA sequencing	**1985** First use of DNA "fingerprinting" in a criminal investigation	**1992** US Army begins collecting blood and tissue from all new recruits as part of a "genetic dog tag" program to give better identification of soldiers killed in combat	**1999** USA announce a 3 year mouse genome project First human chromosome sequenced: chromosome 22
1882 Walter Fleming (embryologist) discovers tiny threads in the nuclei of cells of salamander larvae that appeared to be dividing. These later turn out to be chromosomes.	**1964** Charles Yanofsky and colleagues prove sequence of nucleotides in DNA correspond exactly to the sequence of amino acids in proteins	**1978** First human gene cloned: insulin	**1986** First automated sequencer developed Approval for first genetically engineered vaccine for humans, for hepatitis B	**1993** First rough map of all 23 chromosomes produced Gene for HD cloned	**2000** *Drosophila* (fruit fly) genome sequenced Chromosomes 5, 16 &19 draft sequence Chromosome 21 sequenced
1883 Francis Galton coins the term *eugenics* referring to improving the human race	**1969** First gene in a piece of bacterial DNA isolated. The gene plays a role in the metabolism of sugar	**1980** Mapping human genome proposed using RFLPs (restriction fragment length polymorphisms)	**1989** Creation of the National Centre for Human Genome Research (headed by James Watson) which would oversee the Human Genome Project (HGP) to map and sequence the genes in human DNA by 2005	**1995** *H. influenzae* (virus) sequenced Microarray (CHIP) technology developed	**2000 June** "Working draft" of human genome sequence announced



1910 Thomas Morgan's experiments with the fruit fly (*Drosophila*) reveal some characteristics that are sex-linked: confirms genes reside on chromosomes	**1970** Researchers at the University of Wisconsin synthesis a gene from scratch	**1982** First genetically engineered drug approved: insulin	**1990** Formal launch of the HGP First human gene therapy experiment performed on a 4 yr old girl with an immune deficiency	**1996** *S. cerevisae* (yeast) sequenced	**2001 February** Publication of initial working draft of the human genome published in Science & Nature by the two rival private and public groups
1926 US biologist Hermann Muller discovers X-rays cause genetic mutations in fruit flies	**1973** First genetic engineering experiment: Insertion of a gene from an African clawed toad into a bacterium	**1983** Genetic marker for the genetic condition Huntington disease (HD) located on chromosome 4	**1990** Publication of Michael Crichton's novel "Jurassic Park" in which bio-engineered dinosaurs roam a palaentological theme park: the experiment goes awry	**1997** Cloning of "Dolly"	**2002** Genome of mouse completed
1944 Oswald Avery, Colin McLeod & Maclyn McCarthy discover DNA, not protein, is the hereditary material in most living organisms	**1975** First call for guidelines governing genetic engineering	**1985** Kary Mullis develops PCR (polymerase chain reaction) to rapidly reproduce DNA from a very small sample that enables genetic testing for health and other applications such as forensics and paternity testing	**1991** First gene involved in inherited predisposition to breast cancer and ovarian cancer (BRCA1) located on chromosome 17	**1998** *C. elegans* (worm) sequenced	**April 25th 2003** Completion of the mapping of the genes in the human genome announced setting the stage for determining the function of the then estimated 30, 000 or so genes

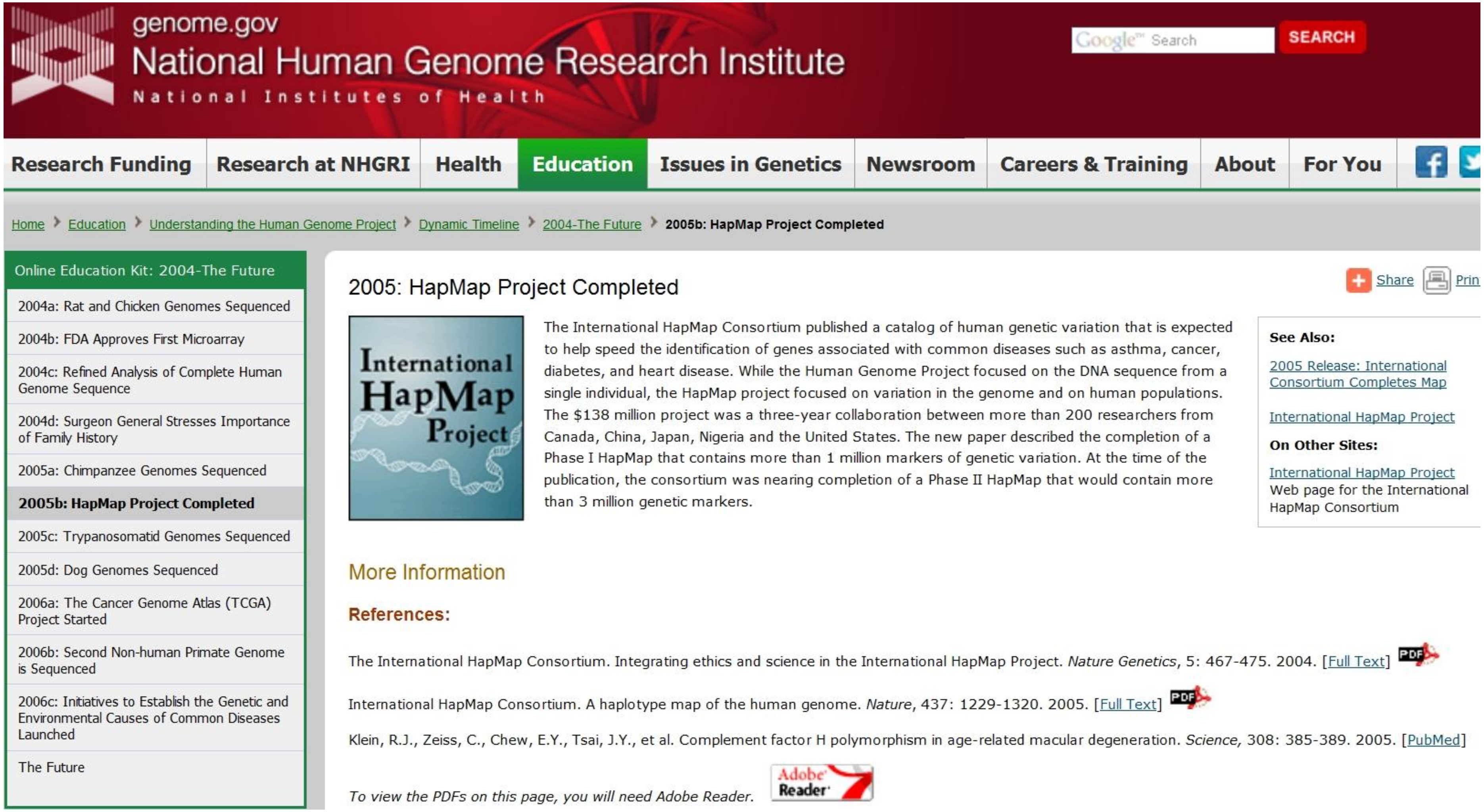


THE HUMAN GENOME

The Sequence of the Human Genome

J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]
Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5]
Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9]
Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1]
Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1]
Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1]
Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1]
Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1]
Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1]
Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1]
Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]
Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1]
Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12]
Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1]
Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1]
Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1]
Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1]
Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1]
Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1]
Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1]
Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1]
Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1]
Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1]
Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1]
Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1]
Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1]
Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1]
Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1]
Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1]
Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1]
Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1]
Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1]
Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1]
Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1]
Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1]
Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1]
Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1]
Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1]
Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1]
Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1]
Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1]
Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1]
Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1]
Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

16 FEBRUARY 2001 VOL 291 SCIENCE www.sciencemag.org


genome.gov
National Human Genome Research Institute
National Institutes of Health
Google™ Search
SEARCH
Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
For You
Home
Education
Understanding the Human Genome Project
Dynamic Timeline
2004-The Future
2005b: HapMap Project Completed
Online Education Kit: 2004-The Future
2004a: Rat and Chicken Genomes Sequenced
2004b: FDA Approves First Microarray
2004c: Refined Analysis of Complete Human Genome Sequence
2004d: Surgeon General Stresses Importance of Family History
2005a: Chimpanzee Genomes Sequenced
2005b: HapMap Project Completed
2005c: Trypanosomatid Genomes Sequenced
2005d: Dog Genomes Sequenced
2006a: The Cancer Genome Atlas (TCGA) Project Started
2006b: Second Non-human Primate Genome is Sequenced
2006c: Initiatives to Establish the Genetic and Environmental Causes of Common Diseases Launched
The Future
Share
Prin
International HapMap Project

2005: HapMap Project Completed

The International HapMap Consortium published a catalog of human genetic variation that is expected to help speed the identification of genes associated with common diseases such as asthma, cancer, diabetes, and heart disease. While the Human Genome Project focused on the DNA sequence from a single individual, the HapMap project focused on variation in the genome and on human populations. The $138 million project was a three-year collaboration between more than 200 researchers from Canada, China, Japan, Nigeria and the United States. The new paper described the completion of a Phase I HapMap that contains more than 1 million markers of genetic variation. At the time of the publication, the consortium was nearing completion of a Phase II HapMap that would contain more than 3 million genetic markers.

See Also:

2005 Release: International Consortium Completes Map

International HapMap Project

On Other Sites:

International HapMap Project
Web page for the International HapMap Consortium

More Information

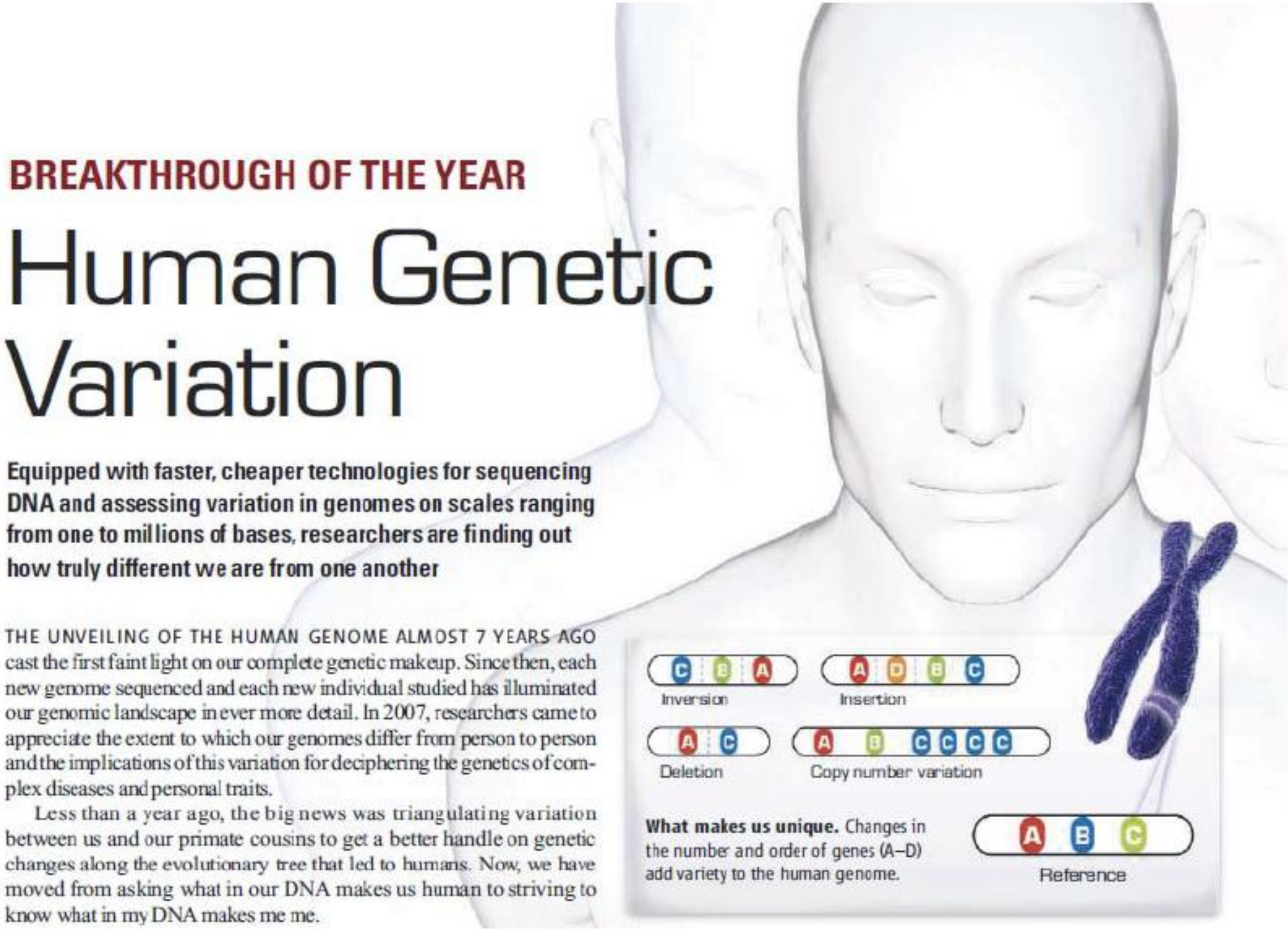
References:

The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Genetics*, 5: 467-475. 2004. [Full Text] PDF

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437: 1229-1320. 2005. [Full Text] PDF

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., et al. Complement factor H polymorphism in age-related macular degeneration. *Science,* 308: 385-389. 2005. [PubMed]

To view the PDFs on this page, you will need Adobe Reader. Adobe Reader



BREAKTHROUGH OF THE YEAR

Human Genetic Variation

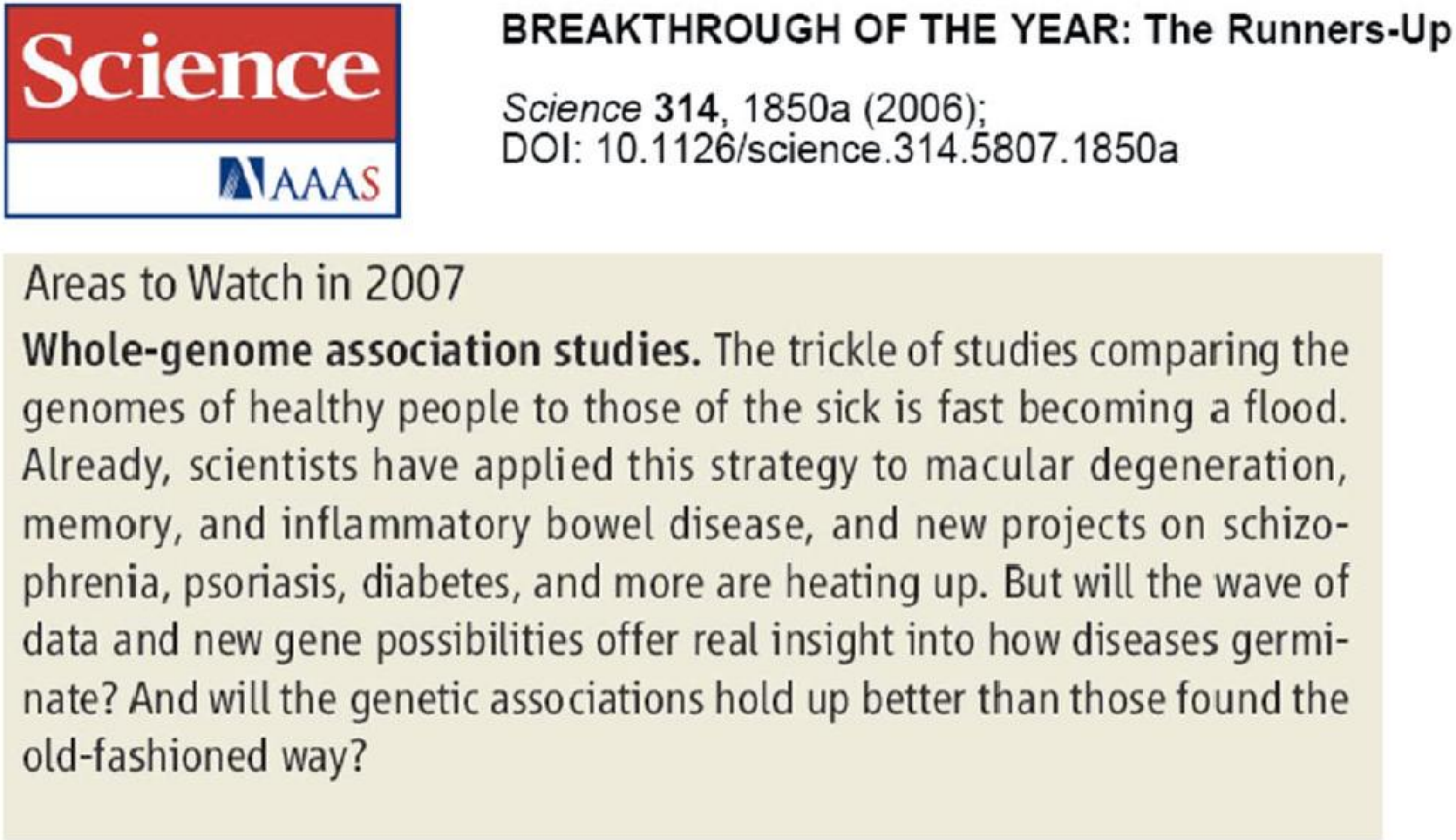
Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.

What makes us unique. Changes in the number and order of genes (A–D) add variety to the human genome.

Pennisi 2007 Science 318:1842-3



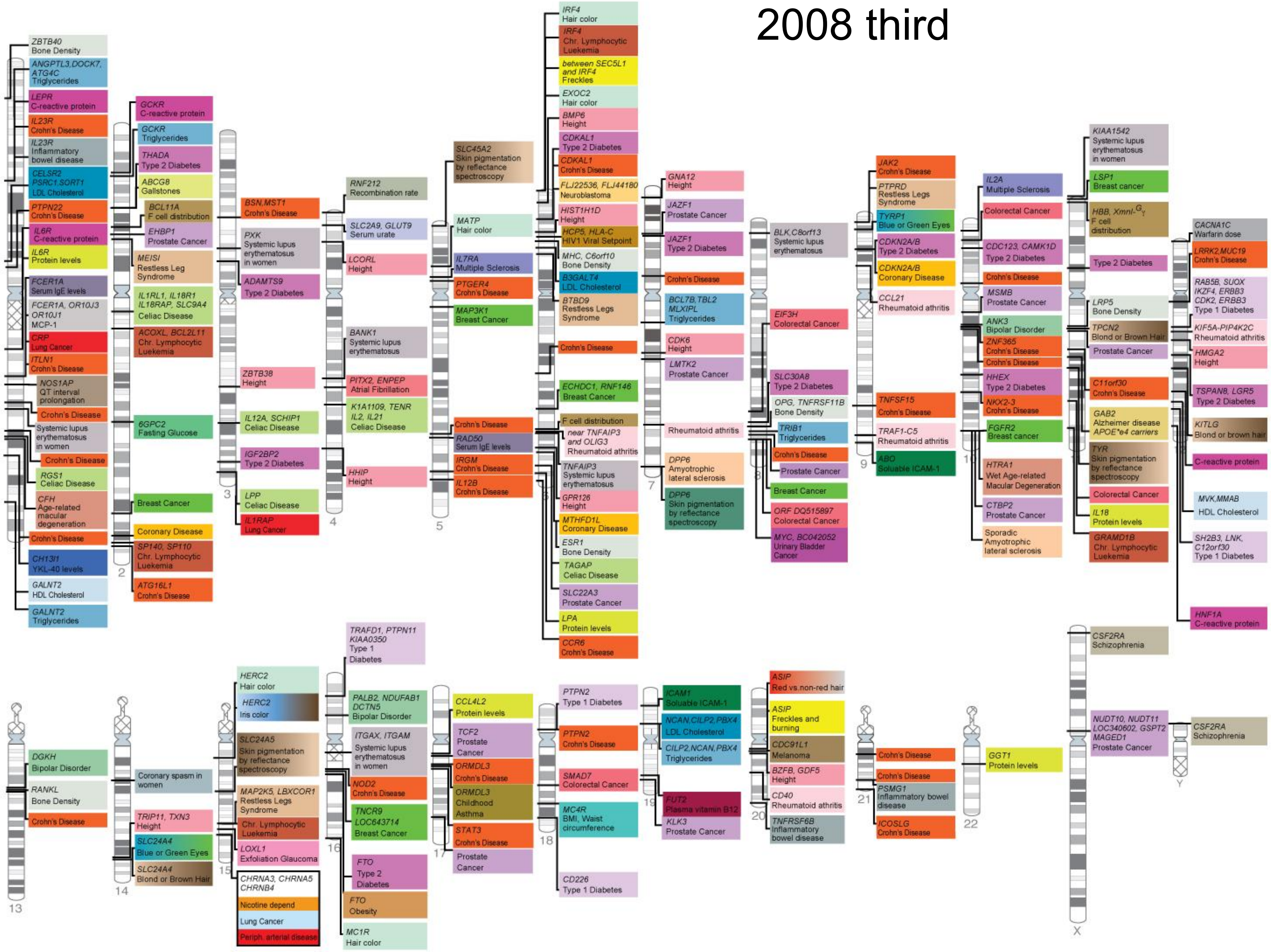
BREAKTHROUGH OF THE YEAR: The Runners-Up

Science **314**, 1850a (2006);
DOI: 10.1126/science.314.5807.1850a

Areas to Watch in 2007

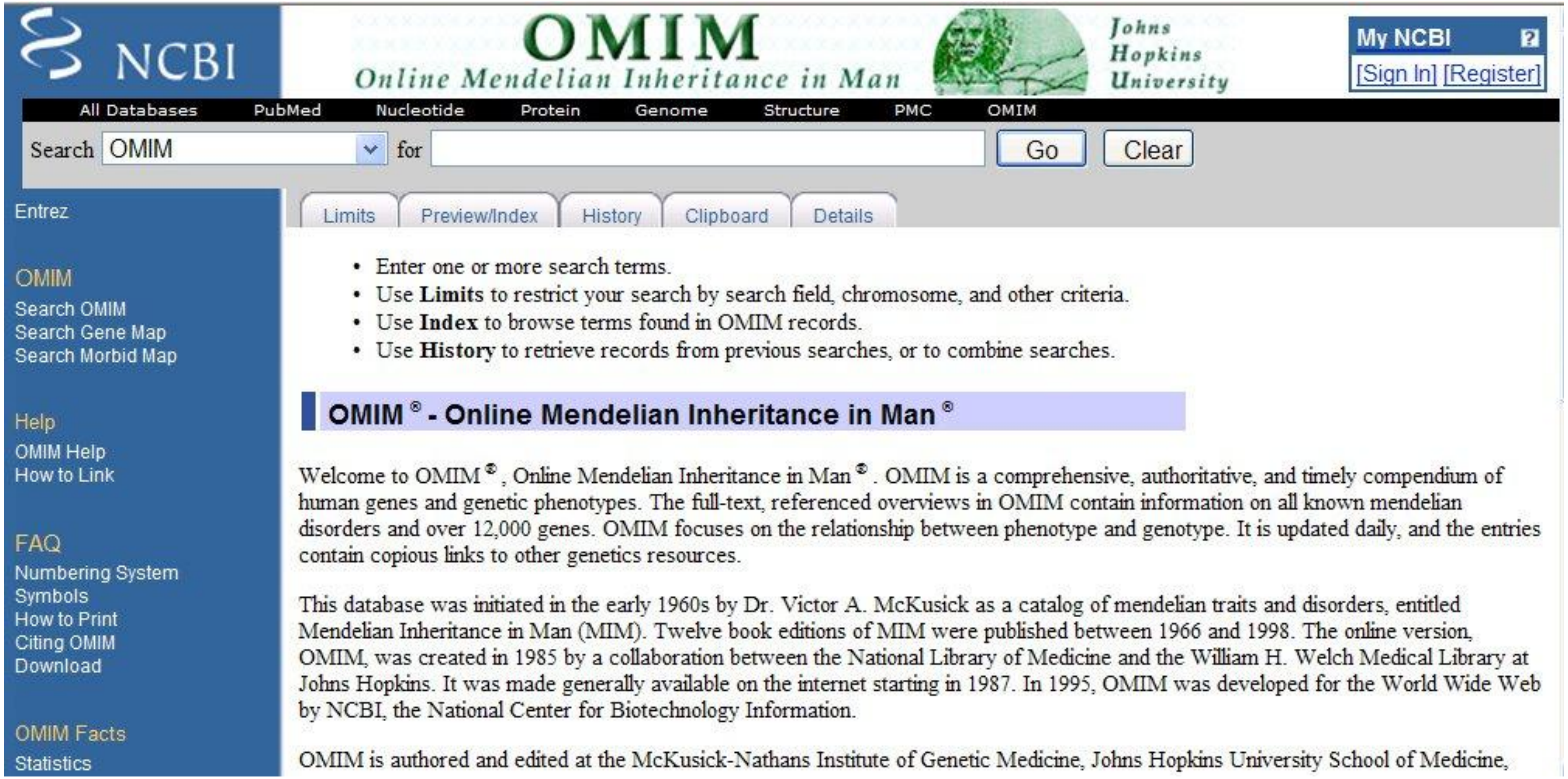
Whole-genome association studies. The trickle of studies comparing the genomes of healthy people to those of the sick is fast becoming a flood. Already, scientists have applied this strategy to macular degeneration, memory, and inflammatory bowel disease, and new projects on schizophrenia, psoriasis, diabetes, and more are heating up. But will the wave of data and new gene possibilities offer real insight into how diseases germinate? And will the genetic associations hold up better than those found the old-fashioned way?

2008 third


ZBTB40 Bone Density
ANGPTL3,DOCK7, ATG4C Triglycerides
LEPR C-reactive protein
IL23R Crohn's Disease
IL23R Inflammatory bowel disease
CELSR2 PSRC1,SORT1 LDL Cholesterol
PTPN22 Crohn's Disease
IL6R C-reactive protein
IL6R Protein levels
FCER1A Serum IgE levels
FCER1A, OR10J3 OR10J1 MCP-1
CRP Lung Cancer
ITLN1 Crohn's Disease
NOS1AP QT interval prolongation
Crohn's Disease
Systemic lupus erythematosus in women
Crohn's Disease
RGS1 Celiac Disease
CFH Age-related macular degeneration
Crohn's Disease
CH13l1 YKL-40 levels
GALNT2 HDL Cholesterol
GALNT2 Triglycerides
1
GCKR C-reactive protein
GCKR Triglycerides
THADA Type 2 Diabetes
ABCG8 Gallstones
BCL11A F cell distribution
EHBP1 Prostate Cancer
MEIS1 Restless Leg Syndrome
IL1RL1, IL18R1 IL18RAP, SLC9A4 Celiac Disease
ACOXL, BCL2L11 Chr. Lymphocytic Luekemia
6GPC2 Fasting Glucose
Breast Cancer
Coronary Disease
SP140, SP110 Chr. Lymphocytic Luekemia
ATG16L1 Crohn's Disease
2
BSN,MST1 Crohn's Disease
PXK Systemic lupus erythematosus in women
ADAMTS9 Type 2 Diabetes
ZBTB38 Height
IL12A, SCHIP1 Celiac Disease
IGF2BP2 Type 2 Diabetes
LPP Celiac Disease
IL1RAP Lung Cancer
3
RNF212 Recombination rate
SLC2A9, GLUT9 Serum urate
LCORL Height
BANK1 Systemic lupus erythematosus
PITX2, ENPEP Atrial Fibrillation
K1A1109, TENR IL2, IL21 Celiac Disease
HHIP Height
4
SLC45A2 Skin pigmentation by reflectance spectroscopy
MATP Hair color
IL7RA Multiple Sclerosis
PTGER4 Crohn's Disease
MAP3K1 Breast Cancer
Crohn's Disease
RAD50 Serum IgE levels
IRGM Crohn's Disease
IL12B Crohn's Disease
5
IRF4 Hair color
IRF4 Chr. Lymphocytic Luekemia
between SEC5L1 and IRF4 Freckles
EXOC2 Hair color
BMP6 Height
CDKAL1 Type 2 Diabetes
CDKAL1 Crohn's Disease
FLJ22536, FLJ44180 Neuroblastoma
HIST1H1D Height
HCP5, HLA-C HIV1 Viral Setpoint
MHC, C6orf10 Bone Density
B3GALT4 LDL Cholesterol
BTBD9 Restless Legs Syndrome
Crohn's Disease
ECHDC1, RNF146 Breast Cancer
F cell distribution
near TNFAIP3 and OLIG3 Rheumatoid athritis
TNFAIP3 Systemic lupus erythematosus
GPR126 Height
MTHFD1L Coronary Disease
ESR1 Bone Density
TAGAP Celiac Disease
SLC22A3 Prostate Cancer
LPA Protein levels
CCR6 Crohn's Disease
6
GNA12 Height
JAZF1 Prostate Cancer
JAZF1 Type 2 Diabetes
Crohn's Disease
BCL7B,TBL2 MLXIPL Triglycerides
CDK6 Height
LMTK2 Prostate Cancer
Rheumatoid athritis
DPP6 Amyotrophic lateral sclerosis
DPP6 Skin pigmentation by reflectance spectroscopy
7
BLK,C8orf13 Systemic lupus erythematosus
EIF3H Colorectal Cancer
SLC30A8 Type 2 Diabetes
OPG, TNFRSF11B Bone Density
TRIB1 Triglycerides
Crohn's Disease
Prostate Cancer
Breast Cancer
ORF DQ515897 Colorectal Cancer
MYC, BC042052 Urinary Bladder Cancer
8
JAK2 Crohn's Disease
PTPRD Restless Legs Syndrome
TYRP1 Blue or Green Eyes
CDKN2A/B Type 2 Diabetes
CDKN2A/B Coronary Disease
CCL21 Rheumatoid athritis
TNFSF15 Crohn's Disease
TRAF1-C5 Rheumatoid athritis
ABO Soluable ICAM-1
9
IL2A Multiple Sclerosis
Colorectal Cancer
CDC123, CAMK1D Type 2 Diabetes
Crohn's Disease
MSMB Prostate Cancer
ANK3 Bipolar Disorder
ZNF365 Crohn's Disease
Crohn's Disease
HHEX Type 2 Diabetes
NKX2-3 Crohn's Disease
FGFR2 Breast cancer
HTRA1 Wet Age-related Macular Degeneration
CTBP2 Prostate Cancer
Sporadic Amyotrophic lateral sclerosis
10
KIAA1542 Systemic lupus erythematosus in women
LSP1 Breast cancer
HBB, Xmnl-G_γ F cell distribution
Type 2 Diabetes
LRP5 Bone Density
TPCN2 Blond or Brown Hair
Prostate Cancer
C11orf30 Crohn's Disease
GAB2 Alzheimer disease APOE*e4 carriers
TYR Skin pigmentation by reflectance spectroscopy
Colorectal Cancer
IL18 Protein levels
GRAMD1B Chr. Lymphocytic Luekemia
11
CACNA1C Warfarin dose
LRRK2,MUC19 Crohn's Disease
RAB5B, SUOX IKZF4, ERBB3 CDK2, ERBB3 Type 1 Diabetes
KIF5A-PIP4K2C Rheumatoid athritis
HMGA2 Height
TSPAN8, LGR5 Type 2 Diabetes
KITLG Blond or brown hair
C-reactive protein
MVK,MMAB HDL Cholesterol
SH2B3, LNK, C12orf30 Type 1 Diabetes
HNF1A C-reactive protein
12
CSF2RA Schizophrenia
NUDT10, NUDT11 LOC340602, GSPT2 MAGED1 Prostate Cancer
X
CSF2RA Schizophrenia
Y
DGKH Bipolar Disorder
RANKL Bone Density
Crohn's Disease
13
Coronary spasm in women
TRIP11, TXN3 Height
SLC24A4 Blue or Green Eyes
SLC24A4 Blond or Brown Hair
14
HERC2 Hair color
HERC2 Iris color
SLC24A5 Skin pigmentation by reflectance spectroscopy
MAP2K5, LBXCOR1 Restless Legs Syndrome
Chr. Lymphocytic Luekemia
LOXL1 Exfoliation Glaucoma
CHRNA3, CHRNA5 CHRNB4
Nicotine depend
Lung Cancer
Periph. arterial disease
15
TRAFD1, PTPN11 KIAA0350 Type 1 Diabetes
PALB2, NDUFAB1 DCTN5 Bipolar Disorder
ITGAX, ITGAM Systemic lupus erythematosus in women
NOD2 Crohn's Disease
TNCR9 LOC643714 Breast Cancer
FTO Type 2 Diabetes
FTO Obesity
MC1R Hair color
16
CCL4L2 Protein levels
TCF2 Prostate Cancer
ORMDL3 Crohn's Disease
ORMDL3 Childhood Asthma
STAT3 Crohn's Disease
Prostate Cancer
17
PTPN2 Type 1 Diabetes
PTPN2 Crohn's Disease
SMAD7 Colorectal Cancer
MC4R BMI, Waist circumference
CD226 Type 1 Diabetes
18
ICAM1 Soluable ICAM-1
NCAN,CILP2,PBX4 LDL Cholesterol
CILP2,NCAN,PBX4 Triglycerides
FUT2 Plasma vitamin B12
KLK3 Prostate Cancer
19
ASIP Red vs.non-red hair
ASIP Freckles and burning
CDC91L1 Melanoma
BZFB, GDF5 Height
CD40 Rheumatoid athritis
TNFRSF6B Inflammatory bowel disease
20
Crohn's Disease
Crohn's Disease
PSMG1 Inflammatory bowel disease
ICOSLG Crohn's Disease
21
GGT1 Protein levels
22

- The pace of the molecular dissection of human disease can be measured by looking at the catalog of human genes and genetic disorders identified so far in *OMIM*, which is updated daily (www.ncbi.nlm.nih.gov/omim).

(V. A. McKusick, Mendelian Inheritance in Man (Johns Hopkins Univ. Press, Baltimore, ed. 12, 1998))



- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and other criteria.
- Use **Index** to browse terms found in OMIM records.
- Use **History** to retrieve records from previous searches, or to combine searches.

OMIM ® - Online Mendelian Inheritance in Man ®

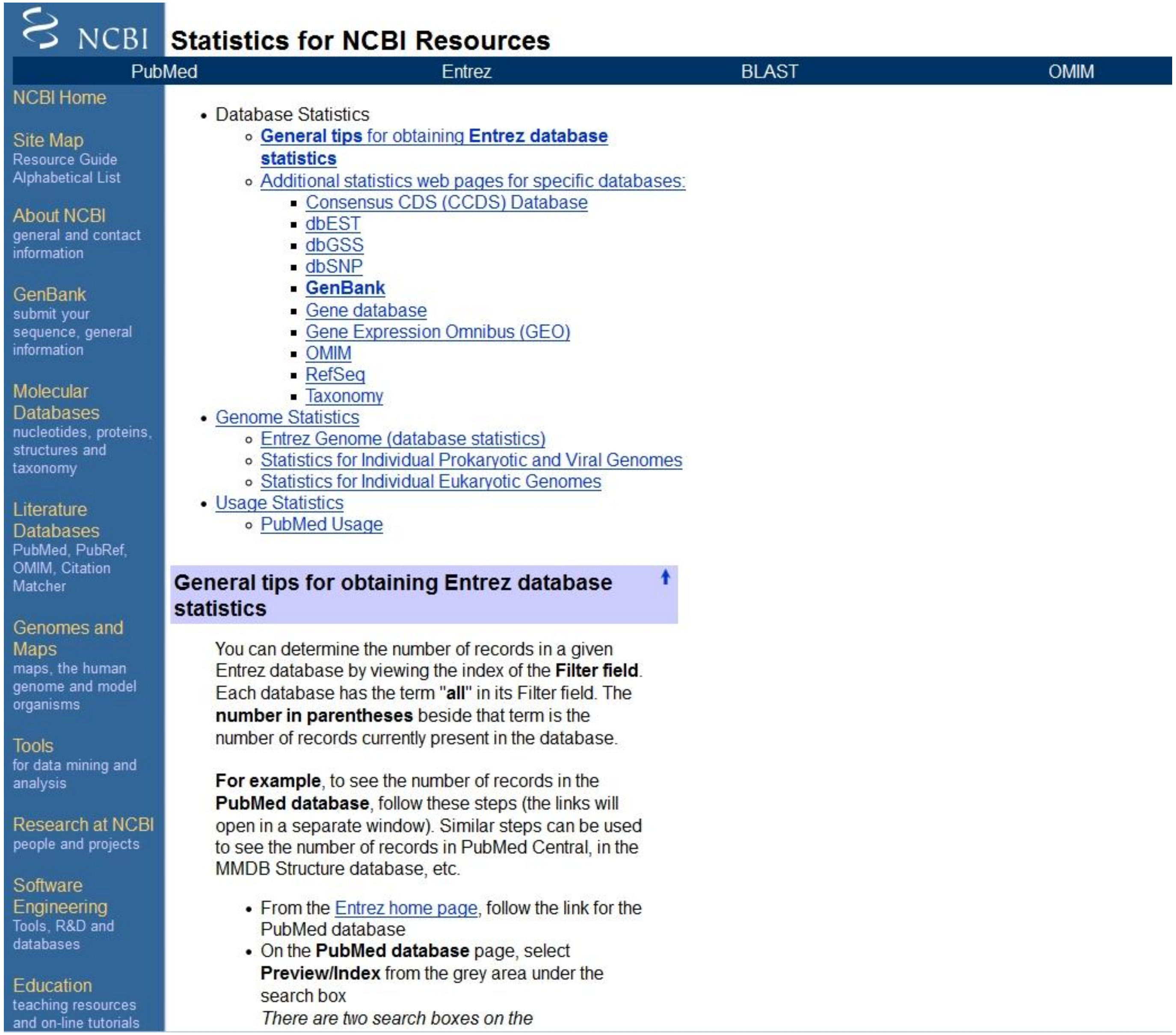
Welcome to OMIM ®, Online Mendelian Inheritance in Man ®. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.

OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine,

What is OMIM?

- Online Mendelian Inheritance in Man (OMIM®) is a continuously updated catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression.
- It is thus considered to be a phenotypic companion to the Human Genome Project. OMIM is a continuation of Dr. Victor A. McKusick's Mendelian Inheritance in Man, which was published through 12 editions, the last in 1998.
- OMIM is currently biocurated at the McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine.
- Frequently asked questions: http://www.omim.org/help/faq



NCBI
PubMed
Entrez
BLAST
OMIM
NCBI Home
Site Map
Resource Guide
Alphabetical List
About NCBI
general and contact information
GenBank
submit your sequence, general information
Molecular Databases
nucleotides, proteins, structures and taxonomy
Literature Databases
PubMed, PubRef, OMIM, Citation Matcher
Genomes and Maps
maps, the human genome and model organisms
Tools
for data mining and analysis
Research at NCBI
people and projects
Software Engineering
Tools, R&D and databases
Education
teaching resources and on-line tutorials

Statistics for NCBI Resources

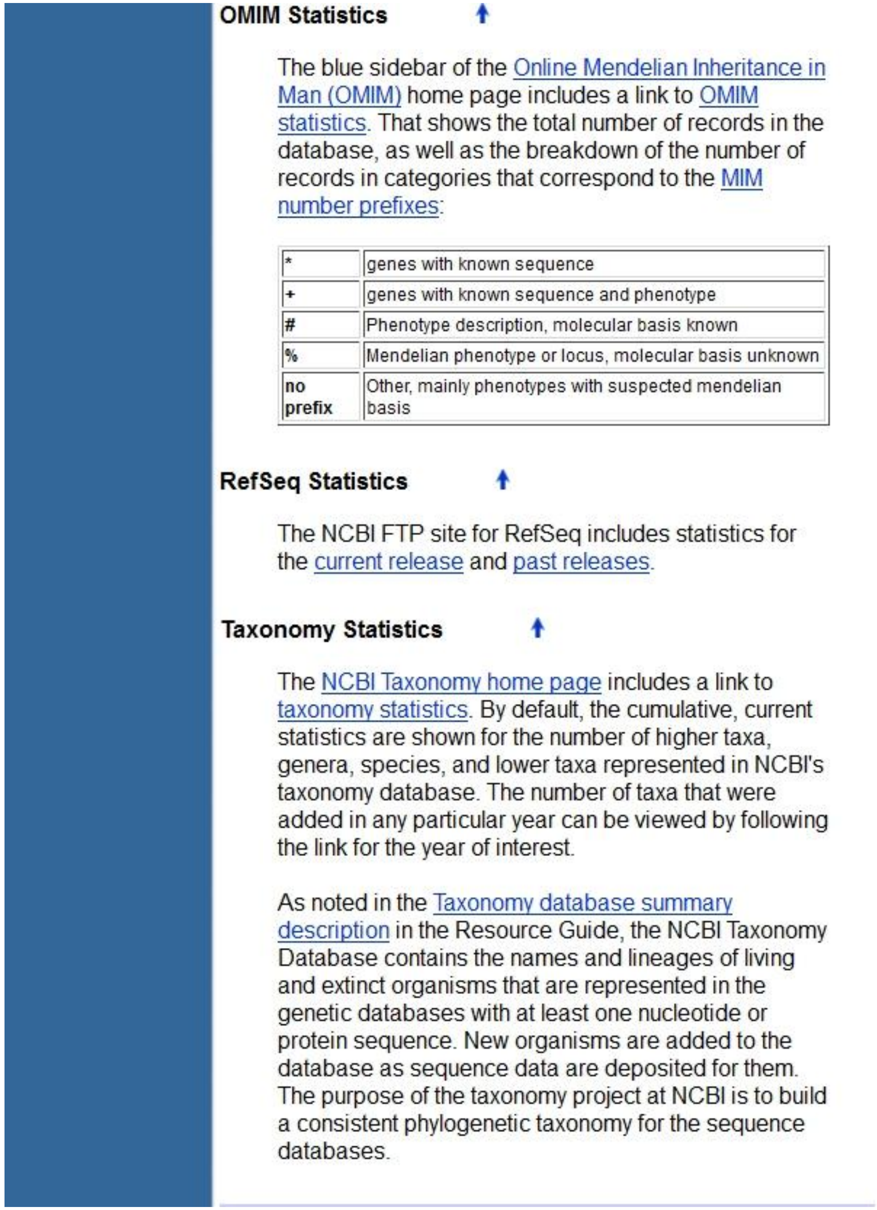
- Database Statistics
  - **General tips** for obtaining **Entrez database statistics**
  - Additional statistics web pages for specific databases:
    - Consensus CDS (CCDS) Database
    - dbEST
    - dbGSS
    - dbSNP
    - **GenBank**
    - Gene database
    - Gene Expression Omnibus (GEO)
    - OMIM
    - RefSeq
    - Taxonomy
- Genome Statistics
  - Entrez Genome (database statistics)
  - Statistics for Individual Prokaryotic and Viral Genomes
  - Statistics for Individual Eukaryotic Genomes
- Usage Statistics
  - PubMed Usage

General tips for obtaining Entrez database statistics

You can determine the number of records in a given Entrez database by viewing the index of the **Filter field**. Each database has the term "**all**" in its Filter field. The **number in parentheses** beside that term is the number of records currently present in the database.

For example, to see the number of records in the **PubMed database**, follow these steps (the links will open in a separate window). Similar steps can be used to see the number of records in PubMed Central, in the MMDB Structure database, etc.

- From the Entrez home page, follow the link for the PubMed database
- On the **PubMed database** page, select **Preview/Index** from the grey area under the search box
  There are two search boxes on the



OMIM Statistics

The blue sidebar of the Online Mendelian Inheritance in Man (OMIM) home page includes a link to OMIM statistics. That shows the total number of records in the database, as well as the breakdown of the number of records in categories that correspond to the MIM number prefixes:

*	genes with known sequence
+	genes with known sequence and phenotype
#	Phenotype description, molecular basis known
%	Mendelian phenotype or locus, molecular basis unknown
no prefix	Other, mainly phenotypes with suspected mendelian basis

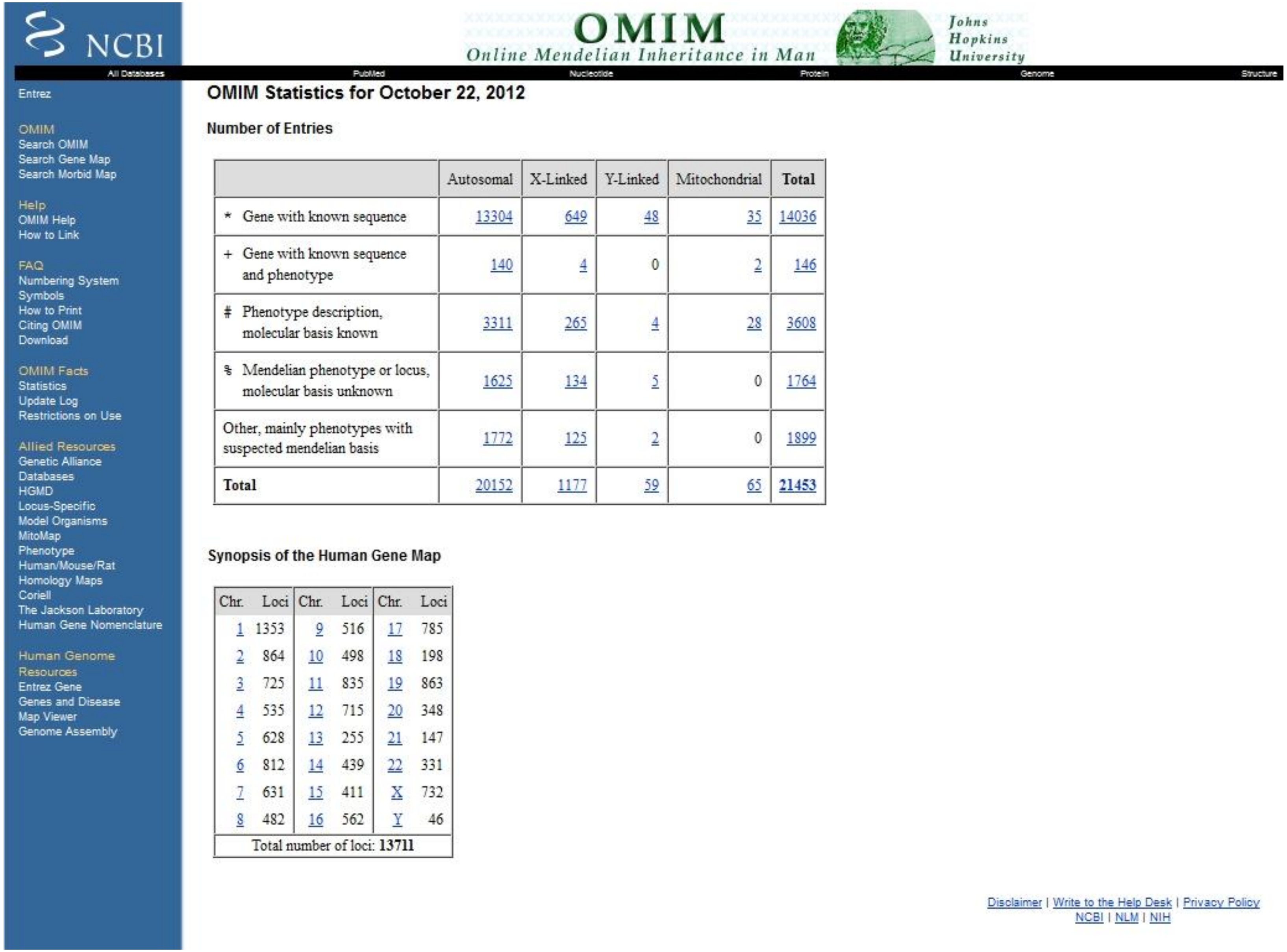
RefSeq Statistics

The NCBI FTP site for RefSeq includes statistics for the current release and past releases.

Taxonomy Statistics

The NCBI Taxonomy home page includes a link to taxonomy statistics. By default, the cumulative, current statistics are shown for the number of higher taxa, genera, species, and lower taxa represented in NCBI's taxonomy database. The number of taxa that were added in any particular year can be viewed by following the link for the year of interest.

As noted in the Taxonomy database summary description in the Resource Guide, the NCBI Taxonomy Database contains the names and lineages of living and extinct organisms that are represented in the genetic databases with at least one nucleotide or protein sequence. New organisms are added to the database as sequence data are deposited for them. The purpose of the taxonomy project at NCBI is to build a consistent phylogenetic taxonomy for the sequence databases.



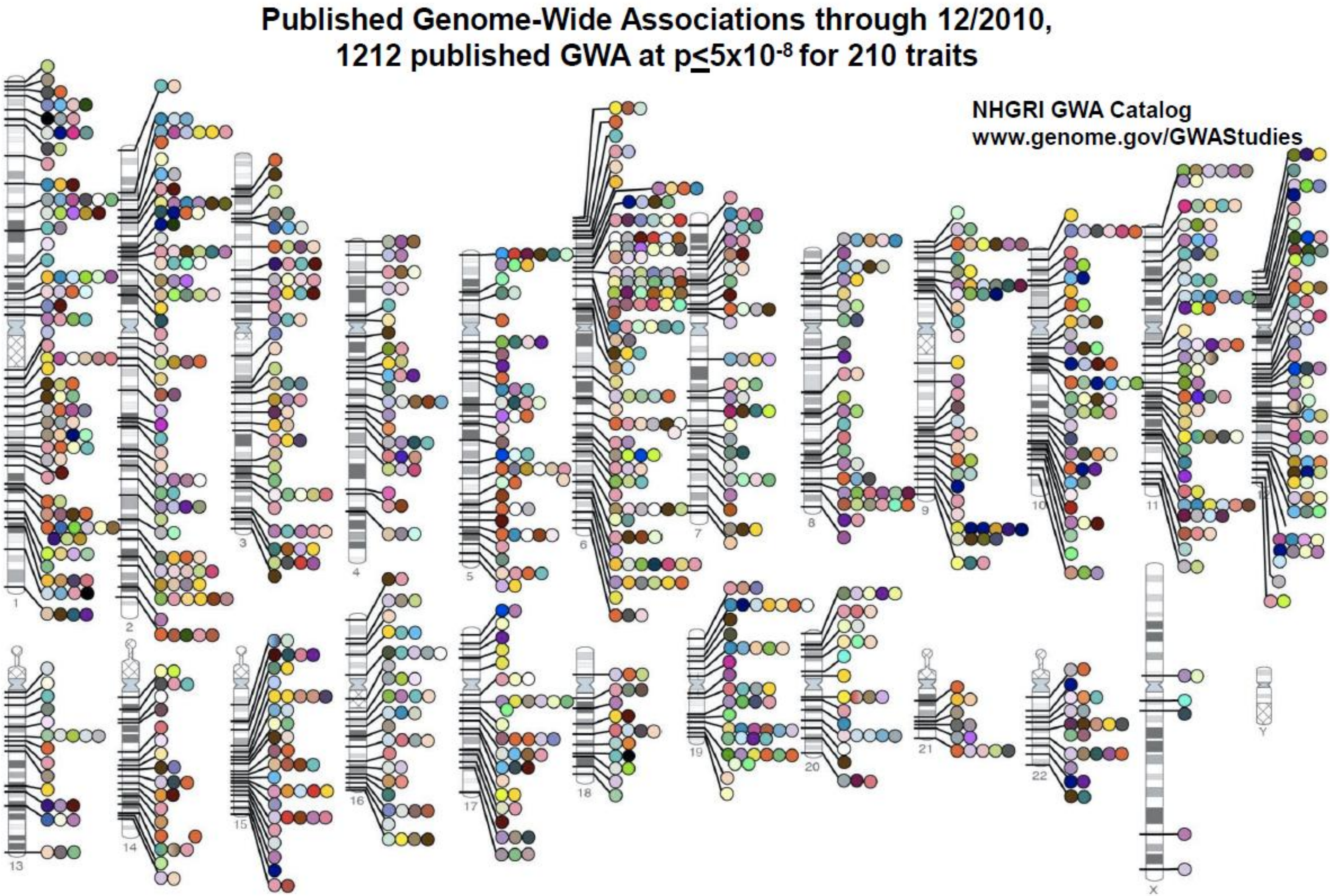
OMIM Statistics for October 22, 2012

Number of Entries

	Autosomal	X-Linked	Y-Linked	Mitochondrial	**Total**
* Gene with known sequence	13304	649	48	35	14036
+ Gene with known sequence and phenotype	140	4	0	2	146
# Phenotype description, molecular basis known	3311	265	4	28	3608
% Mendelian phenotype or locus, molecular basis unknown	1625	134	5	0	1764
Other, mainly phenotypes with suspected mendelian basis	1772	125	2	0	1899
Total	20152	1177	59	65	**21453**

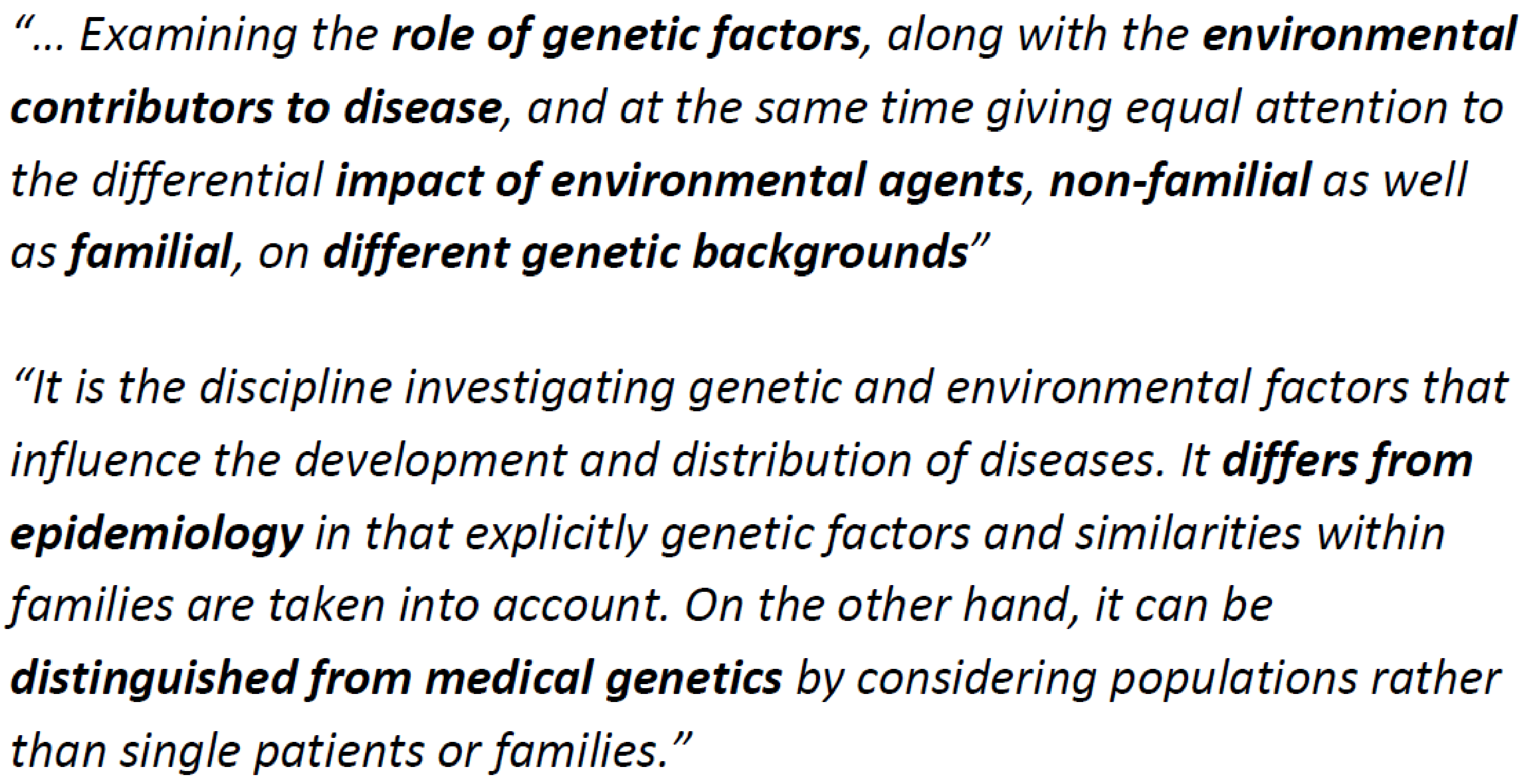
Synopsis of the Human Gene Map

Chr.	Loci	Chr.	Loci	Chr.	Loci
1	1353	9	516	17	785
2	864	10	498	18	198
3	725	11	835	19	863
4	535	12	715	20	348
5	628	13	255	21	147
6	812	14	439	22	331
7	631	15	411	X	732
8	482	16	562	Y	46
Total number of loci: **13711**					


Published Genome-Wide Associations through 12/2010,
1212 published GWA at p≤5x10-8 for 210 traits
NHGRI GWA Catalog
www.genome.gov/GWAStudies
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

1.b Genetic terminology

What is genetic epidemiology?



*"... Examining the **role of genetic factors**, along with the **environmental contributors to disease**, and at the same time giving equal attention to the differential **impact of environmental agents**, **non-familial** as well as **familial**, on **different genetic backgrounds**"*

*"It is the discipline investigating genetic and environmental factors that influence the development and distribution of diseases. It **differs from epidemiology** in that explicitly genetic factors and similarities within families are taken into account. On the other hand, it can be **distinguished from medical genetics** by considering populations rather than single patients or families."*

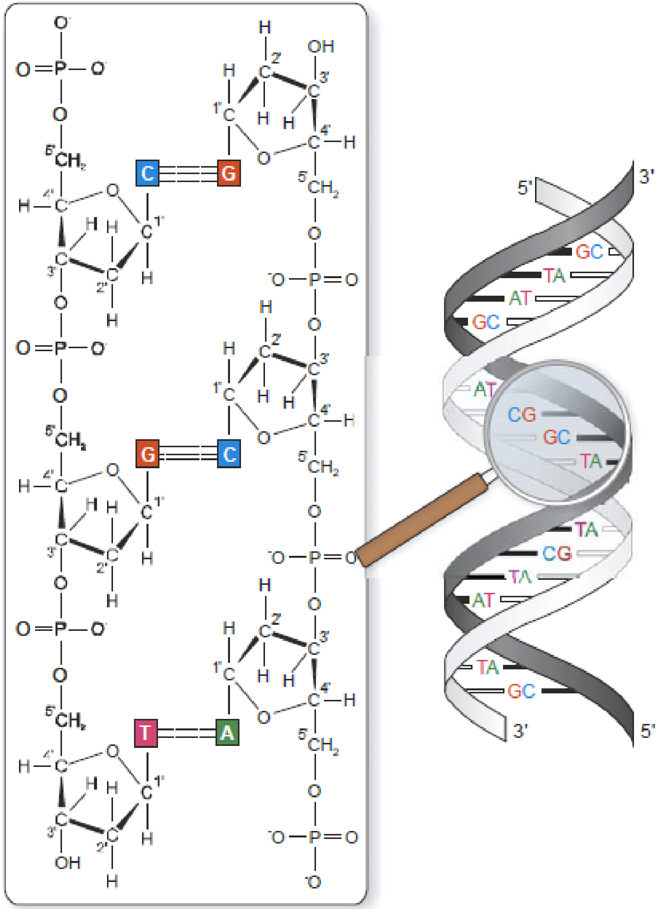
(Ziegler and Van Steen, Brazil 2010)

Where is the genetic information located?

- Cell has nucleus
- Nucleus carries genetic information in chromosomes
- Chromsomes composed of desoxyribonucleic acid (DNA) and proteins
- DNA large molecule consisting in two strands
- Each strand has backbone of sugar and phosphate residues
- Sequence of bases attached to backbone
- Bases: adenine (A), guanine (G), cytosine (C), thymine (T)
- Strands connected through hydrogen bonds
  - A with T (2 hydrogen bonds)
  - C with G (3 hydrogen bonds)

(Ziegler and Van Steen, Brazil 2010)

Where is the genetic information located?



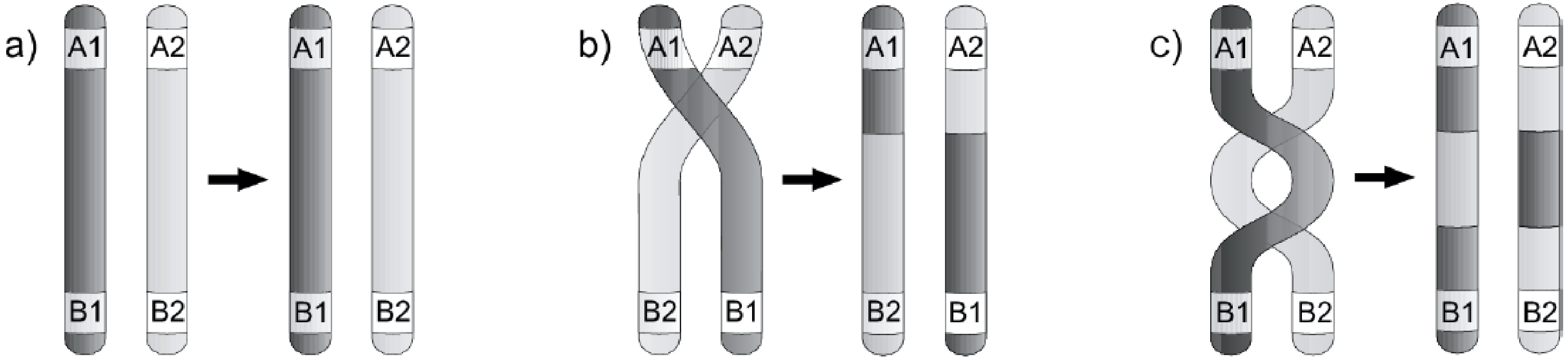
(Ziegler and Van Steen, Brazil 2010)

Where is the genetic information located?

- Chromosomes are
  - Linear arrangements of DNA
  - 22 autosomal pairs in humans
  - 2 sex chromosomes (X and Y)
- Pair of chromosomes called homologs
- Meiosis: special type of cell division
- Crossover: chromosomal segment exchange between homologs during meiosis
- Average # crossovers: 55 × in males, 1.5 × higher in females
- Result of crossover: recombination of non-parental chromosomes in two of the meiotic products

(Ziegler and Van Steen, Brazil 2010)

What is recombination?



- Relevant measure: recombination fraction (probability of odd number of crossovers) between two chromosomal positions
- Strong correlation between recombination fraction and distance in base pairs

(Ziegler and Van Steen, Brazil 2010)

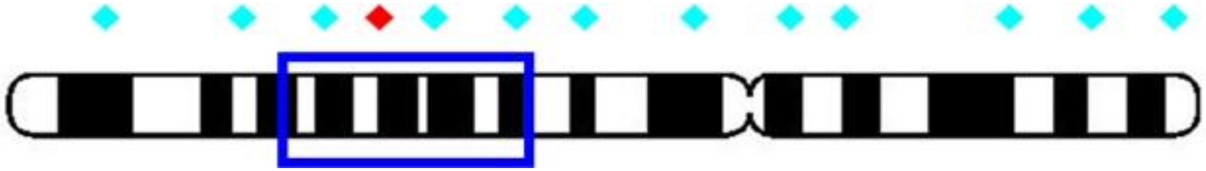
How much do individuals differ with respect to genetic information?

- Allele: one of several alternative forms of DNA sequence at specific chromosomal location (locus)
- Genetic marker: polymorphic DNA sequence at single locus
- Polymorphism: existence of ≥ 2 alleles at single locus
- Homozygosity (homozygous): both alleles identical at locus
- Heterozygosity (heterozygous): different alleles at locus
- Mutation:
  - ○ Changes allele at specific chromosomal position
  - ○ Frequency $\approx 10^{-4}$ to $10^{-6} \Rightarrow$ Individuals differ with freq. of 1/1000 bases

(Ziegler and Van Steen, Brazil 2010)

How much do individuals differ with respect to genetic information?

- **Genotype**: The two alleles inherited at a specific locus. If the alleles are the same, the genotype is homozygous, if different, heterozygous. In genetic association studies, genotypes can be used for analysis as well as alleles or haplotypes.
- **Haplotype**: Linear arrangements of alleles on the same chromosome that have been inherited as a unit. A person has two haplotypes for any such series of loci, one inherited maternally and the other paternally. A haplotype may be characterized by a single allele unless a discrete chromosomal segment flanked by two alleles is meant.



http://www.dorak.info/epi/glosge.html

Are haplotypes always better in association studies for "disease"?

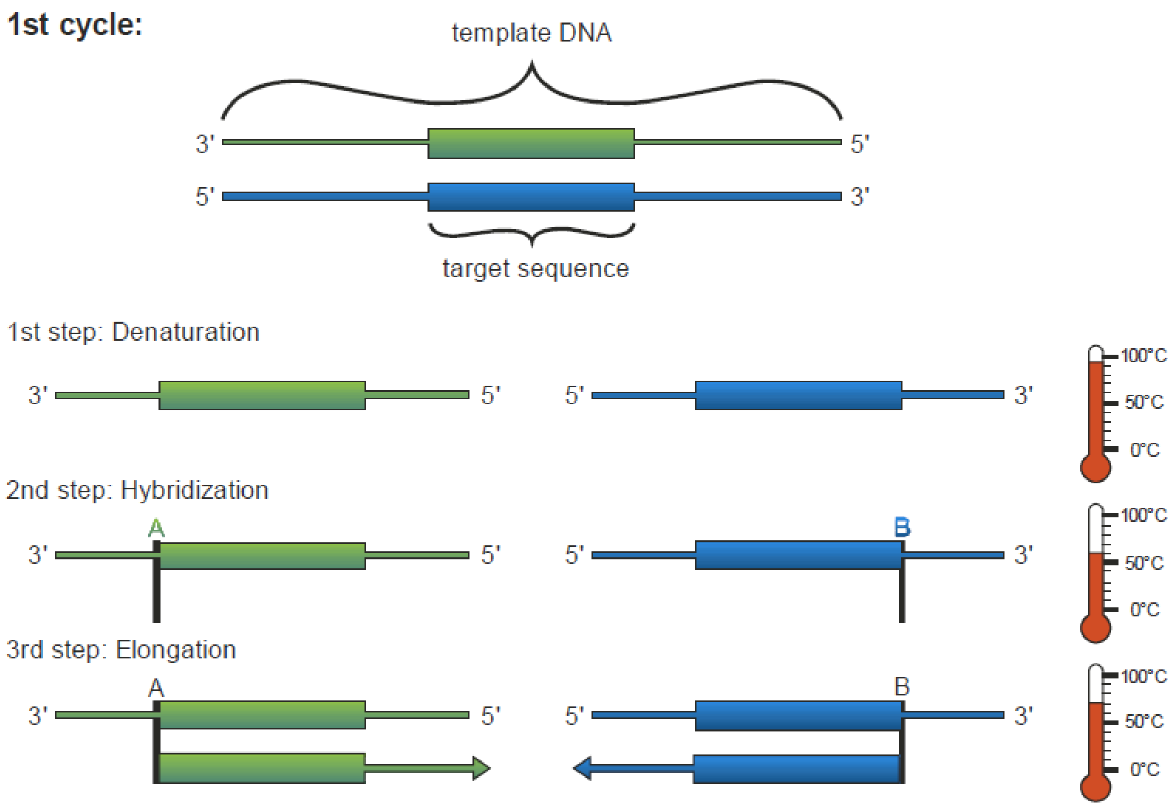
- Analyses based on phased haplotype data rather than unphased genotypes may be *quite powerful*…

M1	1	1	2	2
DSL	D	d	d	d
M2	1	2	1	2

Test 1 vs. 2 for M1:	D + d vs. d
Test 1 vs. 2 for M2:	D + d vs. d
Test haplotype H1 vs. all others:	D vs. d

- If the **Disease Susceptibility Locus** (DSL) is located at a marker, haplotype testing can be *less powerful*

How can individual differences be detected?



(Ziegler and Van Steen, Brazil 2010)

What are microsatellite markers?

- Synonymous: short tandem repeat, STR
- Number of repeats varies between individuals
    - Mononucleotide, dinucleotide, trinucleotide, tetranucleotide, non-integer STRs
- Determine allele length (e.g., 133, 136, 139, 142, …)
- Occurrence in non-coding regions
- High mutation frequency $\approx 10^{-2} - 10^{-4}$ events per locus per generation
- Not easy to score automatically
- Frequent but not dense enough for some applications

(Ziegler and Van Steen, Brazil 2010)

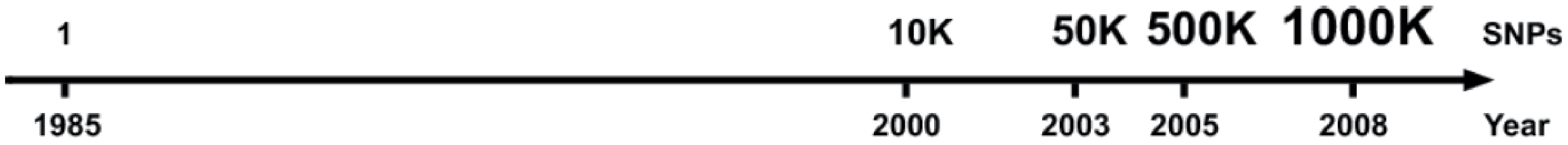
What are single nucleotide polymorphisms?

- Variations in single base, i.e., one base substituted by another base
- In theory: four different nucleotides possible at base
- In practice: generally only two different nucleotides observed
- Definition strict and loose:
  - Strict: minor allele frequency ≥ 1%
  - Loose: ≥ 2 nucleotides observed in two individuals at position
- Nomenclature:
  - ss-number (submitted SNP number)
  - rs-number: searchable in dbSNP, mapped to external resources, unique
  - rs-numbers do not provide information about possible function of SNP
  - Alternative: nomenclature of Human Genome Variation Society

(Ziegler and Van Steen, Brazil 2010)

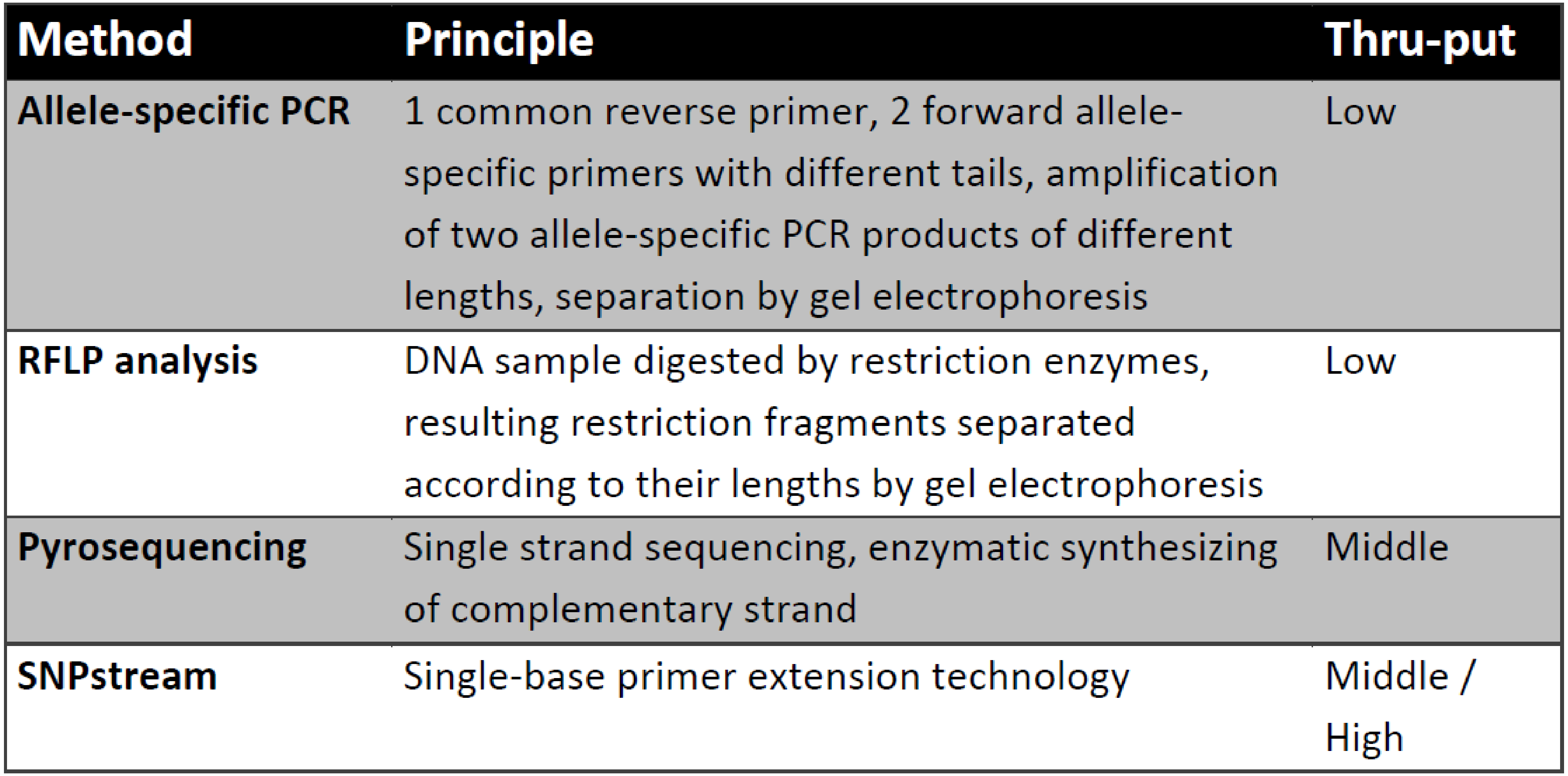
Why are SNPs preferred over STRs?

- SNPs very frequent → dense marker map
- Some SNPs functionally relevant → candidate variations for disease
- SNPs more stable, i.e., lower mutation rate
- Genotyping in highly automated fashion



(Ziegler and Van Steen, Brazil 2010)

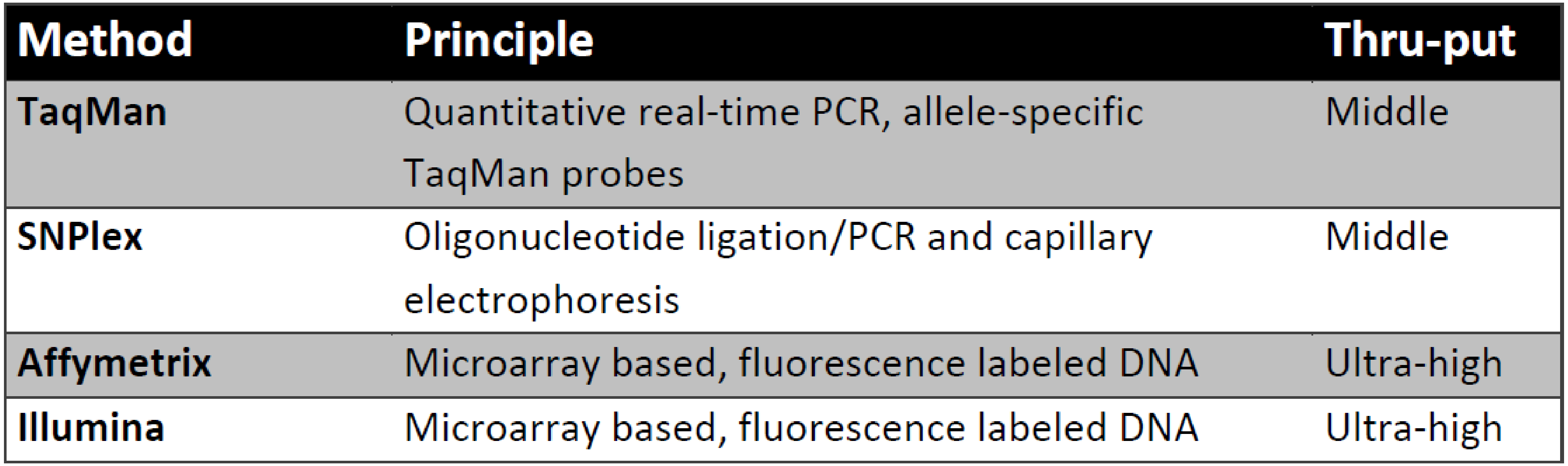
Which genotyping methods are currently being used?



Method	Principle	Thru-put
Allele-specific PCR	1 common reverse primer, 2 forward allele-specific primers with different tails, amplification of two allele-specific PCR products of different lengths, separation by gel electrophoresis	Low
RFLP analysis	DNA sample digested by restriction enzymes, resulting restriction fragments separated according to their lengths by gel electrophoresis	Low
Pyrosequencing	Single strand sequencing, enzymatic synthesizing of complementary strand	Middle
SNPstream	Single-base primer extension technology	Middle / High

(Ziegler and Van Steen, Brazil 2010)

Which genotyping methods are currently being used?



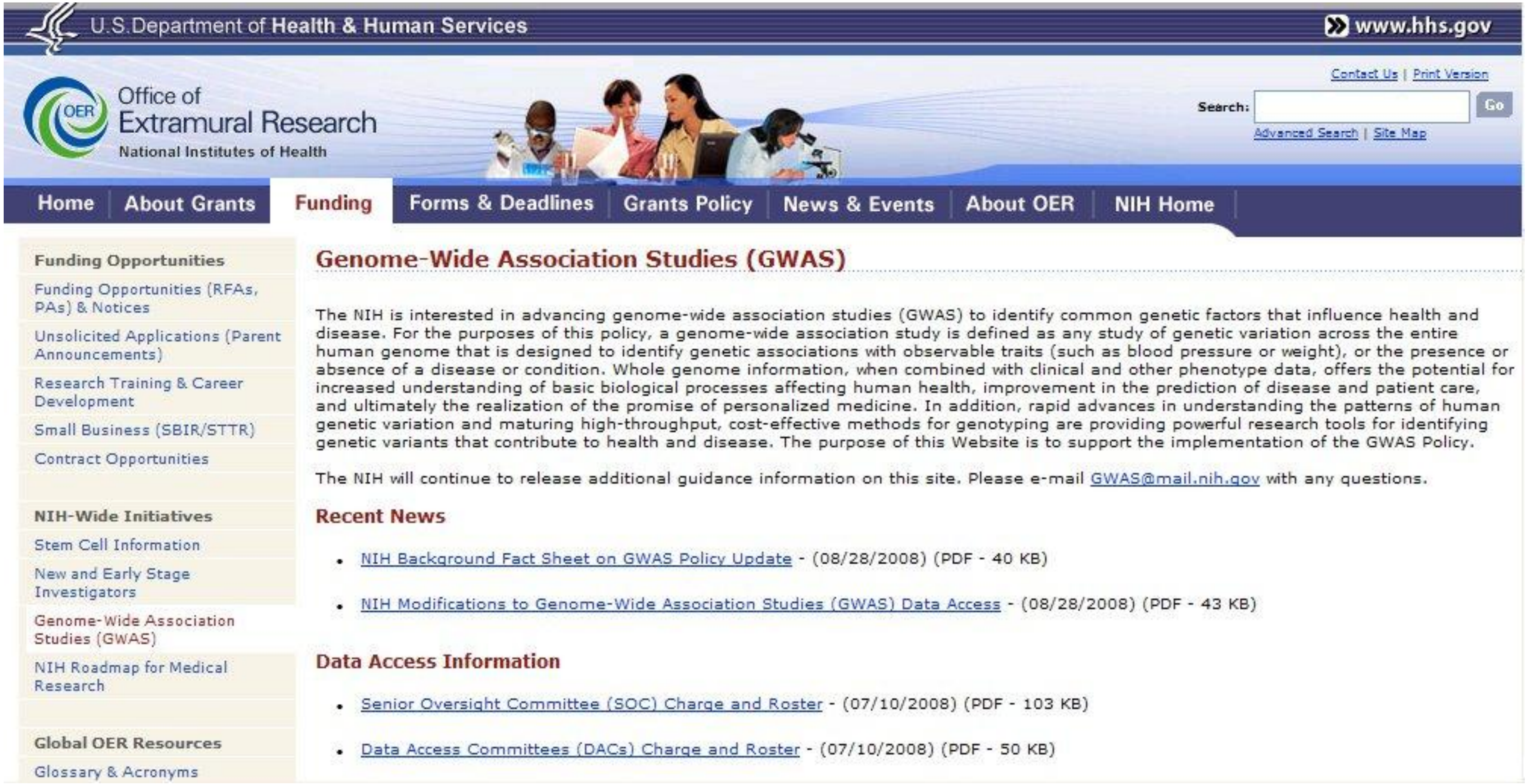
Method	Principle	Thru-put
TaqMan	Quantitative real-time PCR, allele-specific TaqMan probes	Middle
SNPlex	Oligonucleotide ligation/PCR and capillary electrophoresis	Middle
Affymetrix	Microarray based, fluorescence labeled DNA	Ultra-high
Illumina	Microarray based, fluorescence labeled DNA	Ultra-high

(Ziegler and Van Steen, Brazil 2010)

1.c Genetic association studies

What is a genome-wide association study?

- It refers to a method / methodology for interrogating all 10 million variable points across the human genome.
- Since variation is inherited in groups, or blocks, not all 10 million points have to be tested.
- Blocks are shorter though (so need for testing more points) the less closely people are related.


U.S. Department of Health & Human Services
www.hhs.gov
Office of Extramural Research
National Institutes of Health
Contact Us | Print Version
Search:
Go
Advanced Search | Site Map
Home
About Grants
Funding
Forms & Deadlines
Grants Policy
News & Events
About OER
NIH Home
Funding Opportunities
Funding Opportunities (RFAs, PAs) & Notices
Unsolicited Applications (Parent Announcements)
Research Training & Career Development
Small Business (SBIR/STTR)
Contract Opportunities
NIH-Wide Initiatives
Stem Cell Information
New and Early Stage Investigators
Genome-Wide Association Studies (GWAS)
NIH Roadmap for Medical Research
Global OER Resources
Glossary & Acronyms

Genome-Wide Association Studies (GWAS)

The NIH is interested in advancing genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease. For the purposes of this policy, a genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition. Whole genome information, when combined with clinical and other phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of personalized medicine. In addition, rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease. The purpose of this Website is to support the implementation of the GWAS Policy.

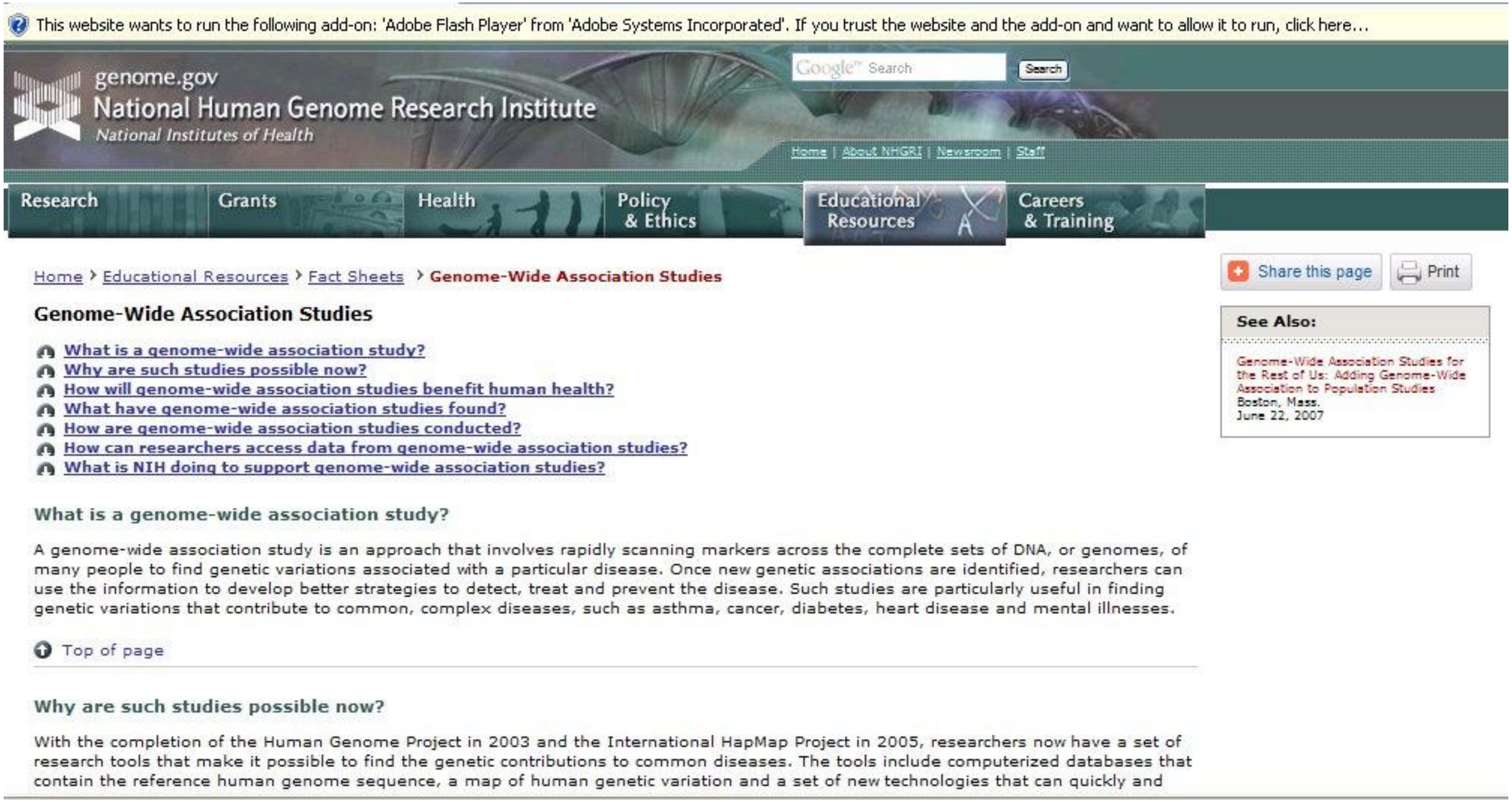
The NIH will continue to release additional guidance information on this site. Please e-mail GWAS@mail.nih.gov with any questions.

Recent News

- NIH Background Fact Sheet on GWAS Policy Update - (08/28/2008) (PDF - 40 KB)
- NIH Modifications to Genome-Wide Association Studies (GWAS) Data Access - (08/28/2008) (PDF - 43 KB)

Data Access Information

- Senior Oversight Committee (SOC) Charge and Roster - (07/10/2008) (PDF - 103 KB)
- Data Access Committees (DACs) Charge and Roster - (07/10/2008) (PDF - 50 KB)



This website wants to run the following add-on: 'Adobe Flash Player' from 'Adobe Systems Incorporated'. If you trust the website and the add-on and want to allow it to run, click here...
genome.gov
National Human Genome Research Institute
National Institutes of Health
Google™ Search
Search
Home | About NHGRI | Newsroom | Staff
Research
Grants
Health
Policy & Ethics
Educational Resources
Careers & Training
Share this page
Print

Home > Educational Resources > Fact Sheets > **Genome-Wide Association Studies**

Genome-Wide Association Studies

- What is a genome-wide association study?
- Why are such studies possible now?
- How will genome-wide association studies benefit human health?
- What have genome-wide association studies found?
- How are genome-wide association studies conducted?
- How can researchers access data from genome-wide association studies?
- What is NIH doing to support genome-wide association studies?

See Also:

Genome-Wide Association Studies for the Rest of Us: Adding Genome-Wide Association to Population Studies
Boston, Mass.
June 22, 2007

What is a genome-wide association study?

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

Top of page

Why are such studies possible now?

With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. The tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation and a set of new technologies that can quickly and

What is a genome-wide association study?

- Hence, a genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease.
- Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.

(http://www.genome.gov/pfv.cfm?pageID=20019523)

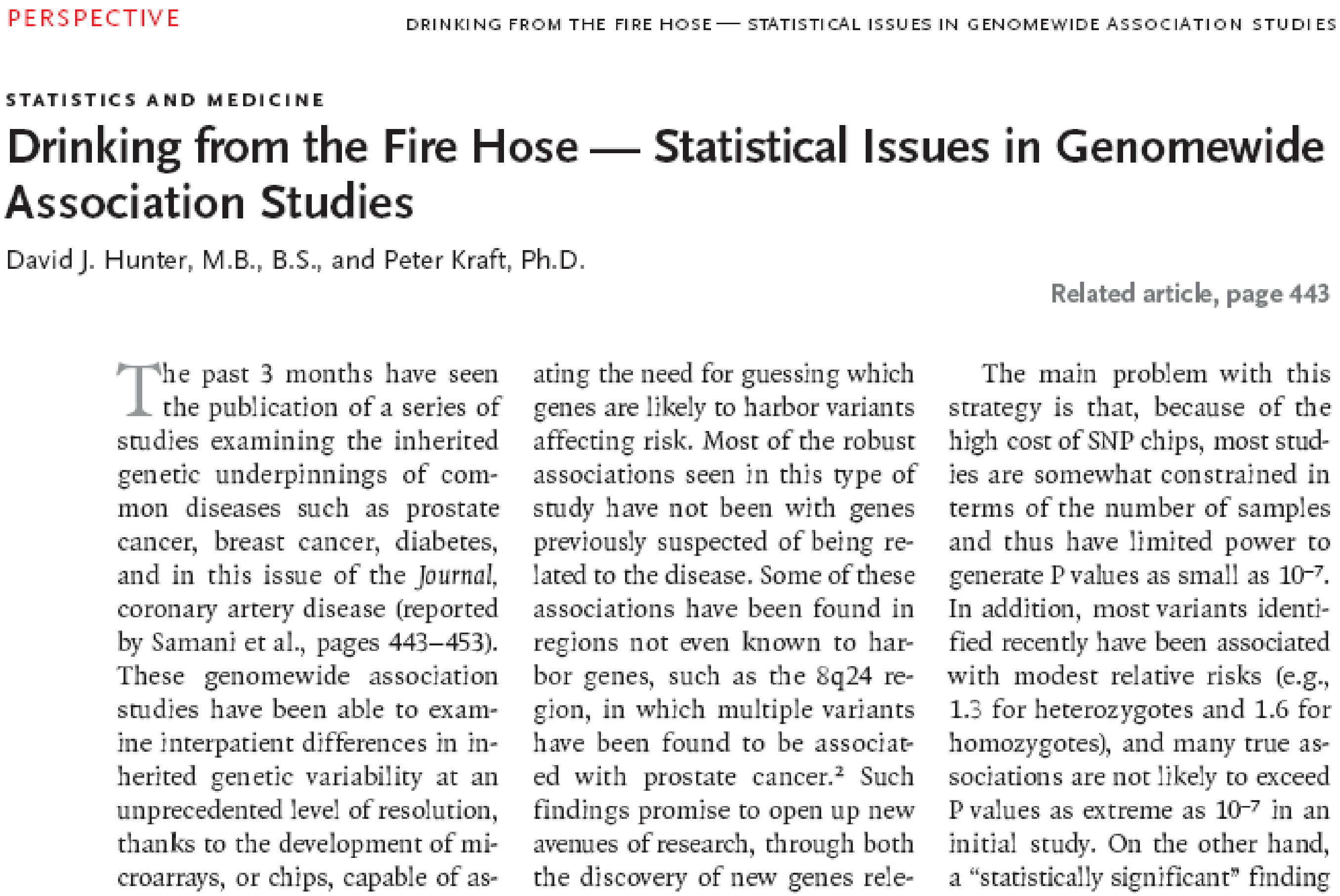
- The impact on medical care from genome-wide association studies could potentially be substantial. Such research is laying the groundwork for the era of personalized medicine, in which the current one size-fits-all approach to medical care will give way to more customized strategies.

What do we need to carry out a genome-wide association study?

- The tools include
  - computerized databases that contain the reference human genome sequence,
  - a map of human genetic variation and
  - a set of new technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease.

(http://www.genome.gov/pfv.cfm?pageID=20019523)

What do we need to carry out a genome-wide association study?



PERSPECTIVE

DRINKING FROM THE FIRE HOSE — STATISTICAL ISSUES IN GENOMEWIDE ASSOCIATION STUDIES

STATISTICS AND MEDICINE

Drinking from the Fire Hose — Statistical Issues in Genomewide Association Studies

David J. Hunter, M.B., B.S., and Peter Kraft, Ph.D.

Related article, page 443

The past 3 months have seen the publication of a series of studies examining the inherited genetic underpinnings of common diseases such as prostate cancer, breast cancer, diabetes, and in this issue of the *Journal*, coronary artery disease (reported by Samani et al., pages 443–453). These genomewide association studies have been able to examine interpatient differences in inherited genetic variability at an unprecedented level of resolution, thanks to the development of microarrays, or chips, capable of associating the need for guessing which genes are likely to harbor variants affecting risk. Most of the robust associations seen in this type of study have not been with genes previously suspected of being related to the disease. Some of these associations have been found in regions not even known to harbor genes, such as the 8q24 region, in which multiple variants have been found to be associated with prostate cancer.[2] Such findings promise to open up new avenues of research, through both the discovery of new genes rele-

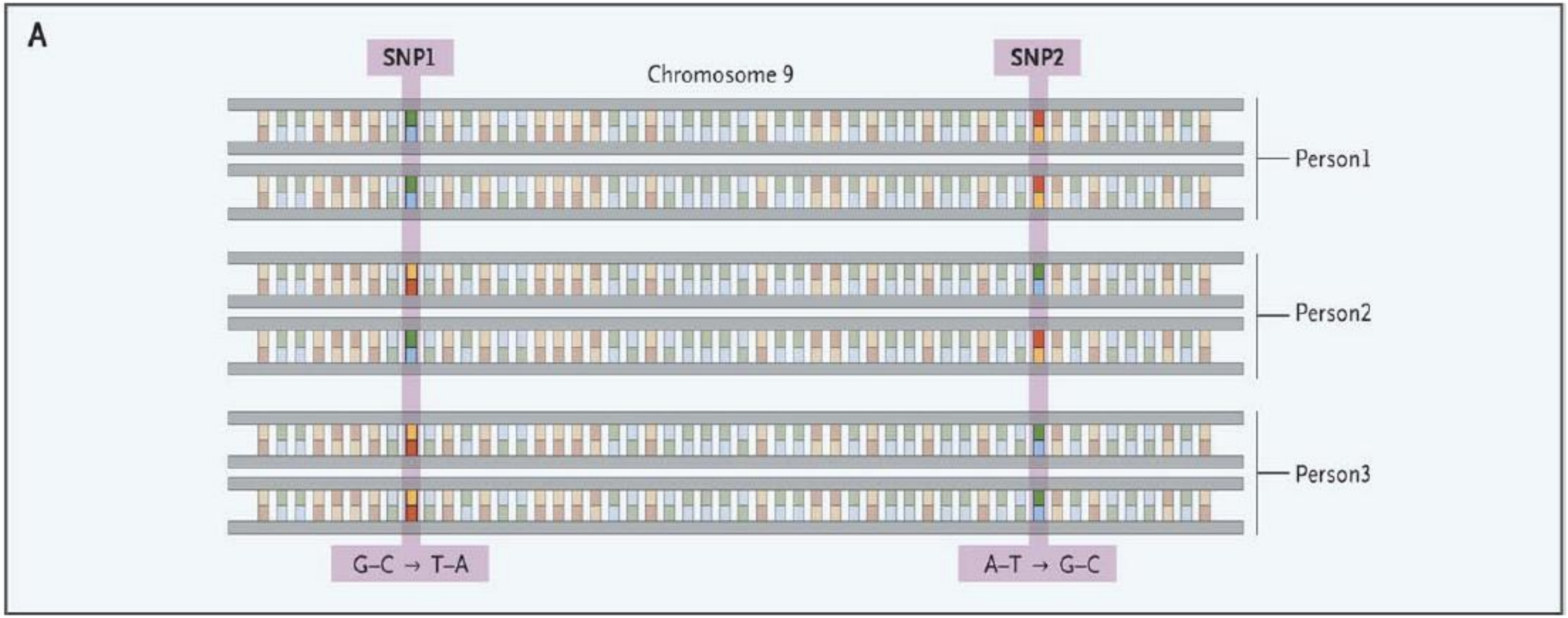
The main problem with this strategy is that, because of the high cost of SNP chips, most studies are somewhat constrained in terms of the number of samples and thus have limited power to generate P values as small as 10^{-7}. In addition, most variants identified recently have been associated with modest relative risks (e.g., 1.3 for heterozygotes and 1.6 for homozygotes), and many true associations are not likely to exceed P values as extreme as 10^{-7} in an initial study. On the other hand, a "statistically significant" finding

What do we need to carry out a genome-wide association study?

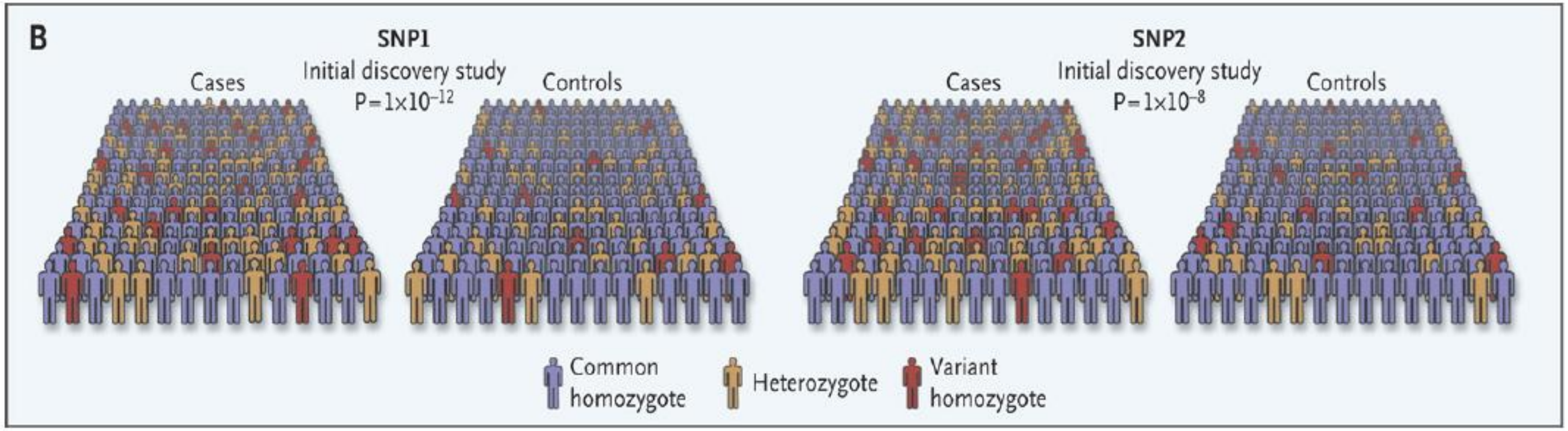
- To distinguish between true and chance effects, there are several routes to be taken:
    - Set **tight standards** for statistical significance
    - Only consider patterns of polymorphisms that could plausibly have been generated by causal genetic variants (**use** understanding of and **insights** into human genetic history or evolutionary processes such as recombination or mutation)
    - Adequately deal with distorting factors, including missing data and genotyping errors (**quality control** measures)

What is the flow of a genome-wide association study?

The genome-wide association study is typically (but not solely!!!) based on a case–control design in which single-nucleotide polymorphisms (SNPs) across the human genome are genotyped ... (Panel A: small fragment)



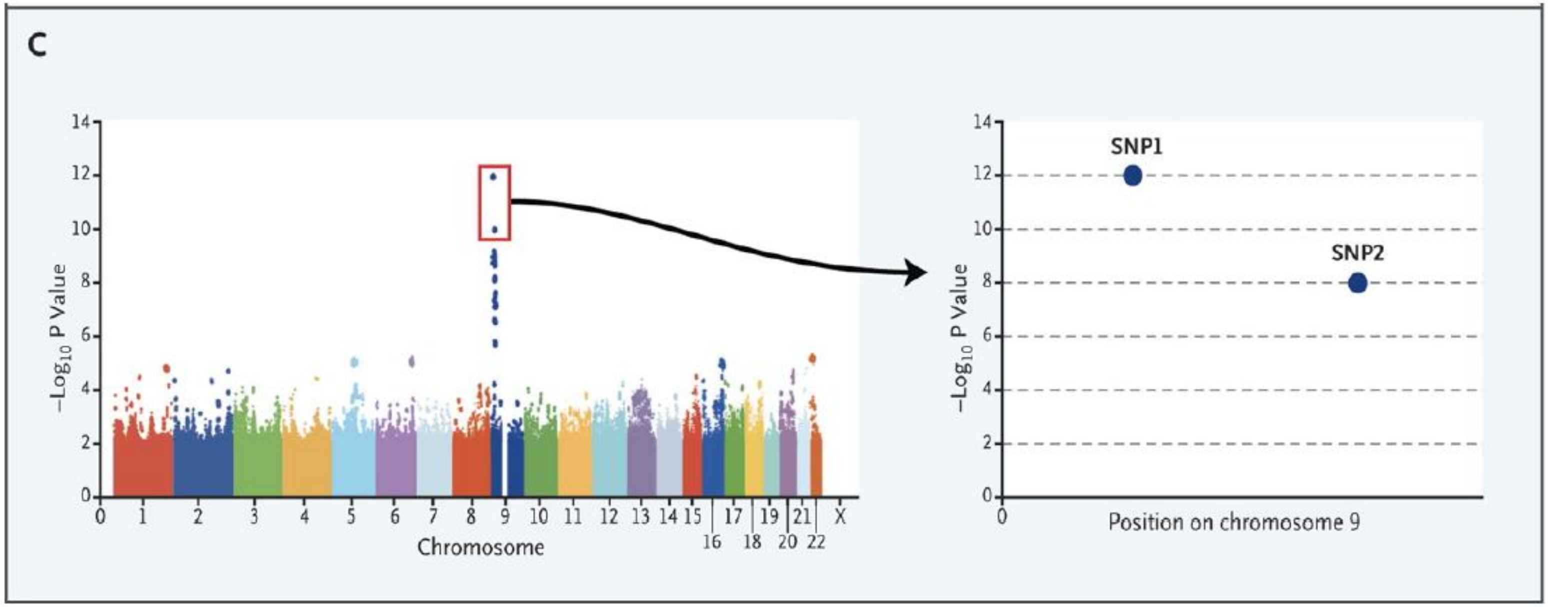
What is the flow of a genome-wide association study?



- Panel B, the strength of association between each SNP and disease is calculated on the basis of the prevalence of each SNP in cases and controls. In this example, SNPs 1 and 2 on chromosome 9 are associated with disease, with P values of 10^{-12} and 10^{-8}, respectively

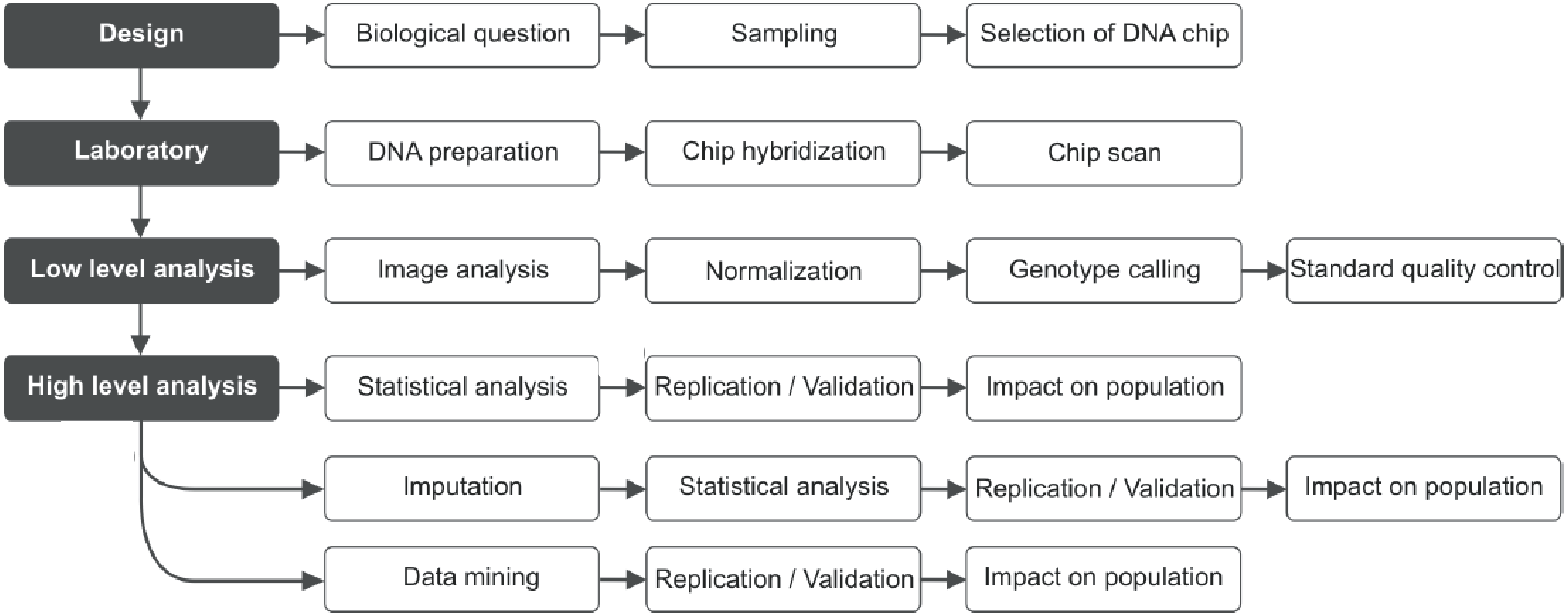
(Manolio 2010)

What is the flow of a genome-wide association study?



- The plot in Panel C shows the P values for all genotyped SNPs that have survived a quality-control screen, with each chromosome shown in
- a different color. The results implicate a locus on chromosome 9, marked by SNPs 1 and 2, which are adjacent to each other (graph at
- right), and other neighboring SNPs. (Manolio 2010)

What is the flow of a genome-wide association study?



Design → Biological question → Sampling → Selection of DNA chip

Laboratory → DNA preparation → Chip hybridization → Chip scan

Low level analysis → Image analysis → Normalization → Genotype calling → Standard quality control

High level analysis → Statistical analysis → Replication / Validation → Impact on population

→ Imputation → Statistical analysis → Replication / Validation → Impact on population

→ Data mining → Replication / Validation → Impact on population

(Ziegler 2009)

2 Study Designs

What are the components of a study design for GWA studies?

- The design of a genetic association study may refer to
  - study scale:
    - Genome-wide
    - Genomic
  - marker design:
    - Which markers are most informative? Microsatellites? SNPs? CNVs?
    - Which platform is the most promising?
  - subject design

Does scale matter?

candidate gene approach

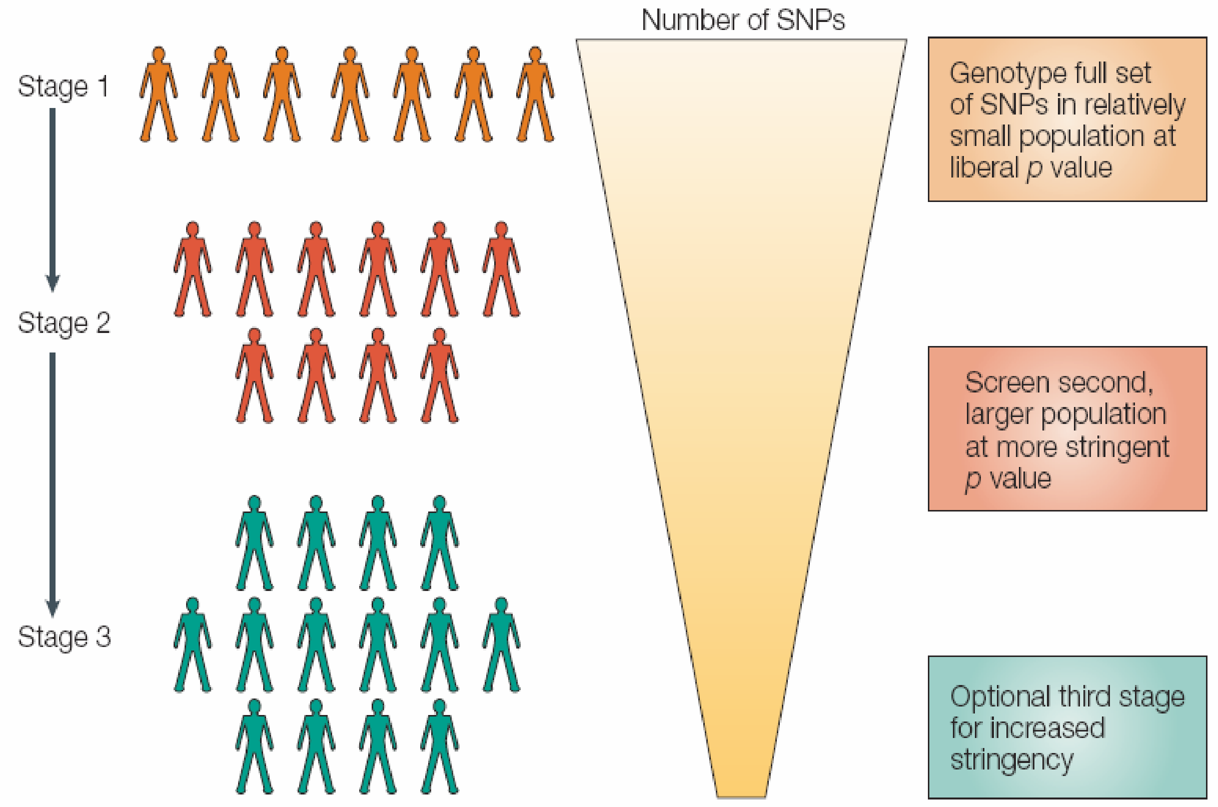
vs

genome-wide screening approach

Can’t see the forest for the trees

Can’t see the trees for the forest

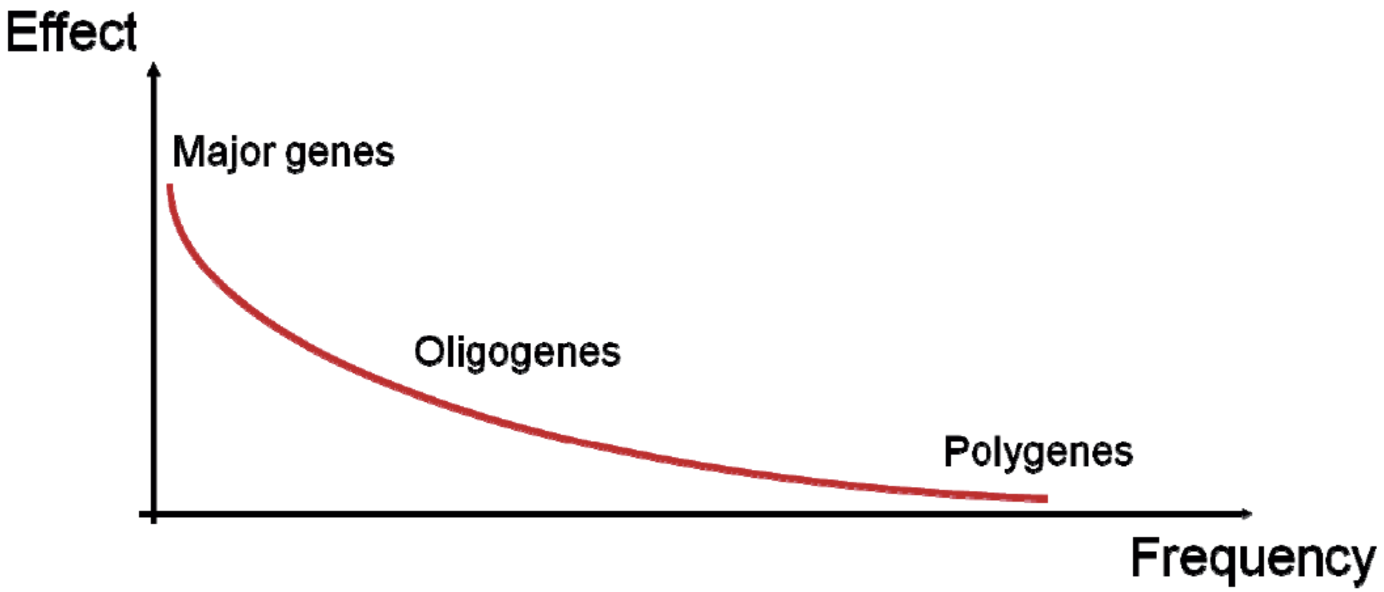
Does scale matter?



Number of SNPs
Stage 1
Stage 2
Stage 3
Genotype full set of SNPs in relatively small population at liberal *p* value
Screen second, larger population at more stringent *p* value
Optional third stage for increased stringency

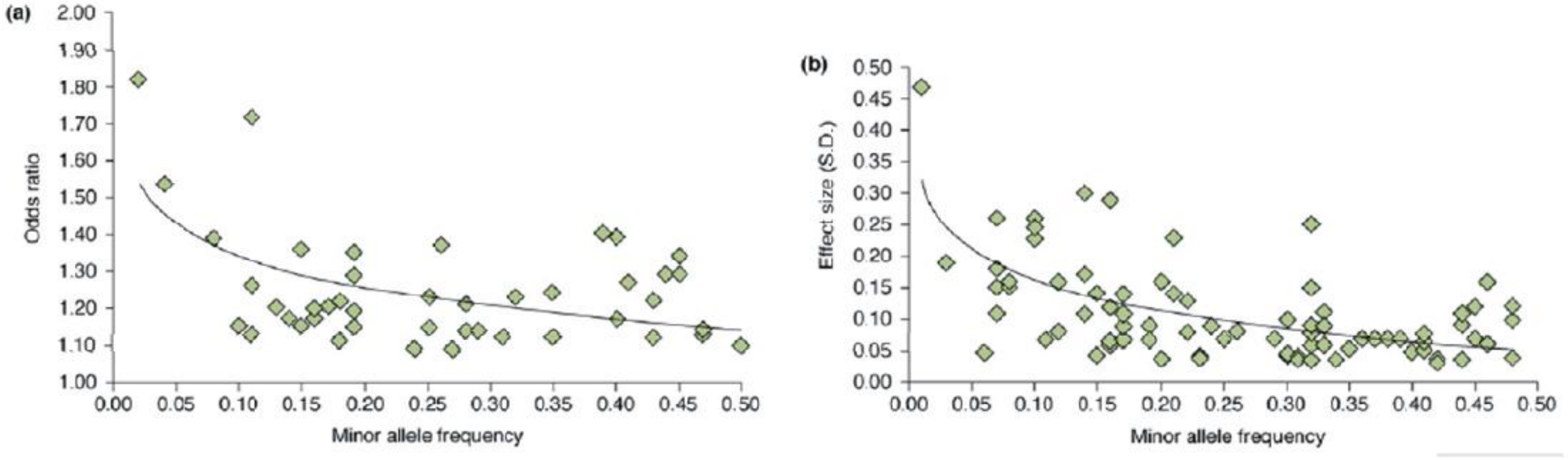
Which genetic markers to select?

The **Common Disease/Common Variant** hypothesis (CDCV)



- Continuous distribution of genetic variants, shaped by mutation and selection

(Ziegler and Van Steen, 2010)



Dichotomous Traits

Quantitative Traits

Arking & Chakravarti 2009 Trends Genet

Observations:

- The higher the MAF (minor allele frequency), the higher the detection rate?
- The higher the MAF, the lower the penetrance?

- There are three types of genetic diseases: Mendelian, oligogenic, polygenic

- **Monogenic diseases** are those in which defects in a single gene produce disease. Often these disease are severe and appear early in life, e.g., cystic fibrosis. For the population as a whole, they are relatively rare. In a sense, these are pure genetic diseases: They do not require any environmental factors to elicit them. Although nutrition is not involved in the causation of monogenic diseases, these diseases can have implications for nutrition. They reveal the effects of particular proteins or enzymes that also are influenced by nutritional factors

(http://www.utsouthwestern.edu)

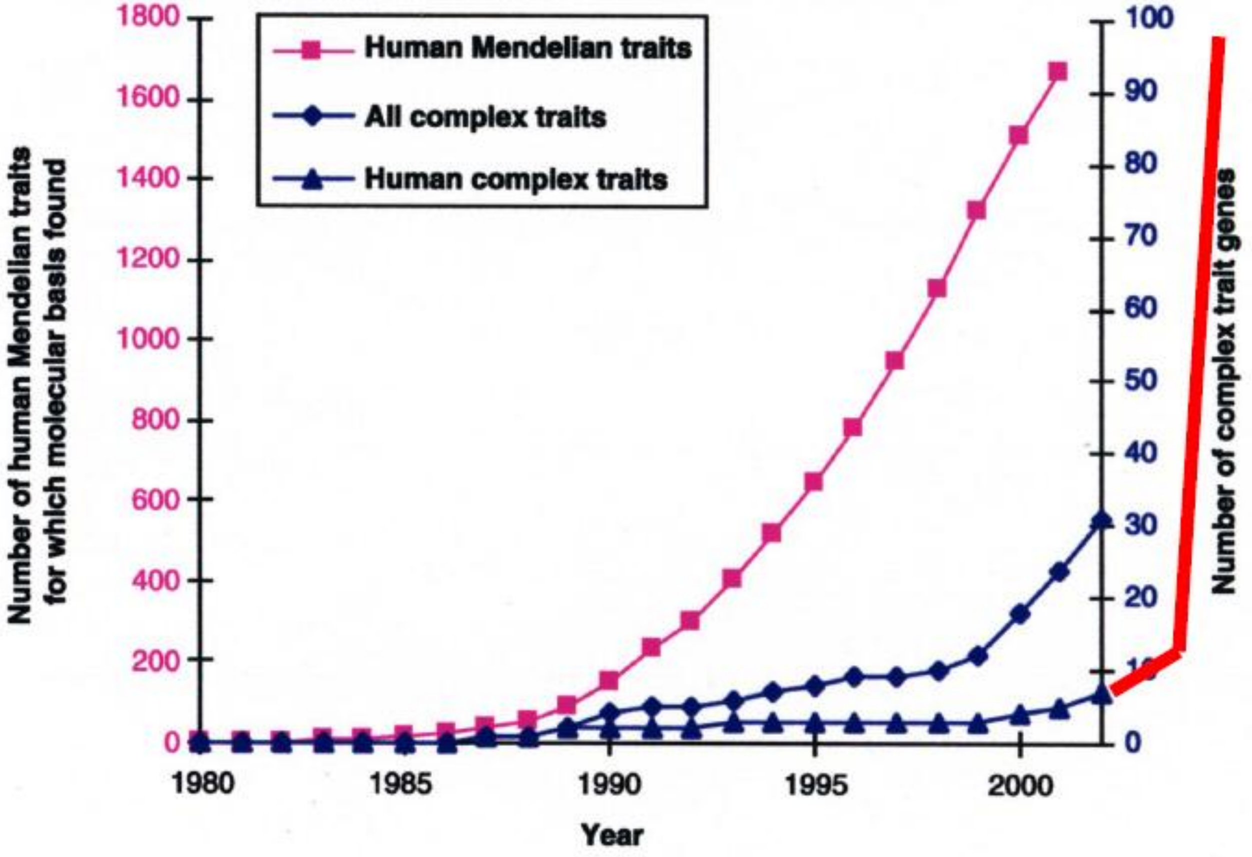
- **Oligogenic diseases** are conditions produced by the combination of two, three, or four defective genes. Often a defect in one gene is not enough to elicit a full-blown disease; but when it occurs in the presence of other moderate defects, a disease becomes clinically manifest. It is the expectation of human geneticists that many chronic diseases can be explained by the combination of defects in a few (major) genes.

- A third category of genetic disorder is **polygenic disease**. According to the polygenic hypothesis, many mild defects in genes conspire to produce some chronic diseases. To date the full genetic basis of polygenic diseases has not been worked out; multiple interacting defects are highly complex !!!

(http://www.utsouthwestern.edu)

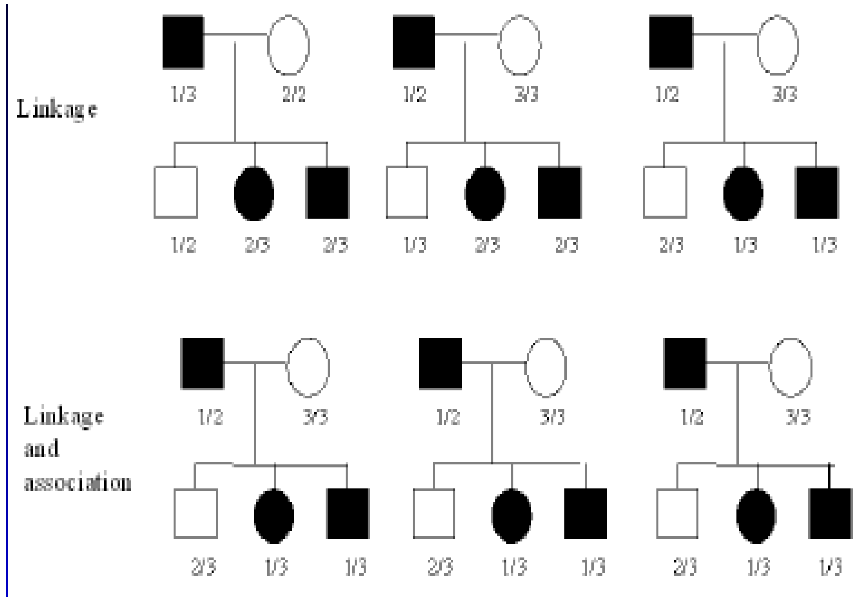
- **Complex diseases** refer to conditions caused by many contributing factors. Such a disease is also called a multifactorial disease.
  - Some disorders, such as sickle cell anemia and cystic fibrosis, are caused by mutations in a single gene.
  - Common medical problems such as heart disease, diabetes, and obesity likely associated with the effects of multiple genes in combination with lifestyle and environmental factors, all of them possibly interacting.

⇒ Challenge for many years to come …



(Glazier et al 2002)

Which genetic markers to select?



(Figure: courtesy of Ed Silverman)

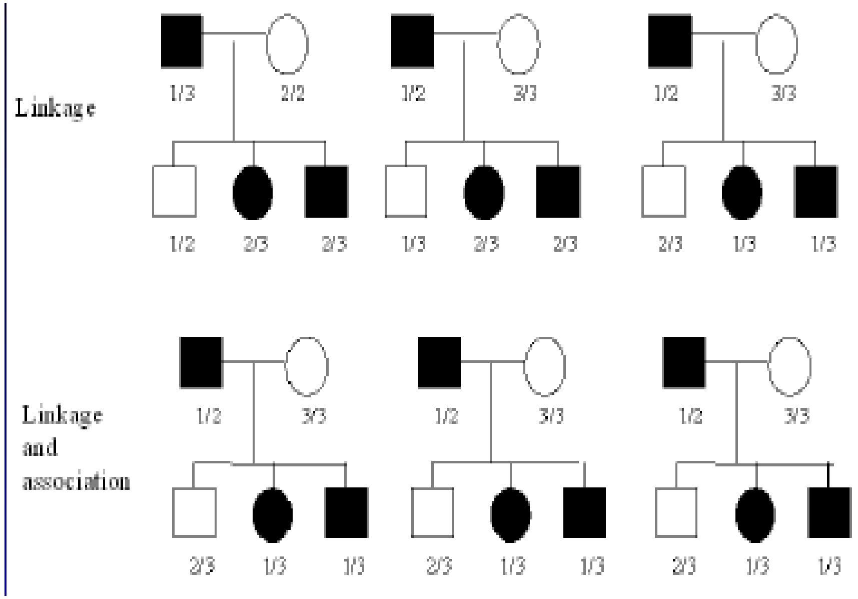
- Linkage exists over a very broad region, entire chromosome can be done using data on only 400-800 DNA markers
- Broad linkage regions imply studies must be followed up with more DNA markers in the region
- Must have family data with more than one affected subject

E.g., microsatellites

Which genetic markers to select?

- Association exists over a narrow region; markers must be close to disease gene
    - The basic concept is linkage disequilibrium (LD) – see later in this chapter
- Initially used for candidate genes or in linked regions
- Can use population-based (unrelated cases) or family-based design

E.g., SNPs



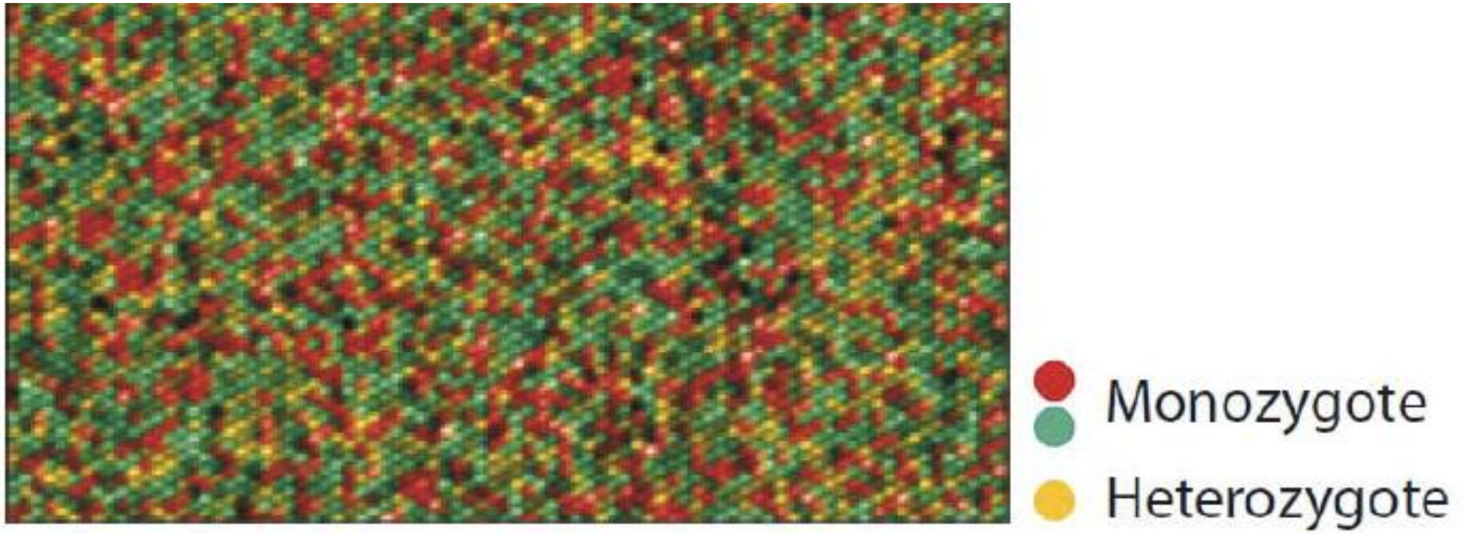
The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

SCIENCE • VOL. 273 • 13 SEPTEMBER 1996

Which DNA SNPs to select?

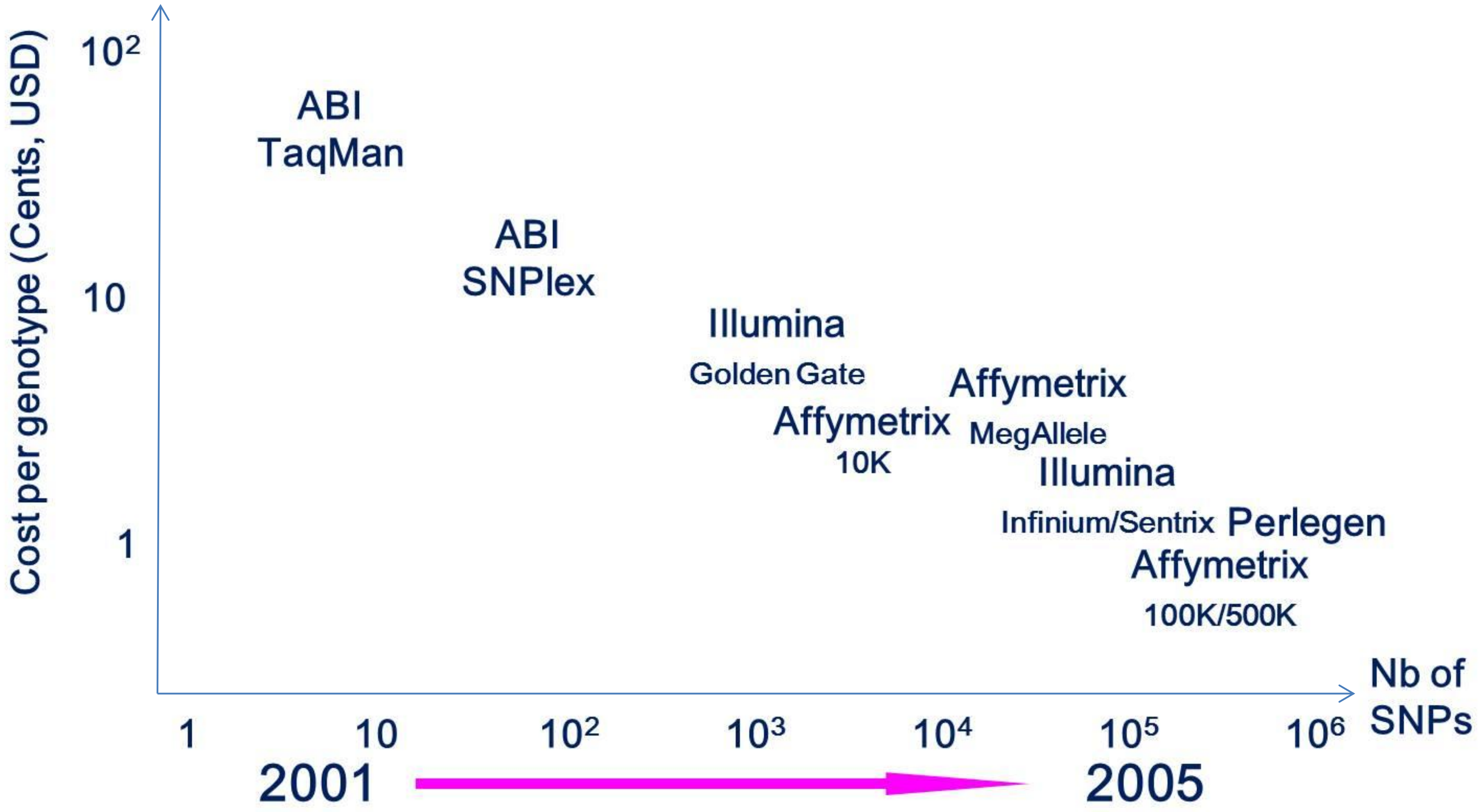
- Costs may play a role, but a balance is needed between costs and chip performance as well as coverage (e.g., exonic regions only?)



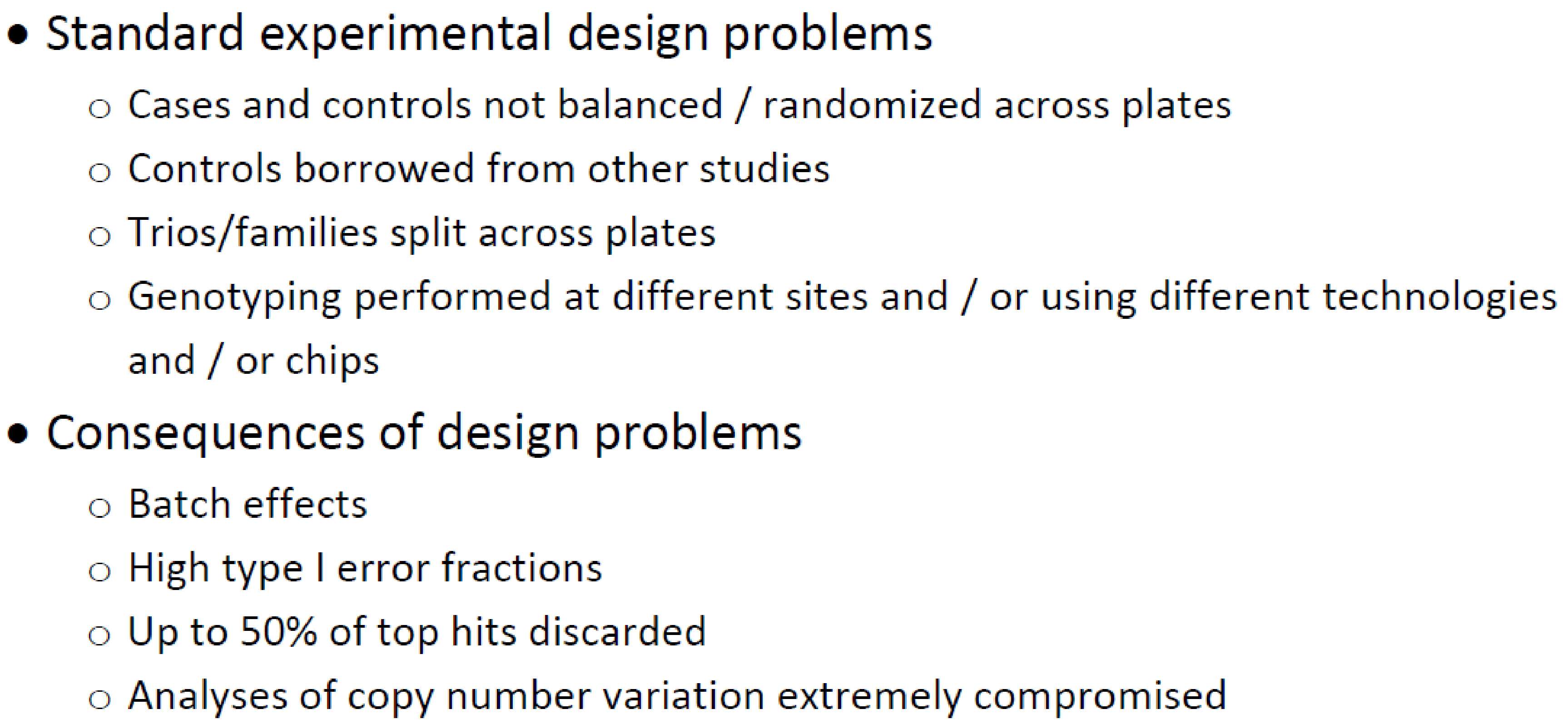
Illumina 610S Quad Beadchip
Ragoussis 2009 Annu Rev Genomics Hum Genet

- Some of the fundamental principles of array technology (see future class)

Which DNA SNPs to select? (adapted from Manolio 2010)



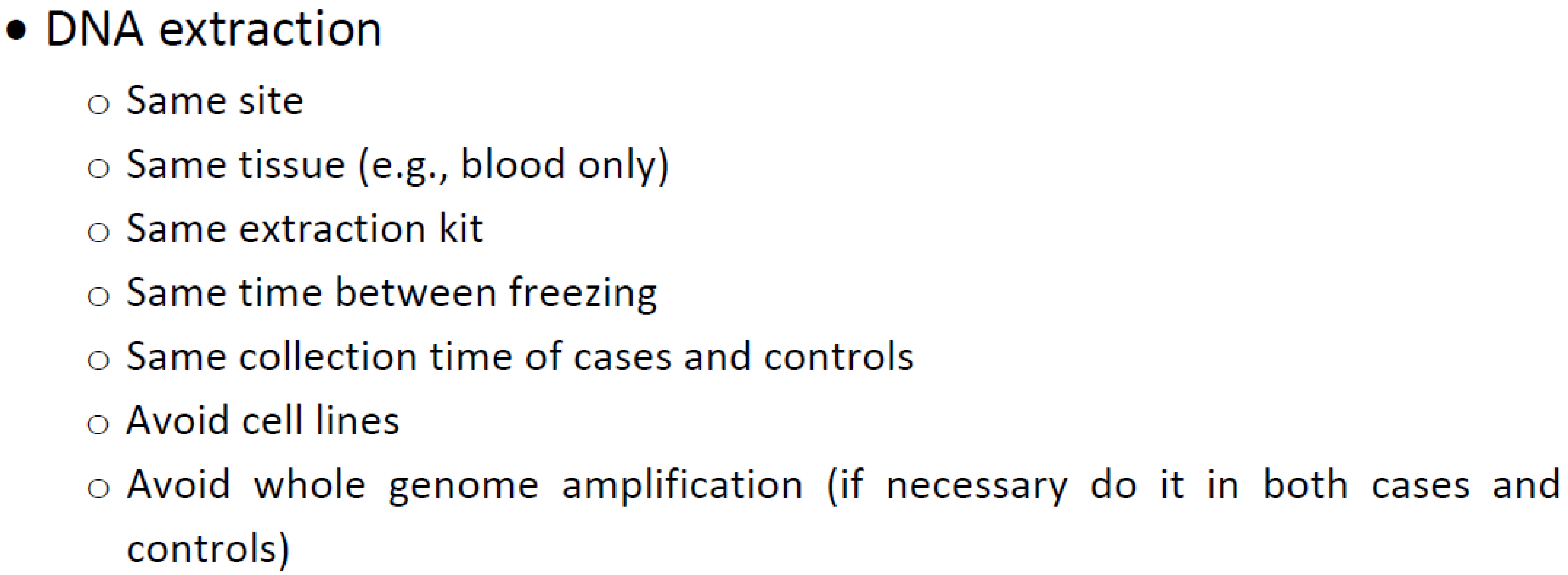
How can technology bias be avoided?



- Standard experimental design problems
  - Cases and controls not balanced / randomized across plates
  - Controls borrowed from other studies
  - Trios/families split across plates
  - Genotyping performed at different sites and / or using different technologies and / or chips
- Consequences of design problems
  - Batch effects
  - High type I error fractions
  - Up to 50% of top hits discarded
  - Analyses of copy number variation extremely compromised

(Ziegler and Van Steen, Brazil 2010)

How can technology bias be avoided?



- DNA extraction
  - ○ Same site
  - ○ Same tissue (e.g., blood only)
  - ○ Same extraction kit
  - ○ Same time between freezing
  - ○ Same collection time of cases and controls
  - ○ Avoid cell lines
  - ○ Avoid whole genome amplification (if necessary do it in both cases and controls)

(Ziegler and Van Steen, Brazil 2010)

How can technology bias be avoided?

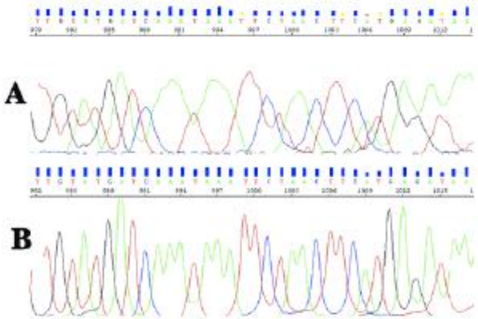
- Plating
  - Randomize phenotype/s across plates using statistical design
  - Stratify by gender
  - Run technical duplicates within and across plates to assess variability
  - Keep families together
  - Do it yourself, do not leave it to the laboratory
- Genotyping
  - All chips from single manufacturing lot
  - Genotype at single site
  - Genotype over shortest period of time possible
  - Avoid day effects, e.g., by using same technician over time
  - Re-genotype bad samples

(Ziegler and Van Steen, Brazil 2010)

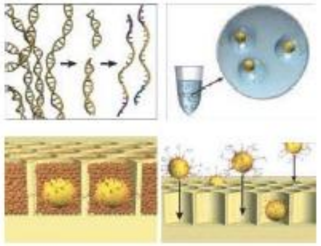
Next generation sequencing will overtake array technology?

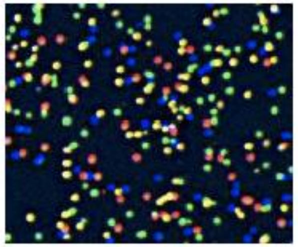
- The competing hypothesis to the CDCV hypothesis is the **Common Disease/Rare Variant** (CDRV) hypothesis.
- It argues that multiple rare DNA sequence variations, each with relatively high penetrance, are the major contributors to genetic susceptibility to common diseases.
- Although some common variants that underlie complex diseases have been identified, and given the recent huge financial and scientific investment in GWA studies, there is no longer a great deal of evidence in support of the CDCV hypothesis and much of it is equivocal...
- Hence, nowadays, both CDCV and CDRV hypotheses have their place in current research efforts.

Next generation sequencing will overtake array technology?



NEW HiSeq 2500
Remarkable speed and flexibility.
MiSeq
Simplicity, integration, and ease-of-use.





Sanger Sequencing
ABI 3730
Electrophoresis
1000 base reads

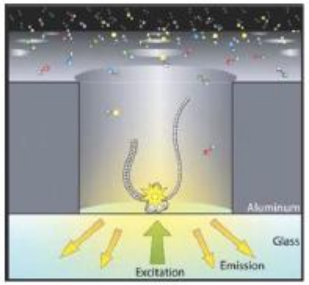
Roche GS-FLX, LifeTechnologies SOLID5500, Illumina HiSeq2500
Clonal amplification

400 base, 75 base, 100 base reads

1st Generation
1977 -

2nd Generation
2005 -

Next generation sequencing will overtake array technology?



IonTorrent, Proton
Pacific Biosciences, RS
Enzyme/Readout
400 base, 1000 base reads, and strobing

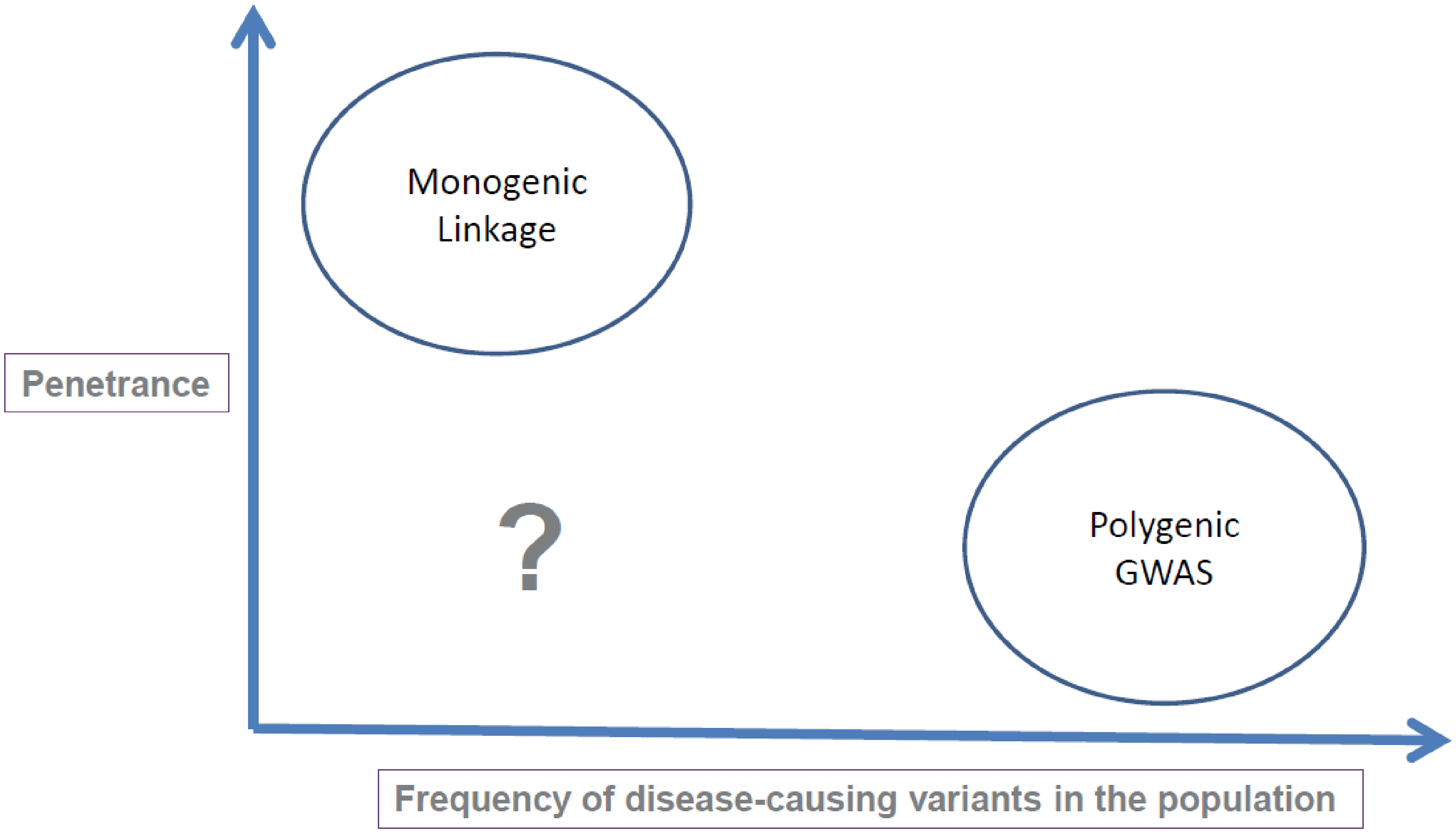
Oxford Nanopore Technologies, GridION
Direct Readout

Very long ~10s kbase reads

2.5th Generation
2011 -

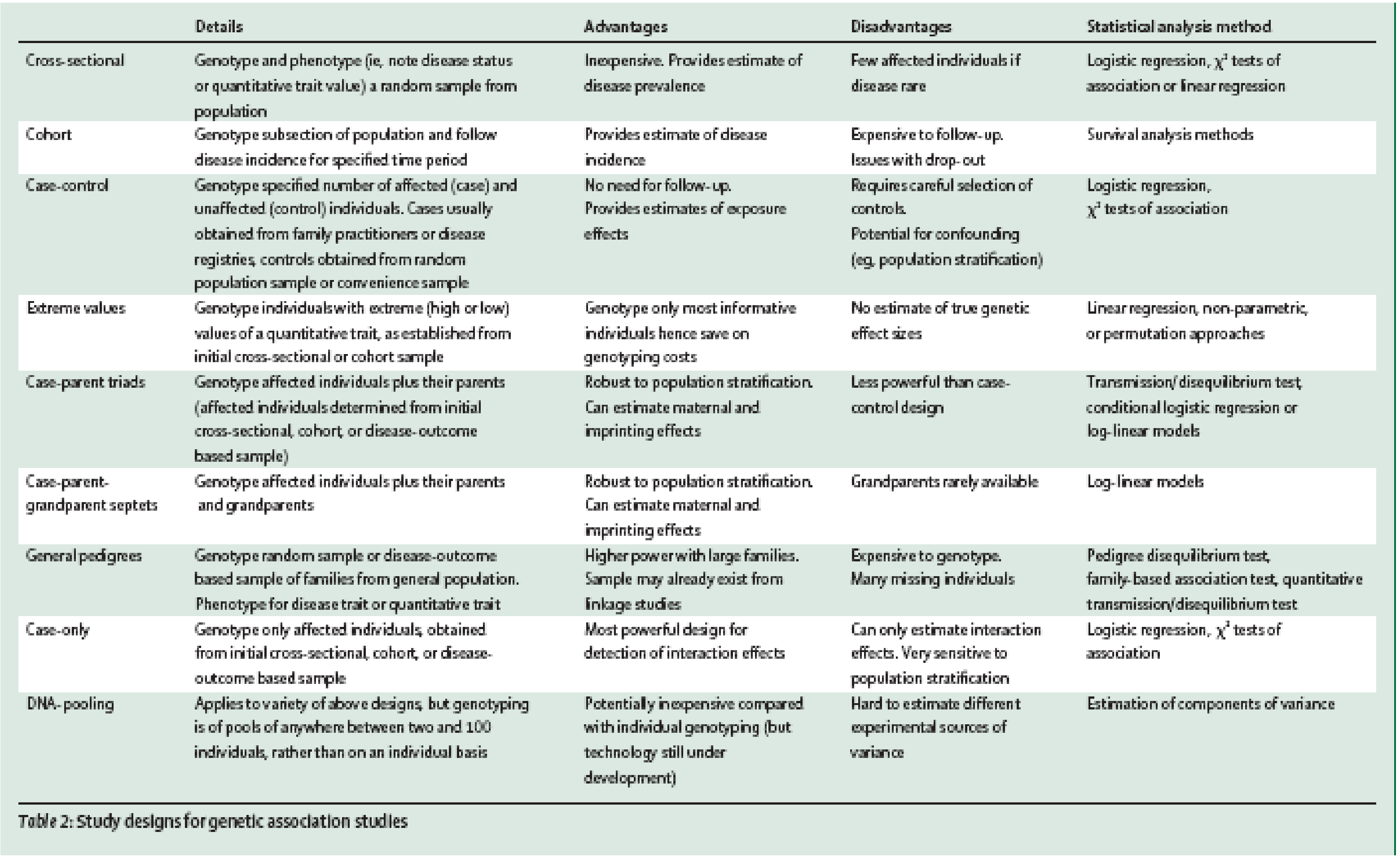
3rd Generation
2013 -

Crucial question: How to best capture disease predisposition?



(Gut 2012)

Which study subjects to select?



	Details	Advantages	Disadvantages	Statistical analysis method
Cross-sectional	Genotype and phenotype (ie, note disease status or quantitative trait value) a random sample from population	Inexpensive. Provides estimate of disease prevalence	Few affected individuals if disease rare	Logistic regression, χ^2 tests of association or linear regression
Cohort	Genotype subsection of population and follow disease incidence for specified time period	Provides estimate of disease incidence	Expensive to follow-up. Issues with drop-out	Survival analysis methods
Case-control	Genotype specified number of affected (case) and unaffected (control) individuals. Cases usually obtained from family practitioners or disease registries, controls obtained from random population sample or convenience sample	No need for follow-up. Provides estimates of exposure effects	Requires careful selection of controls. Potential for confounding (eg, population stratification)	Logistic regression, χ^2 tests of association
Extreme values	Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample	Genotype only most informative individuals hence save on genotyping costs	No estimate of true genetic effect sizes	Linear regression, non-parametric, or permutation approaches
Case-parent triads	Genotype affected individuals plus their parents (affected individuals determined from initial cross-sectional, cohort, or disease-outcome based sample)	Robust to population stratification. Can estimate maternal and imprinting effects	Less powerful than case-control design	Transmission/disequilibrium test, conditional logistic regression or log-linear models
Case-parent-grandparent septets	Genotype affected individuals plus their parents and grandparents	Robust to population stratification. Can estimate maternal and imprinting effects	Grandparents rarely available	Log-linear models
General pedigrees	Genotype random sample or disease-outcome based sample of families from general population. Phenotype for disease trait or quantitative trait	Higher power with large families. Sample may already exist from linkage studies	Expensive to genotype. Many missing individuals	Pedigree disequilibrium test, family-based association test, quantitative transmission/disequilibrium test
Case-only	Genotype only affected individuals, obtained from initial cross-sectional, cohort, or disease-outcome based sample	Most powerful design for detection of interaction effects	Can only estimate interaction effects. Very sensitive to population stratification	Logistic regression, χ^2 tests of association
DNA-pooling	Applies to variety of above designs, but genotyping is of pools of anywhere between two and 100 individuals, rather than on an individual basis	Potentially inexpensive compared with individual genotyping (but technology still under development)	Hard to estimate different experimental sources of variance	Estimation of components of variance

Table 2: Study designs for genetic association studies

(Cordell and Clayton 2005)

Which study subjects to select?

- Cohort studies
  - Assumption I: Participants under study representative for population of interest
  - Assumption II: Phenotypes ascertained similarly in subjects with and without the relevant genetic variants
  - Advantage I: Incident cases, free of survival bias
  - Advantage II: If prevalent cases available, too, comparison of incident and prevalent cases possible
  - Advantage III: Availability of intermediate phenotypes (quantitative traits) with distribution as in population
  - Advantage IV: Direct measure of risk
  - Advantage V: Fewer bias than case-control studies
  - Disadvantage I: Long follow-up required

(Ziegler and Van Steen, 2010)

Which study subjects to select?

- Cohort studies (continued)
  - Disadvantage II: Large sample size required
  - Disadvantage III: Expensive
  - Disadvantage IV: Poorly suited for studying rare diseases
  - Disadvantage VII: Unbalanced distribution of cases and controls
  - Disadvantage V: Consent for GWA genotyping often required
  - Disadvantage VI: Consent for data sharing often required
  - Disadvantage VIII: DNA quality

(Ziegler and Van Steen, 2010)

Which study subjects to select?

- Family-based association studies
  - ○ Assumption I: Families representative for population of interest
  - ○ Assumption II: Same genetic background in both parents
  - ○ Advantage I: Controls immune to population stratification, i.e., no spurious associations, i.e., no association without linkage
  - ○ Advantage II: Checks for Mendelian inheritance possible, i.e., fewer genotyping errors
  - ○ Advantage III: Parental phenotyping not required
  - ○ Advantage IV: Simple logistics for diseases in children
  - ○ Advantage V: Allows investigation of imprinting
  - ○ Disadvantage I: Cost inefficient
  - ○ Disadvantage II: Lower power when compared with case-control studies
  - ○ Disadvantage III: Sensitive to genotyping errors

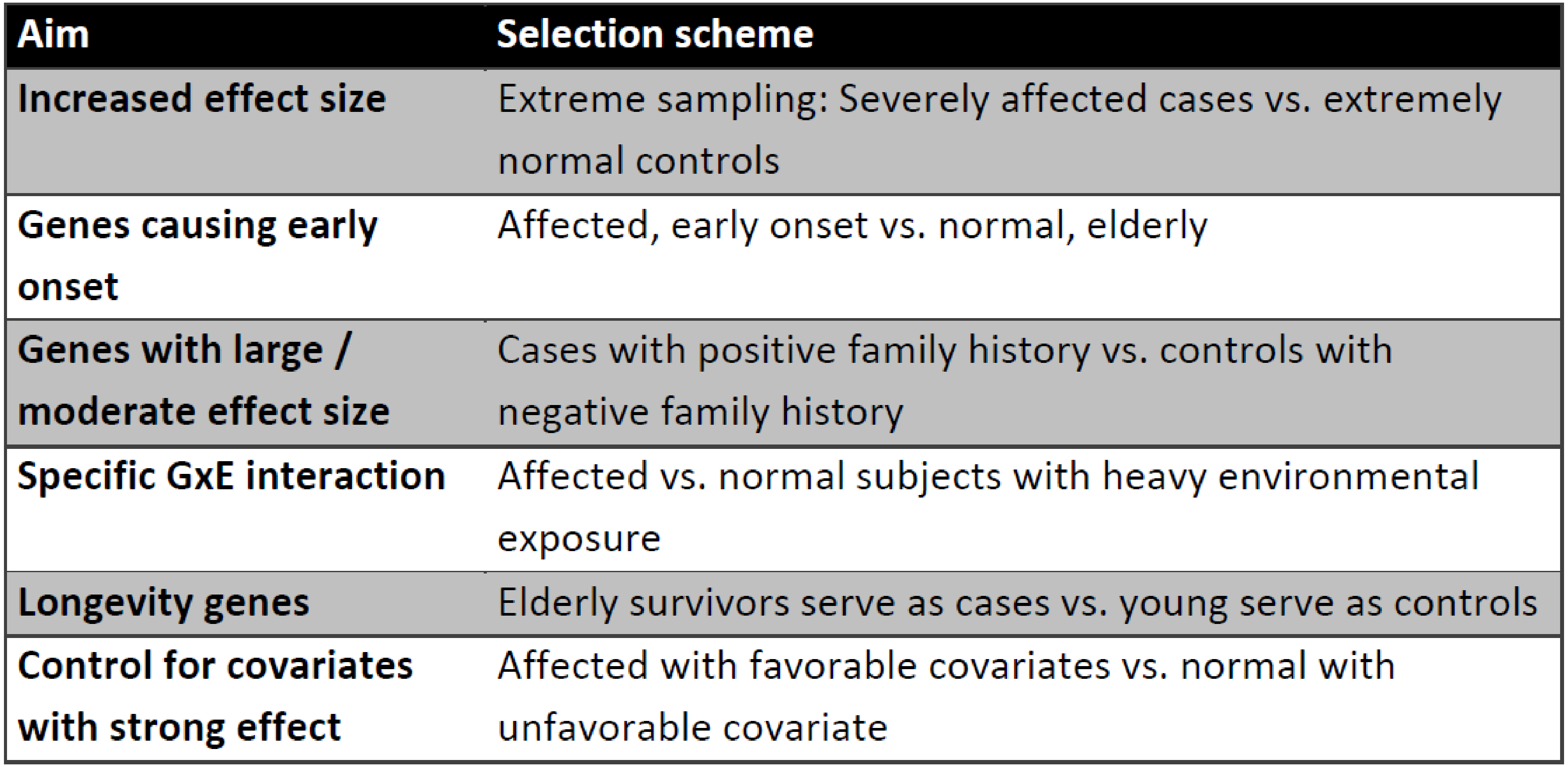
(Ziegler and Van Steen, 2010)

Which study subjects to select?

- Case-control studies
  - Assumption I: Cases and controls drawn from same population
  - Assumption II: Cases representative for all cases in population
  - Assumption III: All data collected similarly in cases and controls
  - Advantage I: Simple
  - Advantage II: Cheap
  - Advantage III: Large number of cases and controls available
  - Advantage IV: Optimal for studying rare diseases
  - Disadvantage I: Prone to population stratification
  - Disadvantage II: Prone to batch effects
  - Disadvantage III: Prone to other biases
  - Disadvantage IV: Cases usually prevalent ↓ fatal, short episodes, mild cases …
  - Disadvantage V: Overestimation of risk for common disease

(Ziegler and Van Steen, 2010)

Which study subjects to select?

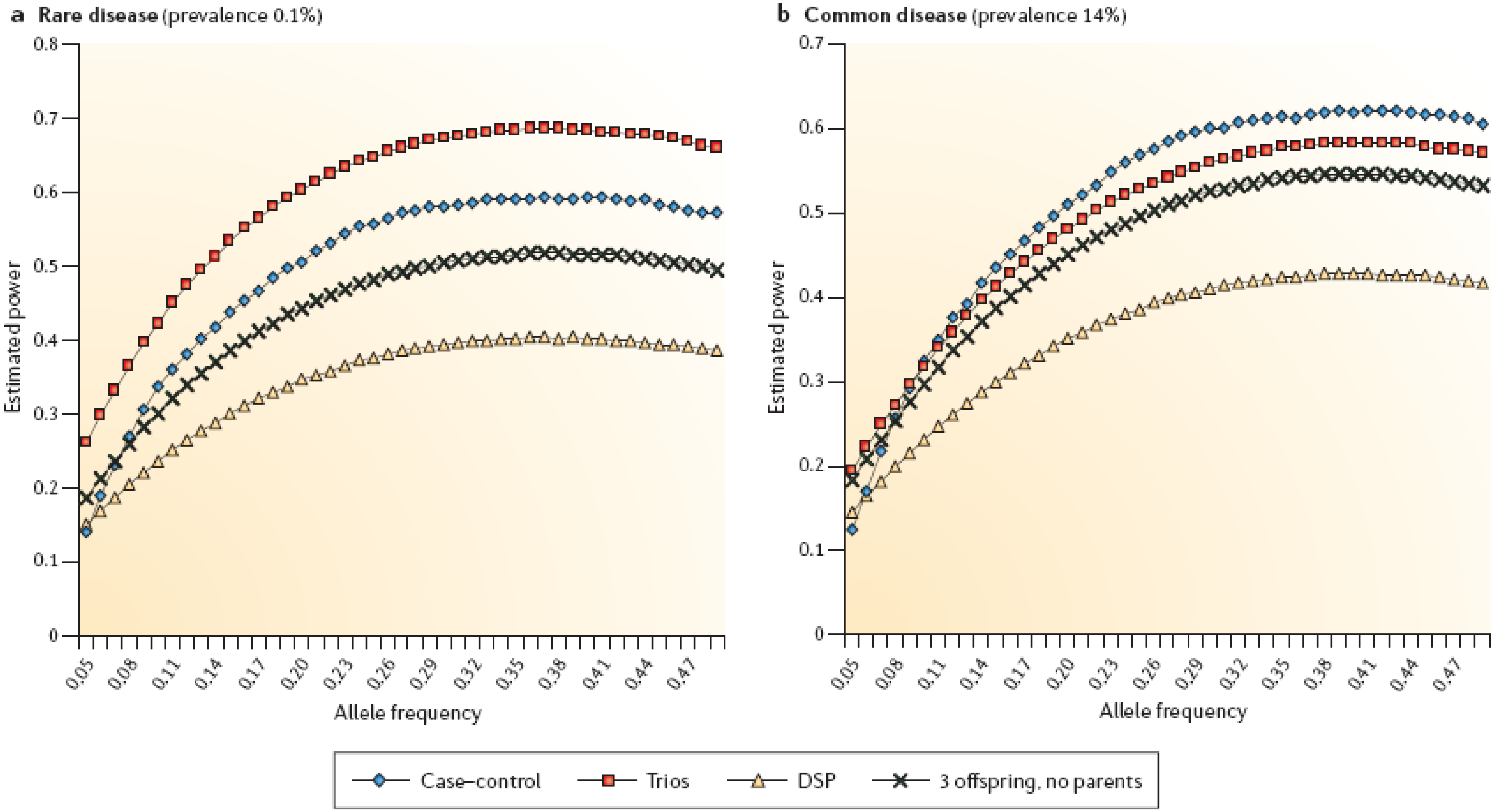


Aim	Selection scheme
Increased effect size	Extreme sampling: Severely affected cases vs. extremely normal controls
Genes causing early onset	Affected, early onset vs. normal, elderly
Genes with large / moderate effect size	Cases with positive family history vs. controls with negative family history
Specific GxE interaction	Affected vs. normal subjects with heavy environmental exposure
Longevity genes	Elderly survivors serve as cases vs. young serve as controls
Control for covariates with strong effect	Affected with favorable covariates vs. normal with unfavorable covariate

Morton & Collins 1998 Proc Natl Acad Sci USA 95:11389

Which study subjects to select?

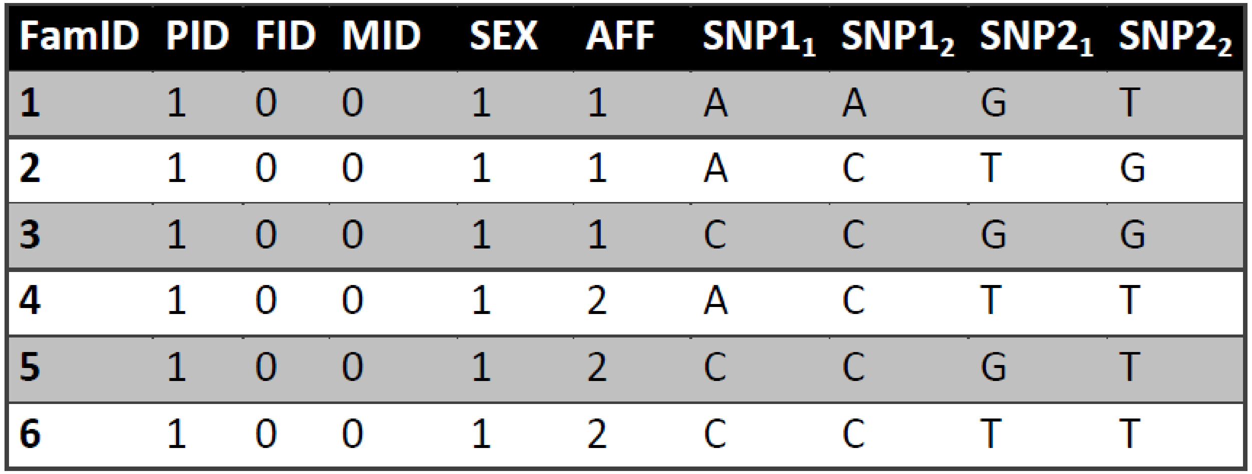
Rare versus common diseases (Lange and Laird 2006)



3 Preliminary analyses

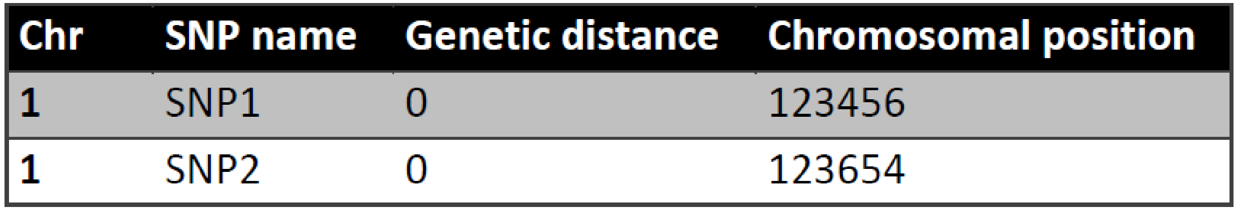
Is there a standard file format for GWA studies?

Standard data format: tped = transposed ped format file



FamID	PID	FID	MID	SEX	AFF	SNP1$_1$	SNP1$_2$	SNP2$_1$	SNP2$_2$
1	1	0	0	1	1	A	A	G	T
2	1	0	0	1	1	A	C	T	G
3	1	0	0	1	1	C	C	G	G
4	1	0	0	1	2	A	C	T	T
5	1	0	0	1	2	C	C	G	T
6	1	0	0	1	2	C	C	T	T

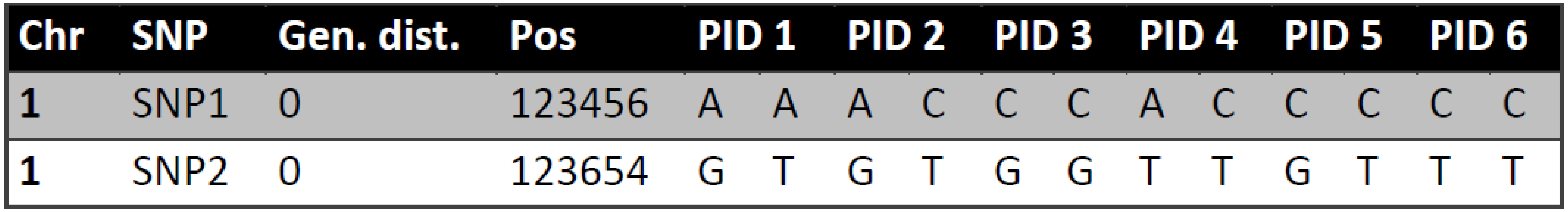
ped file



Chr	SNP name	Genetic distance	Chromosomal position
1	SNP1	0	123456
1	SNP2	0	123654

map file

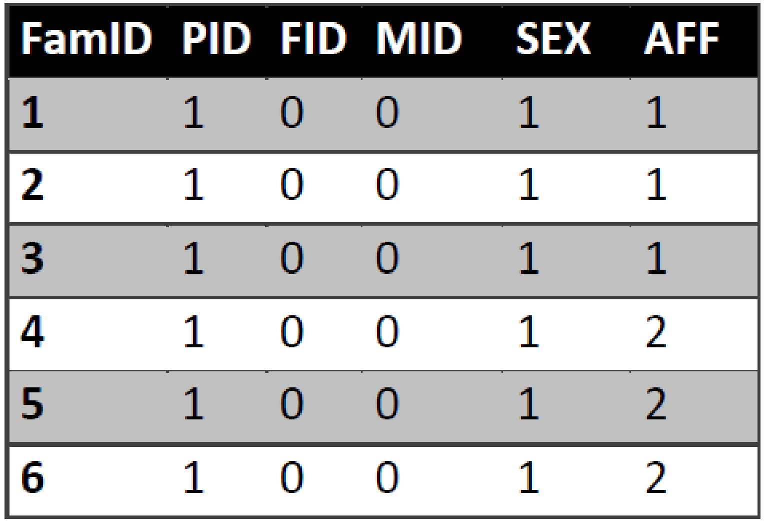
Is there a standard file format for GWA studies?



Chr	SNP	Gen. dist.	Pos	PID 1		PID 2		PID 3		PID 4		PID 5		PID 6	
1	SNP1	0	123456	A	A	A	C	C	C	A	C	C	C	C	C
1	SNP2	0	123654	G	T	G	T	G	G	T	T	G	T	T	T

tped file

tfam file: First 6 columns of standard ped file



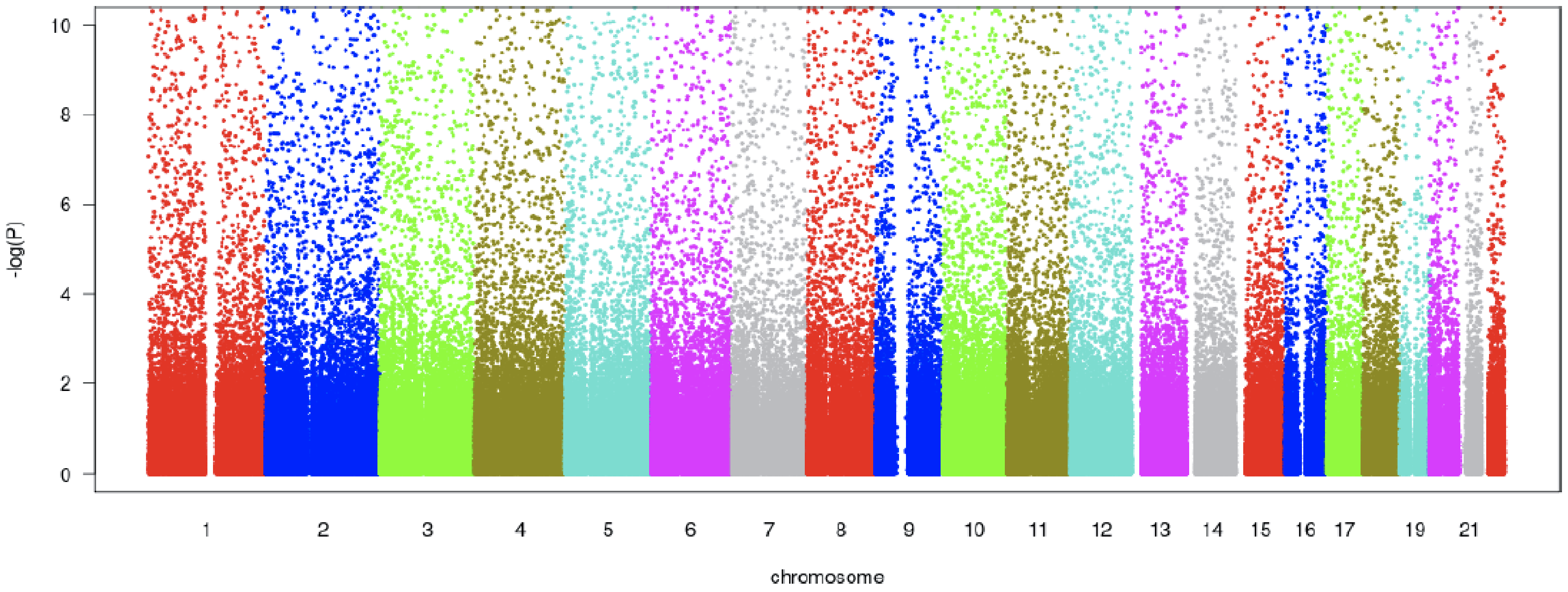
FamID	PID	FID	MID	SEX	AFF
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

tfam file

3.a Quality control

Why is quality control important?

BEFORE (false positives !!!!):

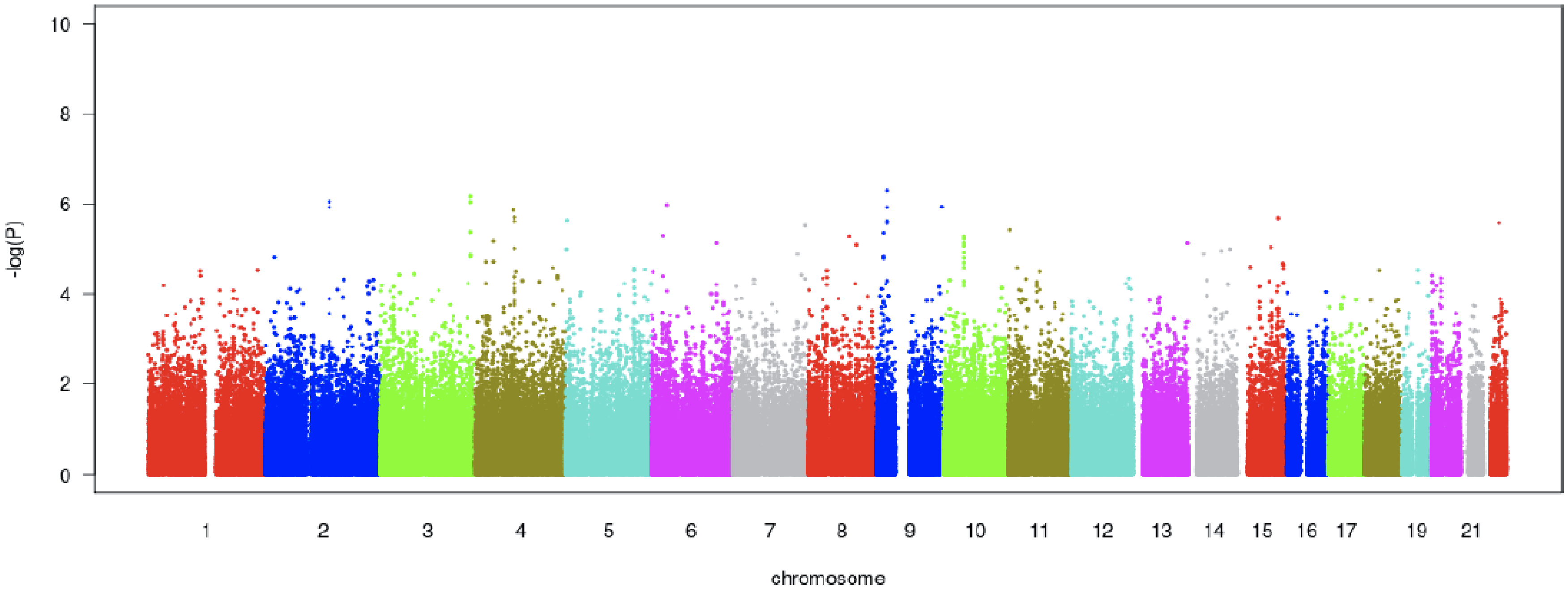


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(Ziegler and Van Steen 2010)

Why is quality control important?

AFTER:



Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

SNPs passing standard quality control: 270,701

(Ziegler and Van Steen 2010)

What is the standard quality control?

- Quality control on different levels:
  - Subject or sample level
  - SNP level
  - X-chromosomal SNP level

What are standard filters on the sample level?

- Call fraction as high as possible
- Cryptic relatedness: if identity by state (IBS) too high, subjects closely related
- Ethnic origin (principal component, multidimensional scaling, non-metric multidimensional scaling): homogeneous study populations required
- No excess or deficiency of heterozygosity (contamination of DNA, hybridization failure)

(Ziegler and Van Steen 2010)

What are standard filters on the SNP level?

- Minor allele frequency (MAF)
    - ○ Genotype calling algorithms perform poorly for SNPs with low MAF
    - ○ Power low for detecting associations to SNPs with low MAF,
- Missing frequency (MiF)
    - ○ Also termed 1 minus SNP call rate
    - ○ Indicator for cluster separation
    - ○ Investigate MiF separately in cases and in controls because of differential missingness
- Hardy-Weinberg equilibrium (HWE)
    - ○ SNPs excluded if substantially more or fewer subjects heterozygous at a SNP than expected (excess heterozygosity or heterozygote deficiency)

(Ziegler and Van Steen 2010)

What is Hardy-Weinberg Equilibrium (HWE)?

Consider diallelic SNP with alleles A_1 and A_2

- Genotype frequencies
  $P(A_1A_1) = p_{11}$, $P(A_1A_2) = p_{12}$, $P(A_2A_2) = p_{22}$
- Allele frequencies $P(A_1) = p = p_{11} + \frac{1}{2}p_{12}$, $P(A_2) = q = p_{22} + \frac{1}{2}p_{12}$

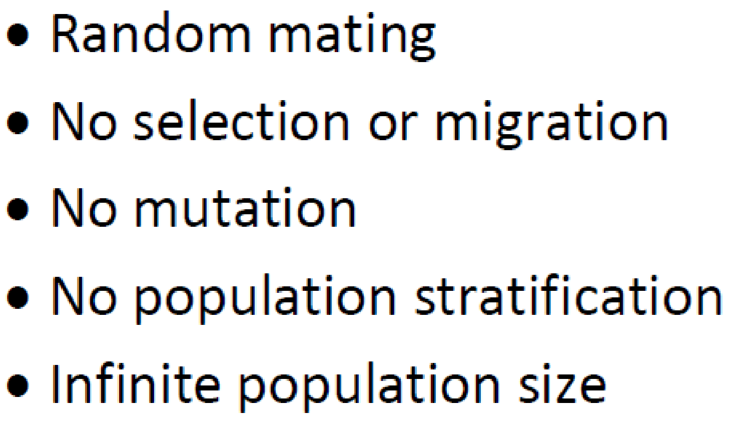
If

- $P(A_1A_1) = p_{11} = p^2$
- $P(A_1A_2) = p_{12} = 2pq$
- $P(A_2A_2) = p_{22} = q^2$

the population is said to be in HWE at the SNP

(Ziegler and Van Steen 2010)

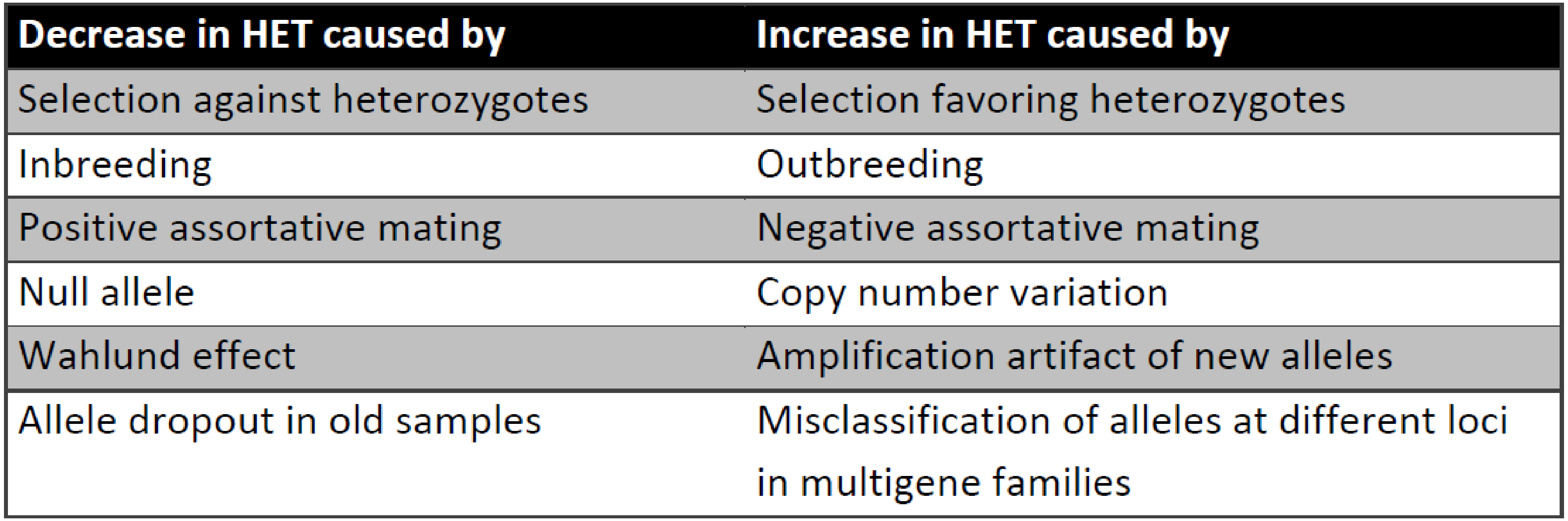
What are the assumptions of HWE?



- Random mating
- No selection or migration
- No mutation
- No population stratification
- Infinite population size

What are signs of deviations from HWE?

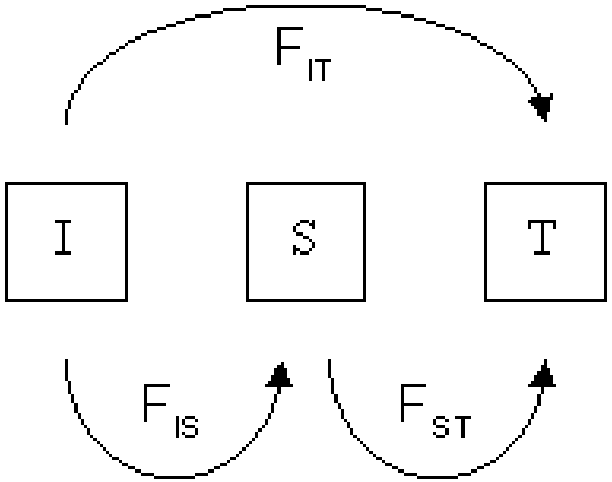
Decreased or increased HET



Decrease in HET caused by	Increase in HET caused by
Selection against heterozygotes	Selection favoring heterozygotes
Inbreeding	Outbreeding
Positive assortative mating	Negative assortative mating
Null allele	Copy number variation
Wahlund effect	Amplification artifact of new alleles
Allele dropout in old samples	Misclassification of alleles at different loci in multigene families

Ziegler & König 2010 ISBN-13 978-3-527-32389-0
Hedrick 2009 9780763757373

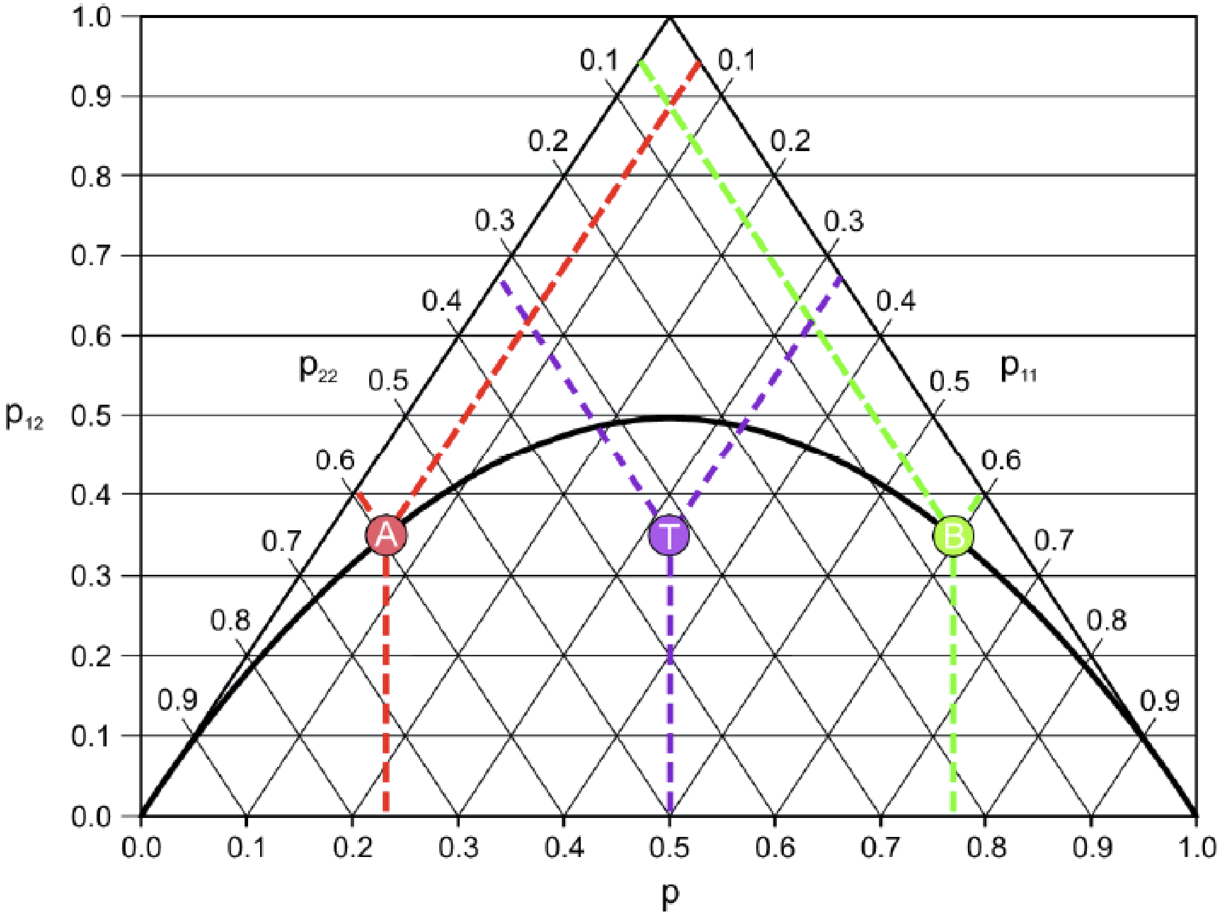
What are signs of deviations from HWE?



F_{IT} is the inbreeding coefficient of an individual (**I**) relative to the total (**T**) population, as above; **F_{IS}** is the inbreeding coefficient of an individual (**I**) relative to the subpopulation (**S**), using the above for subpopulations and averaging them; and **F_{ST}** is the effect of subpopulations (**S**) compared to the total population (**T**)

What are signs of deviations from HWE?

Increased HOM (e.g., in case of population stratification; Wahlund effect)



How can HWE be measured?

- Disequilibrium coefficient: $\mathcal{D}_{A_1} = p_{11} - p^2 = p_{22} - q^2 = -p_{12} + 2pq$
- Inbreeding coefficient:
  - Assume $P(A_1) = p$, probability of 2nd allele to be identical f
  - Prob. of two A_1 alleles equal to $p \cdot f$
  - Prob. for two independent A_1 alleles p^2
  - Ergo: $P(A_1A_1) = p^2(1-f) + pf = p^2 + fpq$

    $P(A_1A_2) = 2pq - 2fpq = 2pq(1-f)$
- Excess heterozygosity: $\gamma = p_{12} \big/ \left(2\sqrt{p_{11}p_{22}}\right)$

- Standard procedure in GWA studies: χ^2 lack of fit test

How can HWE be measured?

- The Pearson test is easy to compute, but the χ2 approximation can be poor when there are low genotype counts, in which case it is better to use a Fisher exact test, which does not rely on the χ2 approximation.
- Discard loci that, for example, deviate from HWE among controls at significance level $\alpha = 10^{-3}$ or 10^{-4}. But be flexible !
- The open-source data-analysis software R includes the "*SNPassoc*" package that implements an exact SNP test of Hardy-Weinberg Equilibrium (http://www.sph.umich.edu/csg/abecasis/Exact/snp_hwe.r)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

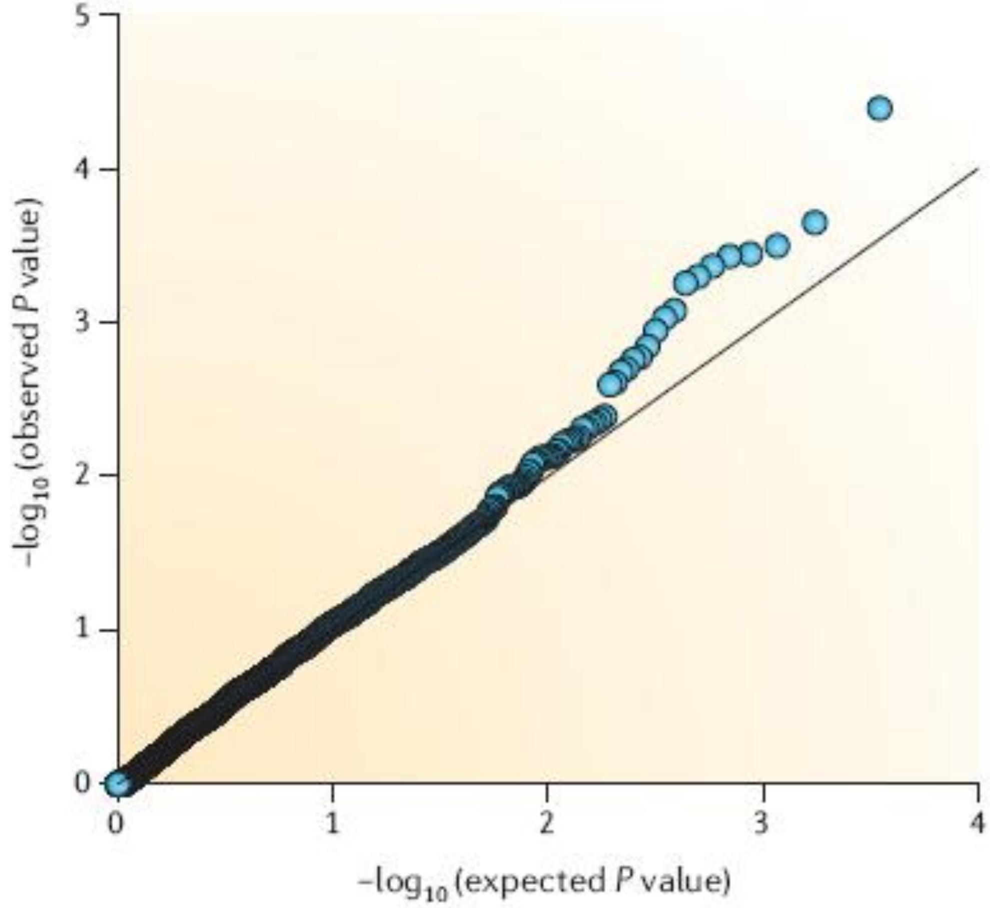
Expectations computed under the null of HWE

Nr of degrees of freedom is 1 (p+q=1)

How can HWE be measured?

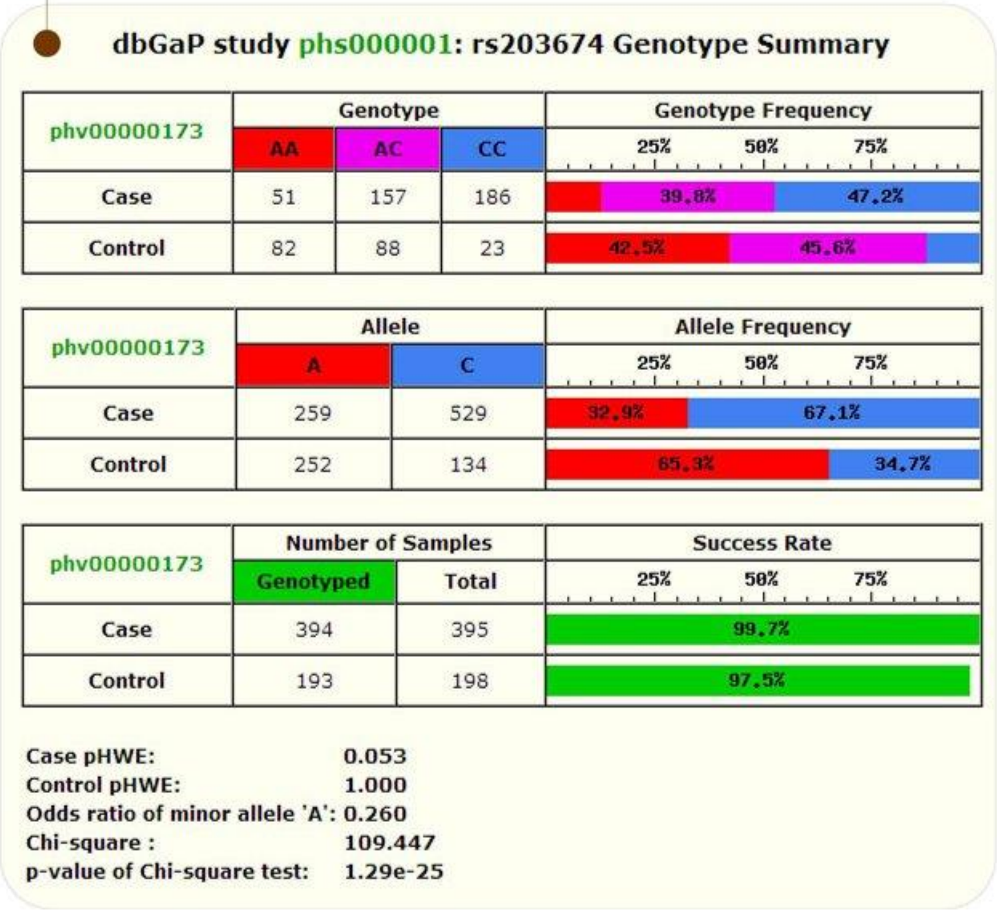
- A useful tool for interpreting the results of HWE and other tests on many SNPs is the log quantile–quantile (QQ) *p*-value plot:
    - the negative logarithm of the *i*-th smallest *p*-value is plotted against $-\log (i / (L + 1))$, where *L* is the number of SNPs.
- The 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- A 45-degree reference line is also plotted as visualization tool:
    - If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
    - The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

How can HWE be measured?



(Balding 2006)

Why is cluster plot reading important?



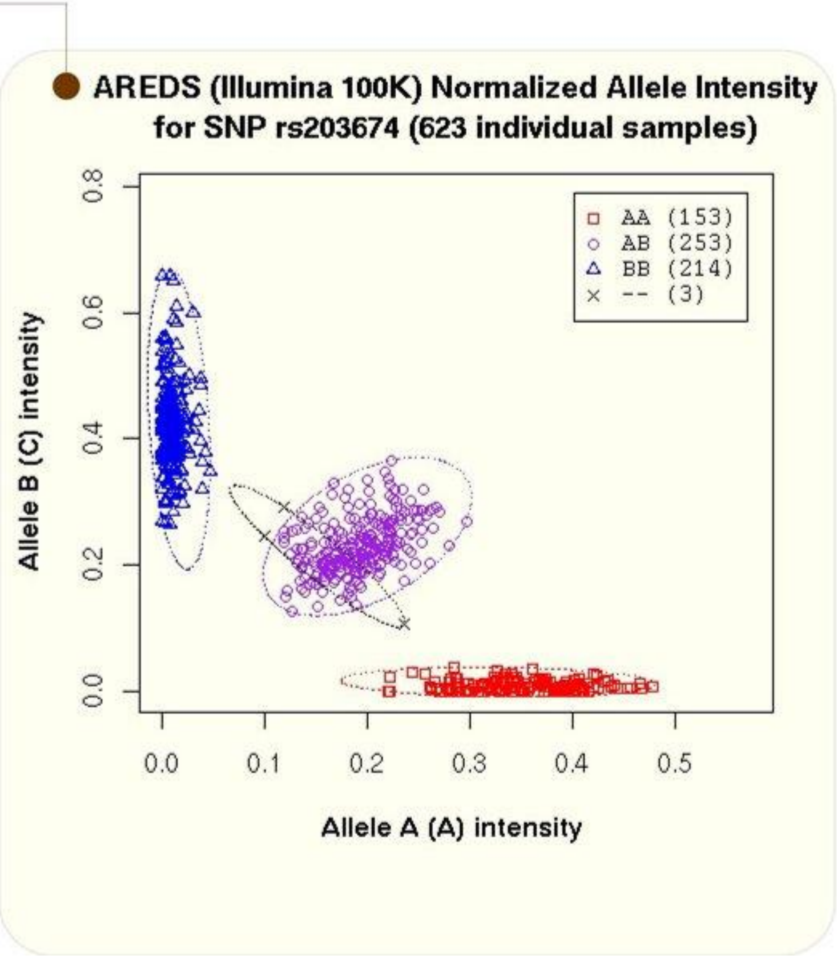
dbGaP study phs000001: rs203674 Genotype Summary

phv00000173	Genotype			Genotype Frequency
	AA	AC	CC	25% 50% 75%
Case	51	157	186	39.8% 47.2%
Control	82	88	23	42.5% 45.6%

phv00000173	Allele		Allele Frequency
	A	C	25% 50% 75%
Case	259	529	32.9% 67.1%
Control	252	134	65.3% 34.7%

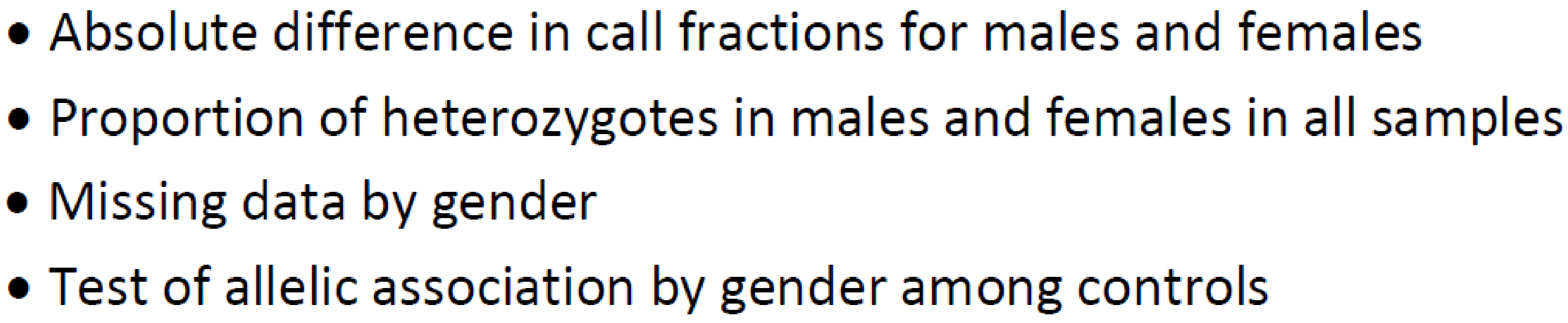
phv00000173	Number of Samples		Success Rate
	Genotyped	Total	25% 50% 75%
Case	394	395	99.7%
Control	193	198	97.5%

Case pHWE:	0.053
Control pHWE:	1.000
Odds ratio of minor allele 'A':	0.260
Chi-square :	109.447
p-value of Chi-square test:	1.29e-25



AREDS (Illumina 100K) Normalized Allele Intensity for SNP rs203674 (623 individual samples)

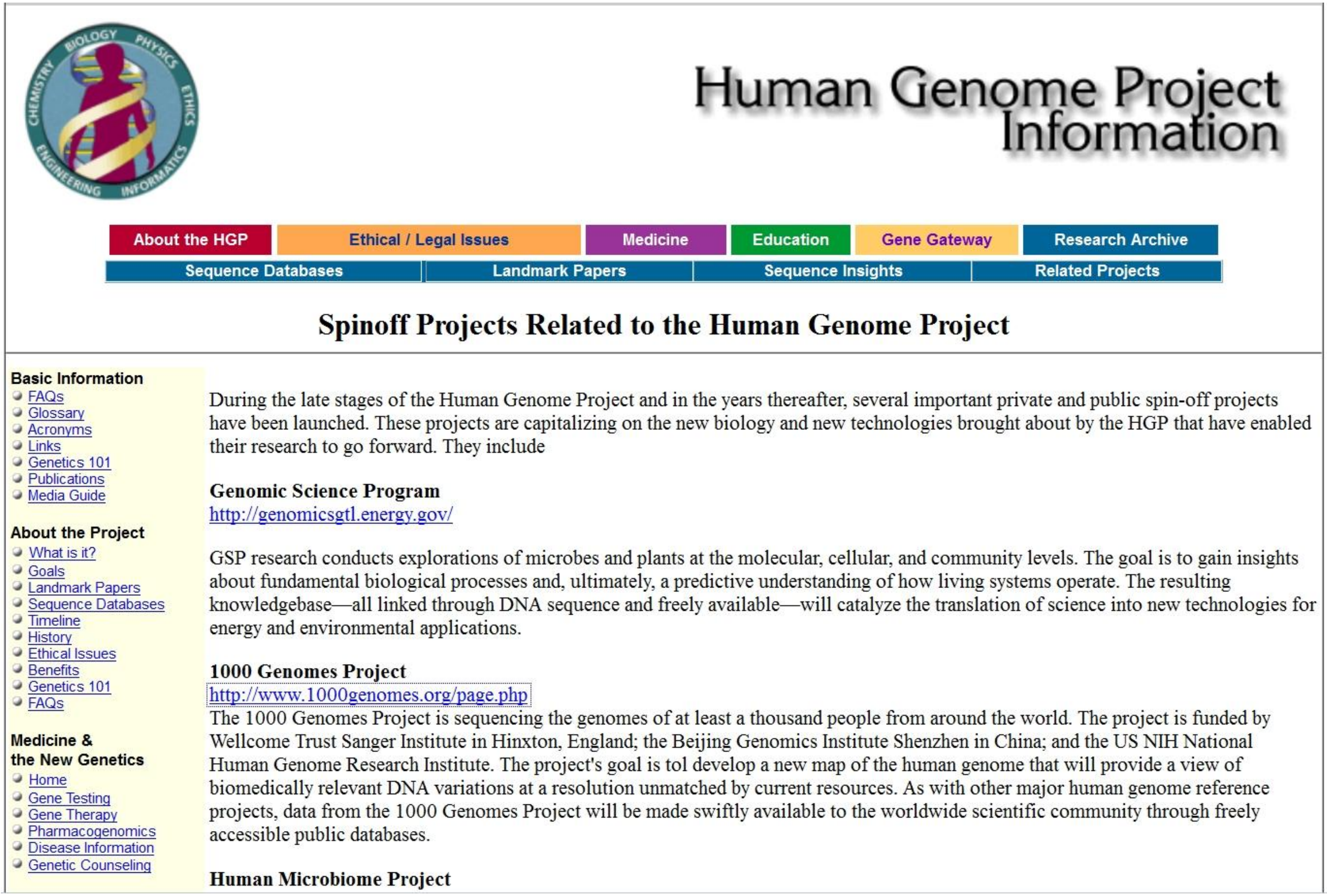
What are standard filters on the gender level?



- Absolute difference in call fractions for males and females
- Proportion of heterozygotes in males and females in all samples
- Missing data by gender
- Test of allelic association by gender among controls

(Ziegler and Van Steen 2010)

Is there a power advantage in imputing?



Spinoff Projects Related to the Human Genome Project

Basic Information
- FAQs
- Glossary
- Acronyms
- Links
- Genetics 101
- Publications
- Media Guide

About the Project
- What is it?
- Goals
- Landmark Papers
- Sequence Databases
- Timeline
- History
- Ethical Issues
- Benefits
- Genetics 101
- FAQs

Medicine & the New Genetics
- Home
- Gene Testing
- Gene Therapy
- Pharmacogenomics
- Disease Information
- Genetic Counseling

During the late stages of the Human Genome Project and in the years thereafter, several important private and public spin-off projects have been launched. These projects are capitalizing on the new biology and new technologies brought about by the HGP that have enabled their research to go forward. They include

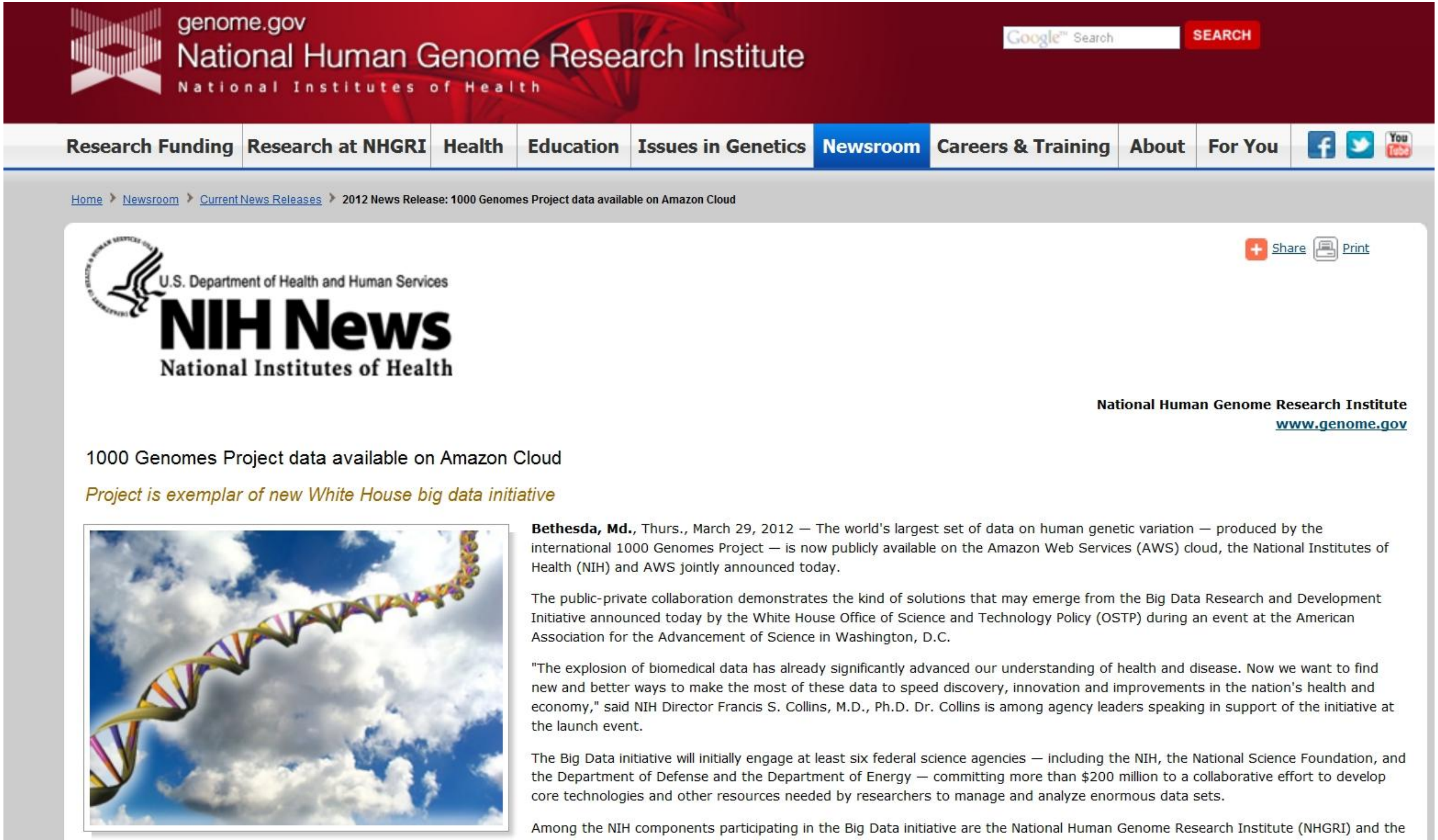
Genomic Science Program
http://genomicsgtl.energy.gov/

GSP research conducts explorations of microbes and plants at the molecular, cellular, and community levels. The goal is to gain insights about fundamental biological processes and, ultimately, a predictive understanding of how living systems operate. The resulting knowledgebase—all linked through DNA sequence and freely available—will catalyze the translation of science into new technologies for energy and environmental applications.

1000 Genomes Project
http://www.1000genomes.org/page.php
The 1000 Genomes Project is sequencing the genomes of at least a thousand people from around the world. The project is funded by Wellcome Trust Sanger Institute in Hinxton, England; the Beijing Genomics Institute Shenzhen in China; and the US NIH National Human Genome Research Institute. The project's goal is tol develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

Human Microbiome Project

Is there a power advantage in imputing?


genome.gov
National Human Genome Research Institute
National Institutes of Health
Google™ Search
SEARCH
Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
For You
Home > Newsroom > Current News Releases > 2012 News Release: 1000 Genomes Project data available on Amazon Cloud
Share
Print
U.S. Department of Health and Human Services
NIH News
National Institutes of Health

National Human Genome Research Institute
www.genome.gov

1000 Genomes Project data available on Amazon Cloud

Project is exemplar of new White House big data initiative

Bethesda, Md., Thurs., March 29, 2012 — The world's largest set of data on human genetic variation — produced by the international 1000 Genomes Project — is now publicly available on the Amazon Web Services (AWS) cloud, the National Institutes of Health (NIH) and AWS jointly announced today.

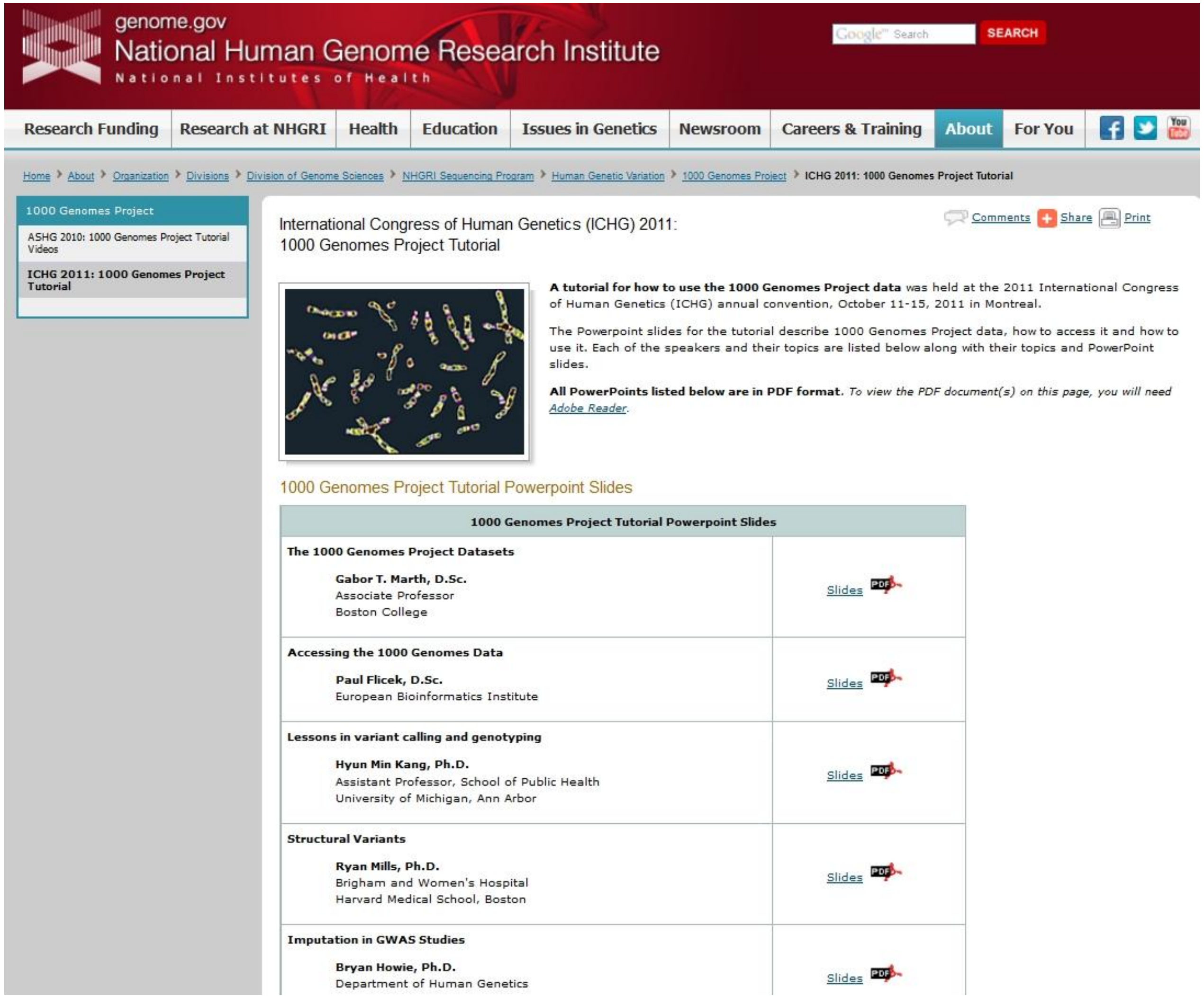
The public-private collaboration demonstrates the kind of solutions that may emerge from the Big Data Research and Development Initiative announced today by the White House Office of Science and Technology Policy (OSTP) during an event at the American Association for the Advancement of Science in Washington, D.C.

"The explosion of biomedical data has already significantly advanced our understanding of health and disease. Now we want to find new and better ways to make the most of these data to speed discovery, innovation and improvements in the nation's health and economy," said NIH Director Francis S. Collins, M.D., Ph.D. Dr. Collins is among agency leaders speaking in support of the initiative at the launch event.

The Big Data initiative will initially engage at least six federal science agencies — including the NIH, the National Science Foundation, and the Department of Defense and the Department of Energy — committing more than $200 million to a collaborative effort to develop core technologies and other resources needed by researchers to manage and analyze enormous data sets.

Among the NIH components participating in the Big Data initiative are the National Human Genome Research Institute (NHGRI) and the

Is there a power advantage in imputing?


genome.gov
National Human Genome Research Institute
National Institutes of Health
Google Search
SEARCH
Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
For You
Home > About > Organization > Divisions > Division of Genome Sciences > NHGRI Sequencing Program > Human Genetic Variation > 1000 Genomes Project > ICHG 2011: 1000 Genomes Project Tutorial
1000 Genomes Project
ASHG 2010: 1000 Genomes Project Tutorial Videos
ICHG 2011: 1000 Genomes Project Tutorial
Comments
Share
Print

International Congress of Human Genetics (ICHG) 2011:
1000 Genomes Project Tutorial

A tutorial for how to use the 1000 Genomes Project data was held at the 2011 International Congress of Human Genetics (ICHG) annual convention, October 11-15, 2011 in Montreal.

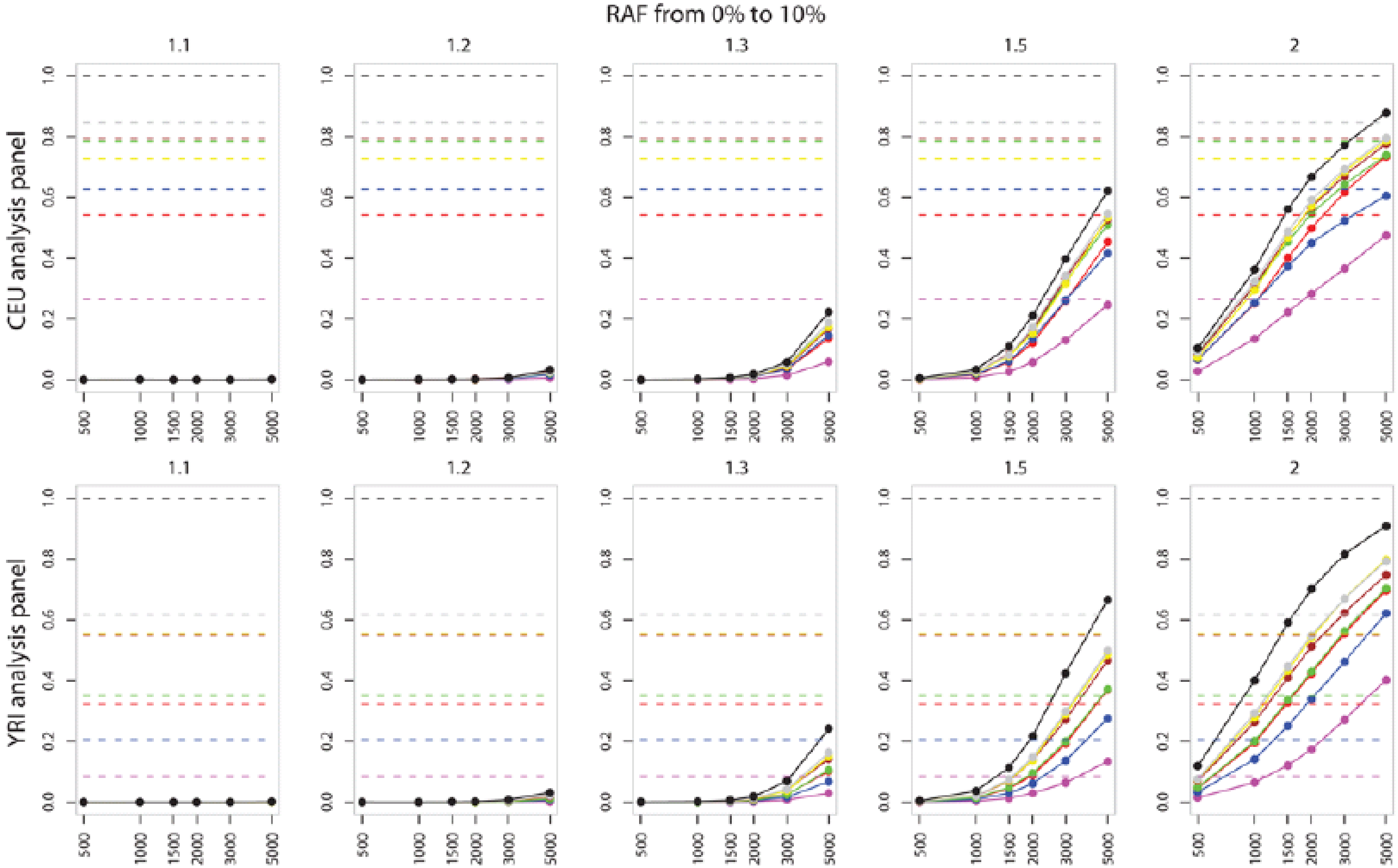
The Powerpoint slides for the tutorial describe 1000 Genomes Project data, how to access it and how to use it. Each of the speakers and their topics are listed below along with their topics and PowerPoint slides.

All PowerPoints listed below are in PDF format. *To view the PDF document(s) on this page, you will need Adobe Reader.*

1000 Genomes Project Tutorial Powerpoint Slides

1000 Genomes Project Tutorial Powerpoint Slides	
The 1000 Genomes Project Datasets **Gabor T. Marth, D.Sc.** Associate Professor Boston College	Slides PDF
Accessing the 1000 Genomes Data **Paul Flicek, D.Sc.** European Bioinformatics Institute	Slides PDF
Lessons in variant calling and genotyping **Hyun Min Kang, Ph.D.** Assistant Professor, School of Public Health University of Michigan, Ann Arbor	Slides PDF
Structural Variants **Ryan Mills, Ph.D.** Brigham and Women's Hospital Harvard Medical School, Boston	Slides PDF
Imputation in GWAS Studies **Bryan Howie, Ph.D.** Department of Human Genetics	Slides PDF

Is there a power advantage in imputing? (Spencer et al 2009)

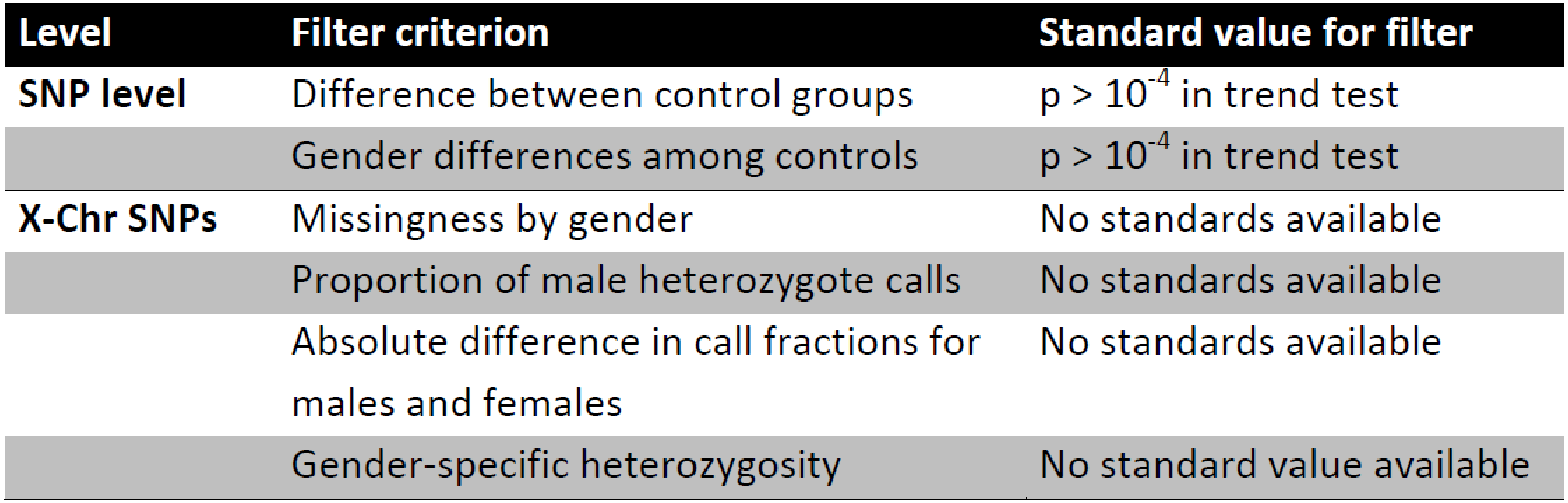


What are the Travemünde criteria?

Level	Filter criterion	Standard value for filter
Sample level	Call fraction	≥ 97%
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Mean ± 3 std.dev. over all samples
	Heterozygosity by gender	Mean ± 3 std.dev. within gender group
SNP level	MAF	≥ 1%
	MiF	≤ 2% in any study group, e.g., in both cases and controls
	MiF by gender	≤ 2% in any gender
	HWE	$p < 10^{-4}$

(Ziegler 2009)

What are the Travemünde criteria?

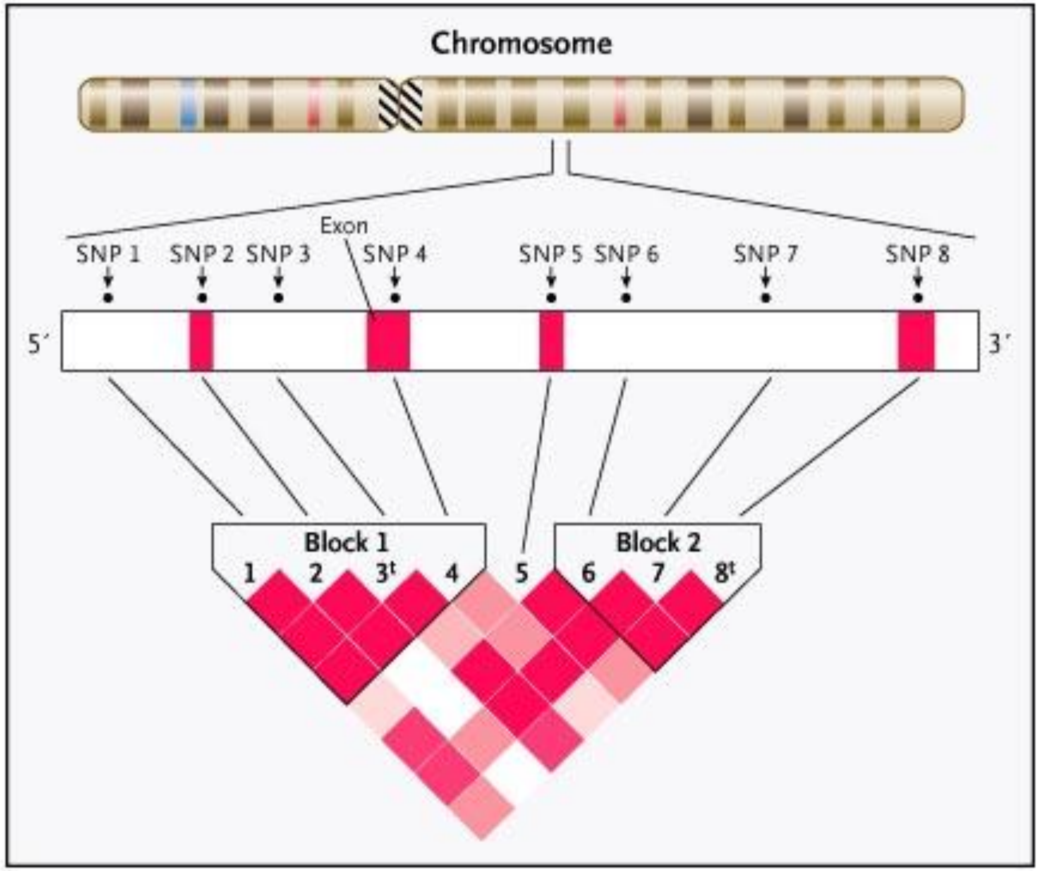


Level	Filter criterion	Standard value for filter
SNP level	Difference between control groups	$p > 10^{-4}$ in trend test
	Gender differences among controls	$p > 10^{-4}$ in trend test
X-Chr SNPs	Missingness by gender	No standards available
	Proportion of male heterozygote calls	No standards available
	Absolute difference in call fractions for males and females	No standards available
	Gender-specific heterozygosity	No standard value available

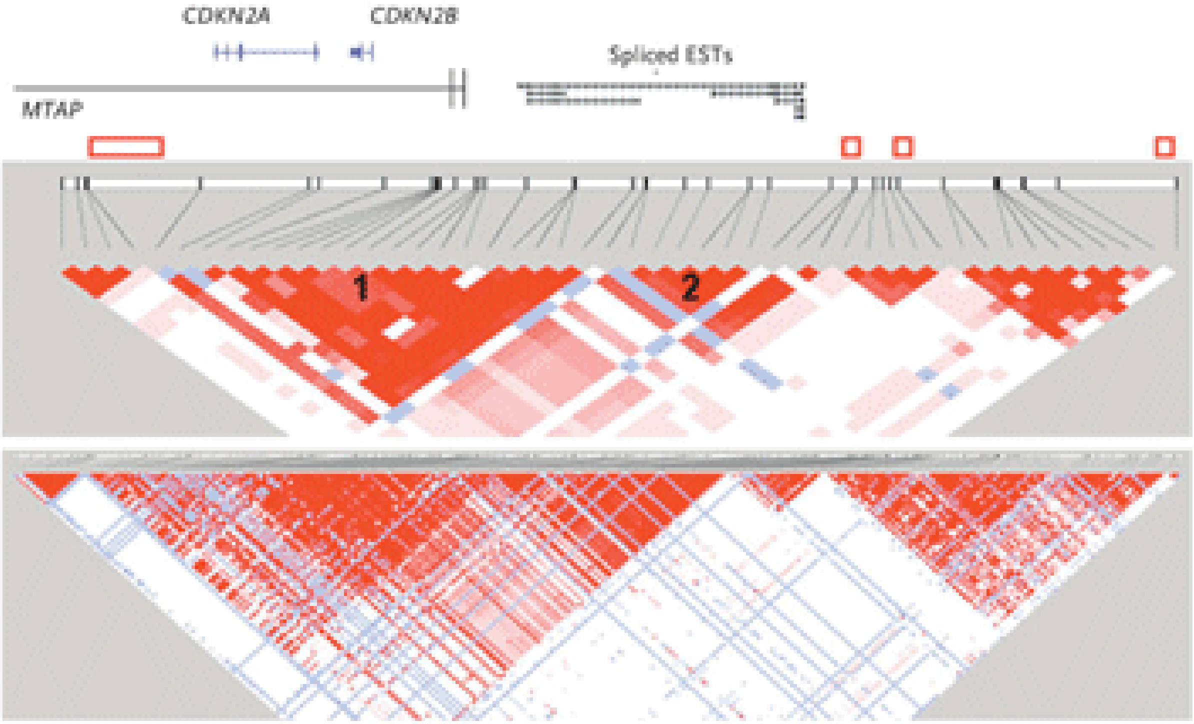
(Ziegler 2009)

3.b Linkage disequilibrium, haplotypes and SNP tagging

Mapping the relationships among SNPs (Christensen and Murray 2007)



Relationships among SNPs induce multiple signals



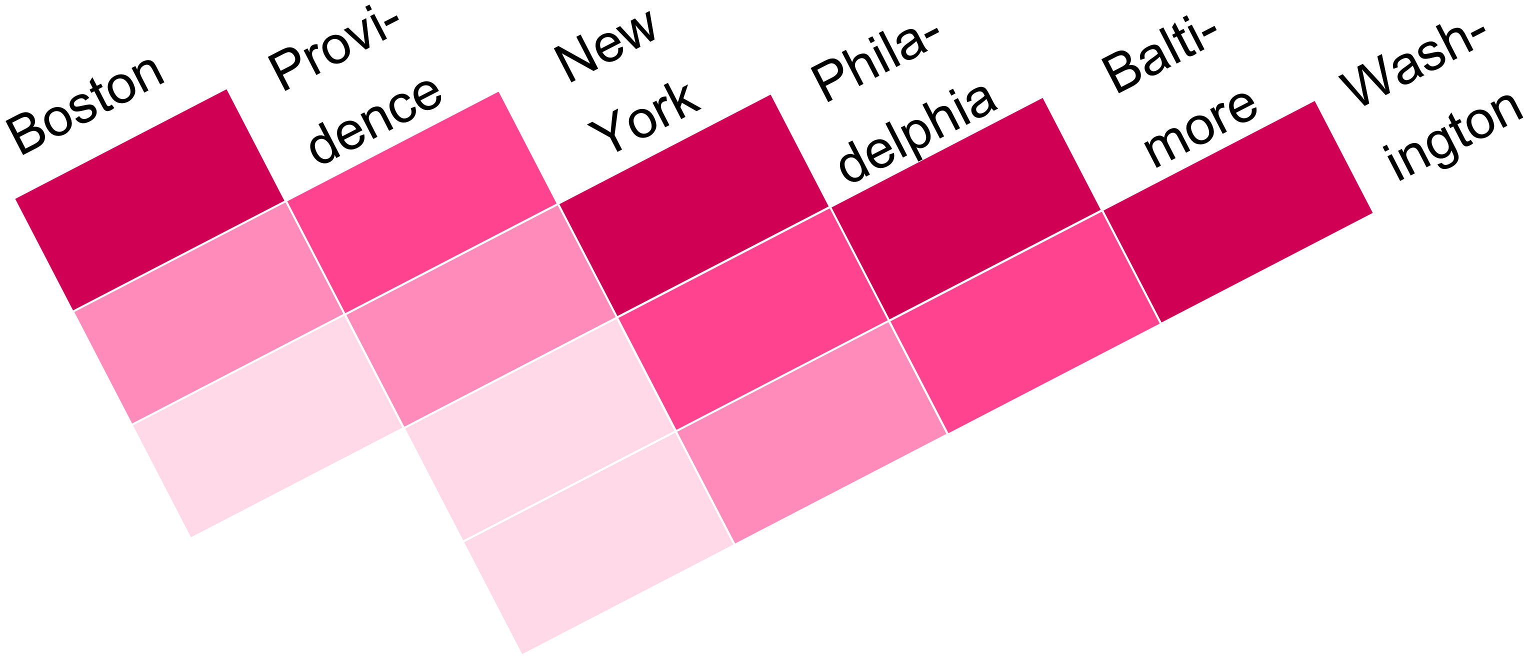
(Samani et al 2007))

- These plots can be generated using the free software “*Haploview*”, but also in R!

Distances among cities

	Boston	Providence	New York	Philadelphia	Baltimore
Providence	59				
New York	210	152			
Philadelphia	320	237	86		
Baltimore	430	325	173	87	
Washington	450	358	206	120	34

Distances among cities



Distances among SNPs

- If a causal polymorphism is not genotyped, we can still hope to detect its effects through **Linkage Disequilibrium** (LD) with polymorphisms that are typed (key principle behind doing genetic association analysis …).
- LD is a measure of co-segregation of alleles in a population: Two alleles at different loci that occur together on the same chromosome (or gamete) more often than would be predicted by random chance. In general, LD is taken to be a measure of allelic association.
- Among the measures that have been proposed for two-locus haplotype data, the two most important are *D*' (Lewontin's *D* prime) and r^2 (the square correlation coefficient between the two loci under study).

- Sample size must be increased by a factor of $1/r^2$ to detect an unmeasured variant, compared with the sample size for testing the variant itself.

(Jorgenson and Witte 2006)

Distances among SNPs

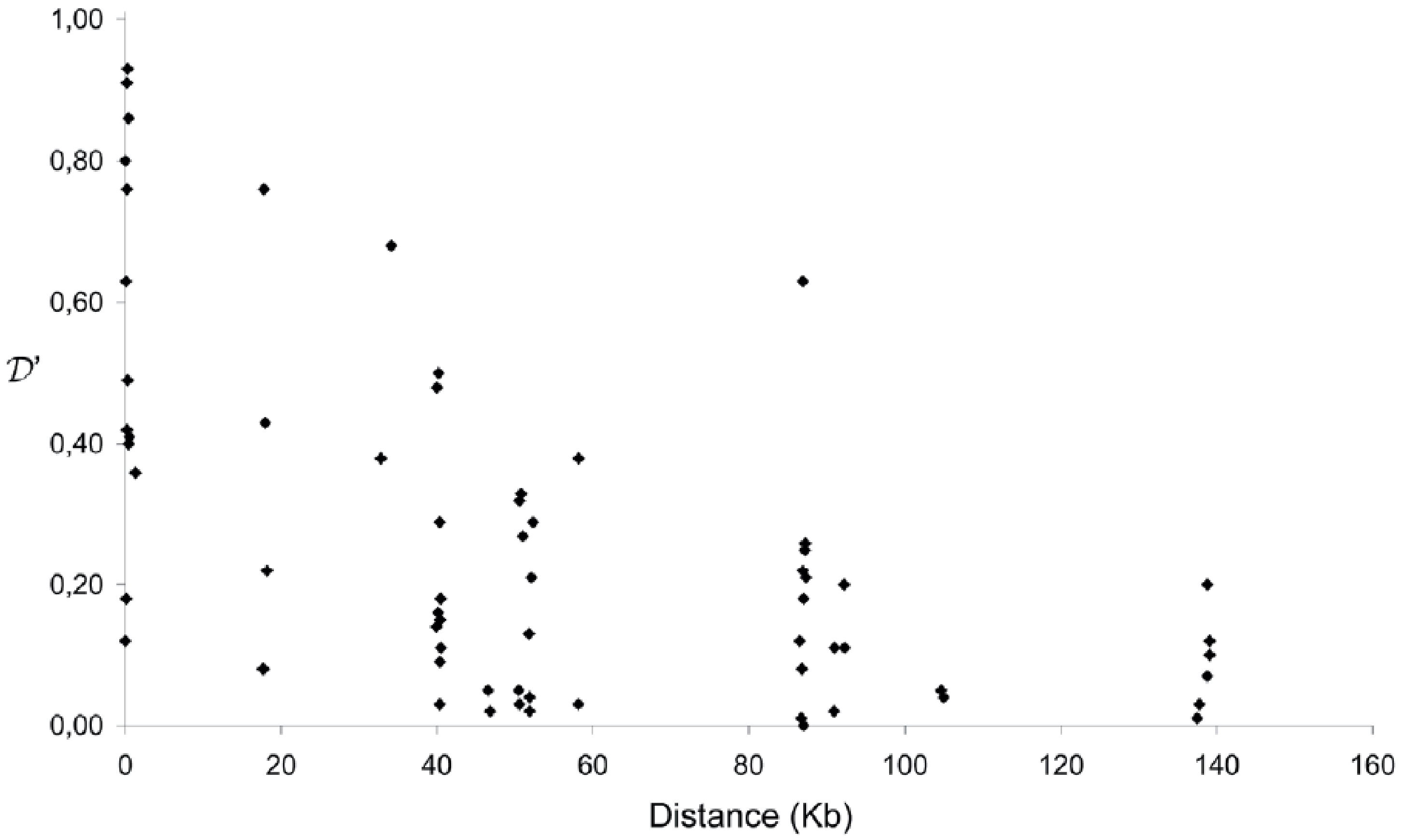
- The measure D is defined as the difference between the observed and expected (under the null hypothesis of independence) proportion of haplotypes bearing specific alleles at two loci: $p_{AB} - p_A p_B$

	A	*a*
B	p_{AB}	p_{aB}
b	p_{Ab}	p_{ab}

  - D' is the absolute ratio of D compared with its maximum value.
  - D' =1 : complete LD
- R^2 is the statistical correlation of two markers :
  - When R^2=1, knowing the genotypes of alleles of one SNP is directly predictive of genotype of another SNP

$$R^2 = \frac{D^2}{P(A)P(a)P(B)P(b)}$$

How far does linkage disequilibrium extend?



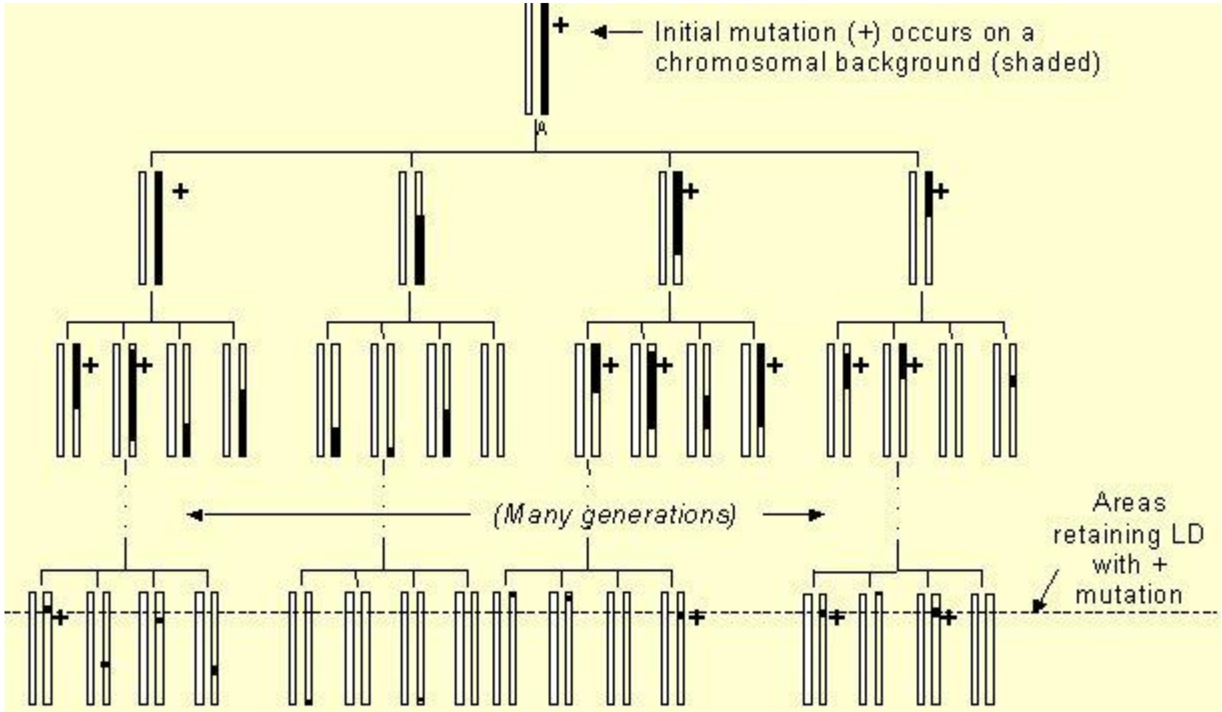
(Hecker et al 2003)

How to interpret LD data?

- The patterns of LD observed in natural populations are the result of a complex interplay between genetic factors and the population's demographic history (Pritchard, 2001).
- LD is usually a function of distance between the two loci. This is mainly because recombination acts to break down LD in successive generations (Hill, 1966).
- When a mutation first occurs it is in complete LD with the nearest marker (D' = 1.0). Given enough time and as a function of the distance between the mutation and the marker, LD tends to decay and in complete equilibrium reached D' = 0 value. Thus, it decreases at every generation of random mating unless some process is opposing to the approach to linkage 'equilibrium'.

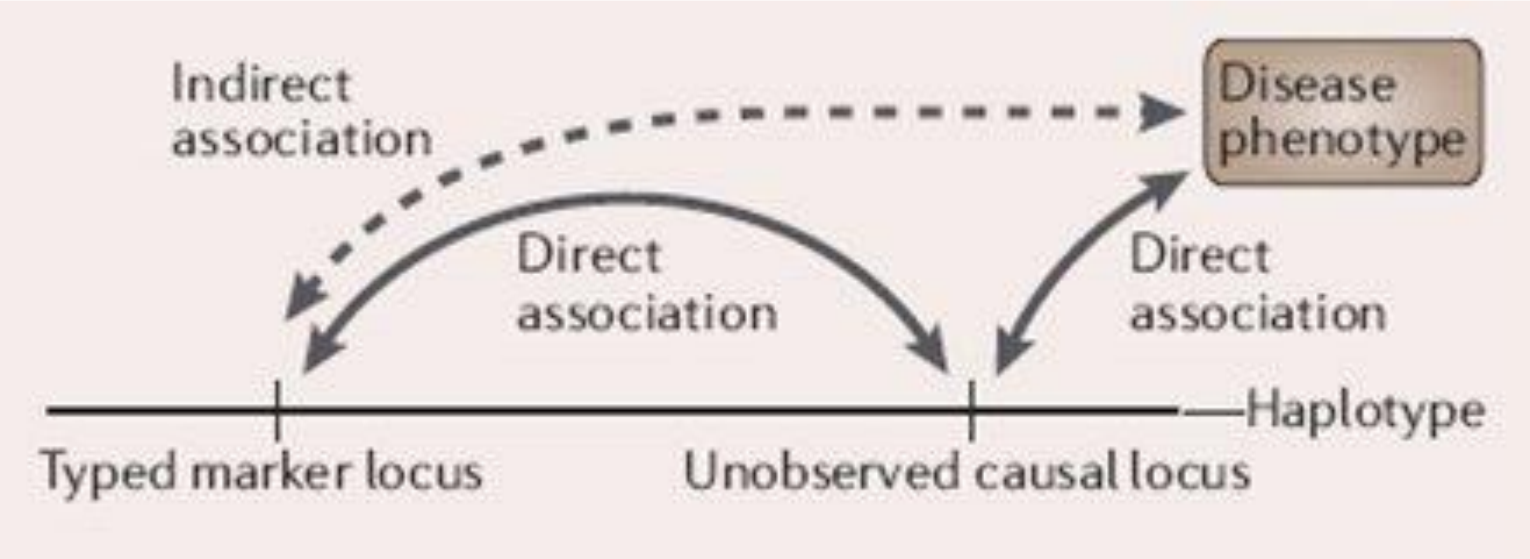
How to interpret LD data?

- Therefore, the key concept in a (population-based) genetic association study is linkage disequilibrium.


Initial mutation (+) occurs on a chromosomal background (shaded)
+
(Many generations)
Areas retaining LD with + mutation

How to interpret LD data?

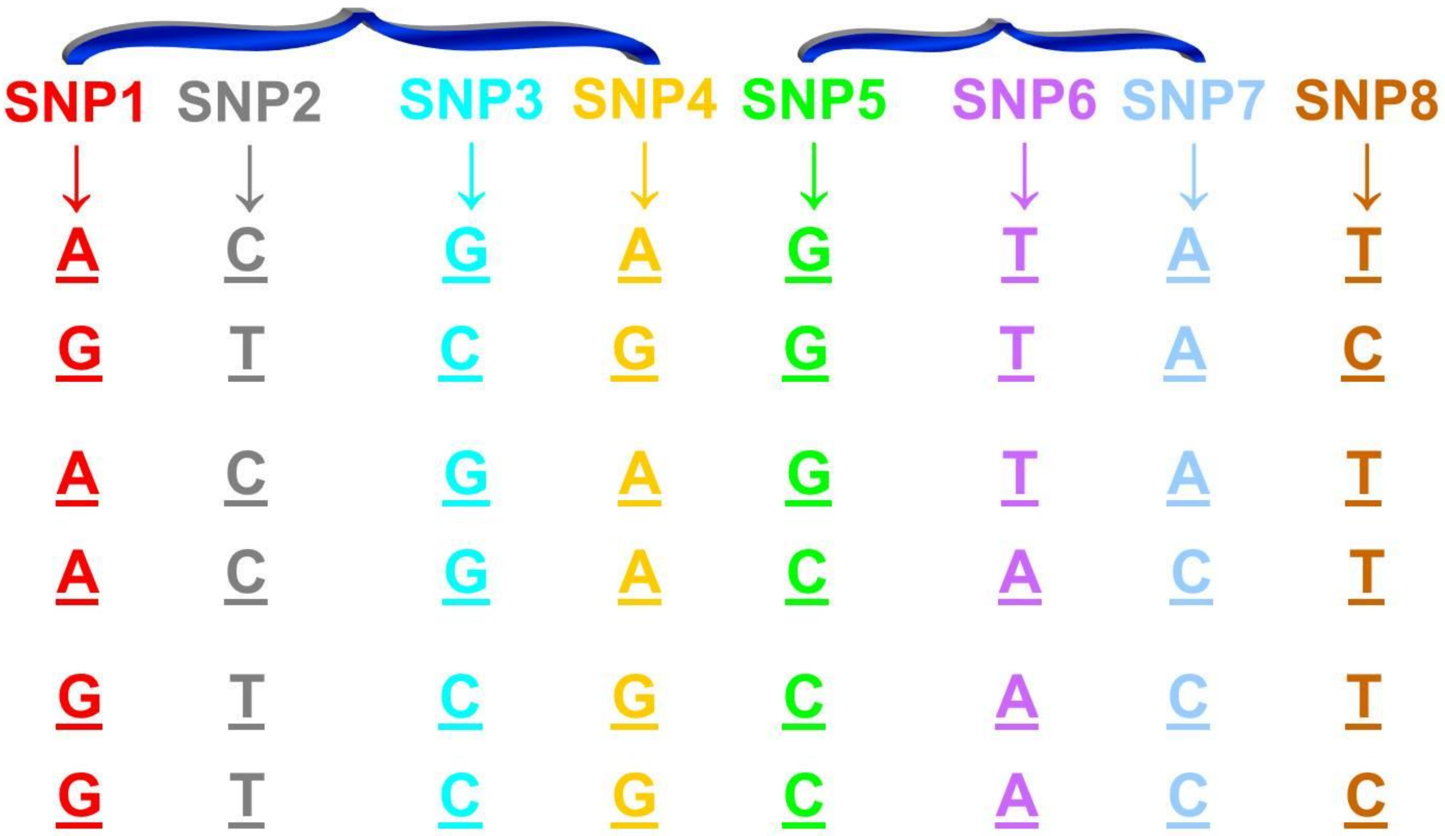
- It gives the rational for performing genetic association studies



Phenotype: The visible or measurable (expressed) characteristic of an organism

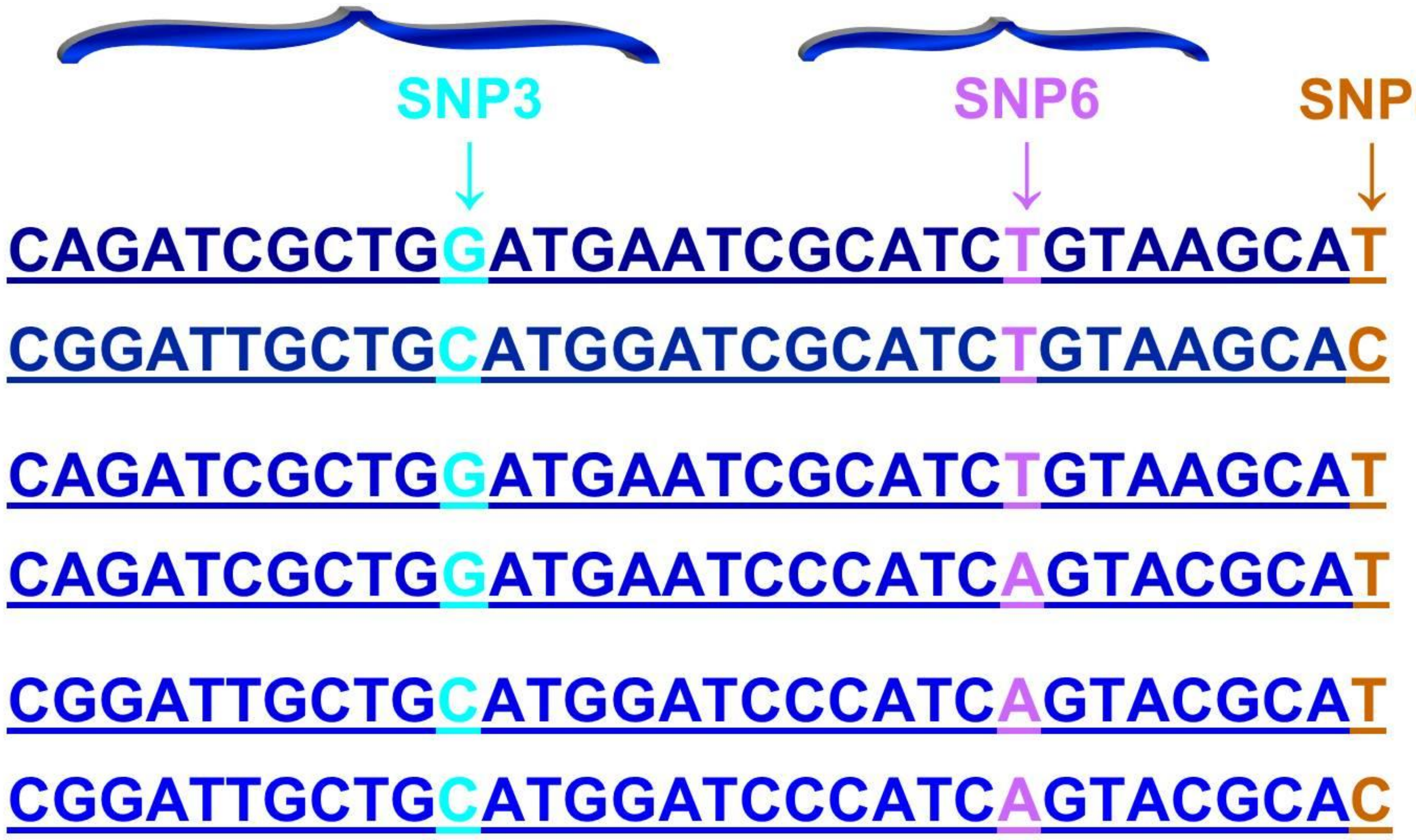
Trait: Coded phenotype

How can one tag SNP serve as proxy for many? (adapted from Manolio 2010)



SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8
A	C	G	A	G	T	A	T
G	T	C	G	G	T	A	C
A	C	G	A	G	T	A	T
A	C	G	A	C	A	C	T
G	T	C	G	C	A	C	T
G	T	C	G	C	A	C	C

How can one tag SNP serve as proxy for many? (adapted from Manolio 2010)



CAGATCGCTGGATGAATCGCATCTGTAAGCAT

CGGATTGCTGCATGGATCGCATCTGTAAGCAC

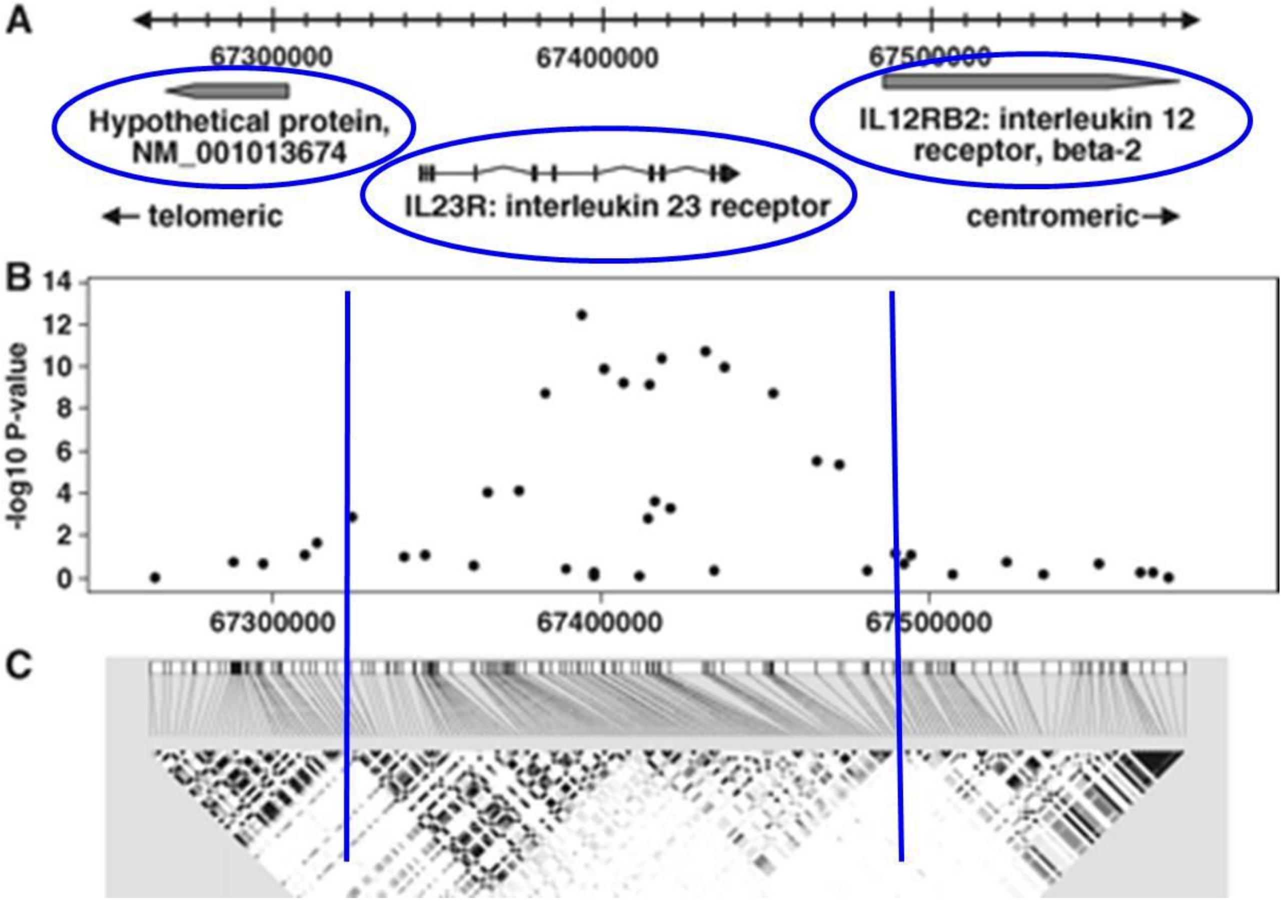
CAGATCGCTGGATGAATCGCATCTGTAAGCAT

CAGATCGCTGGATGAATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAC

Where is the true causal variant?

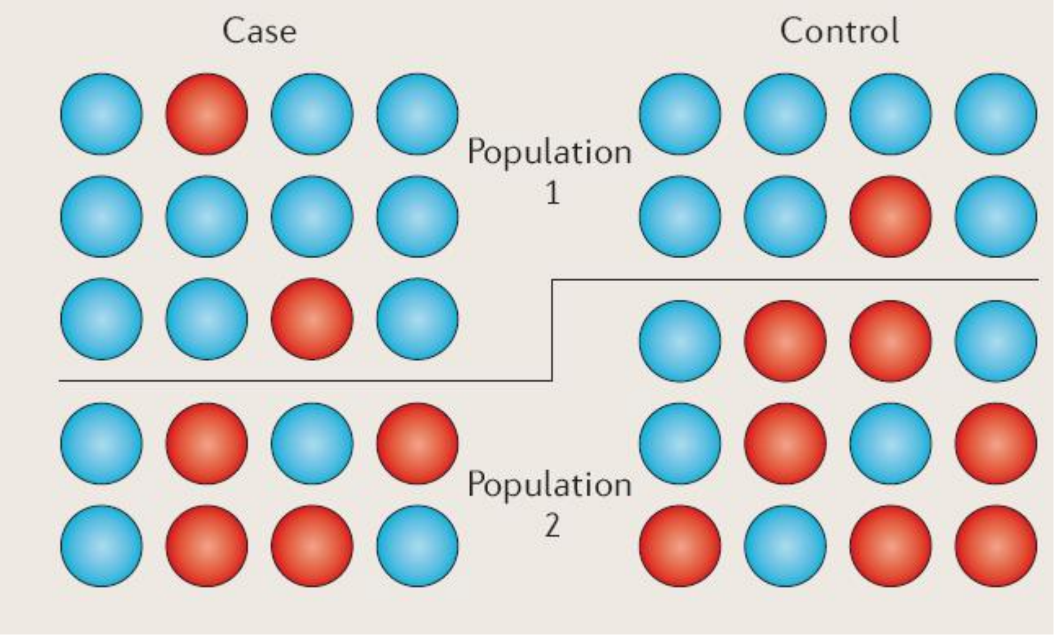


(Duerr et al 2006)

3.c Confounding

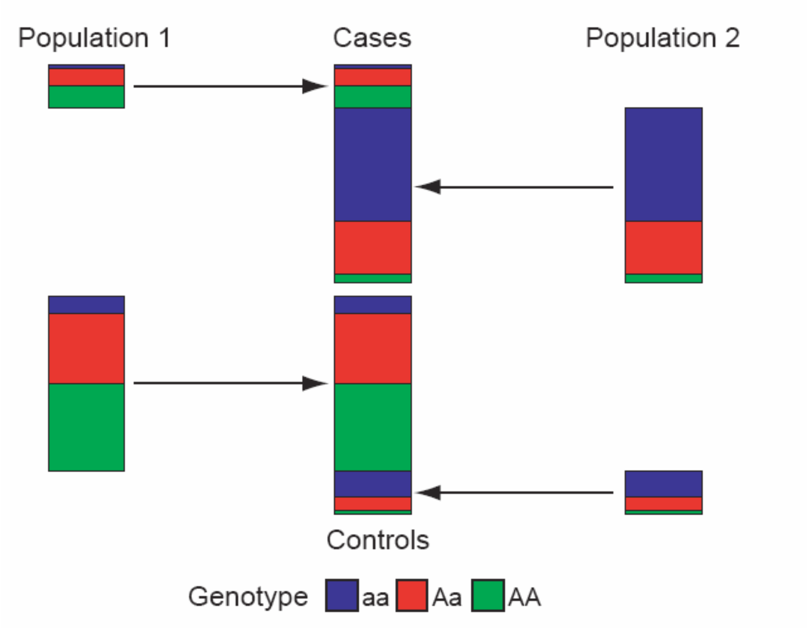
What is spurious association?

- Spurious association refers to false positive association results due to not having accounted for population substructure as a confounding factor in the analysis



What is spurious association?

- Typically, there are two characteristics present:
  - A difference in proportion of individual from two (or more) subpopulation in case and controls
  - Subpopulations have different allele frequencies at the locus.

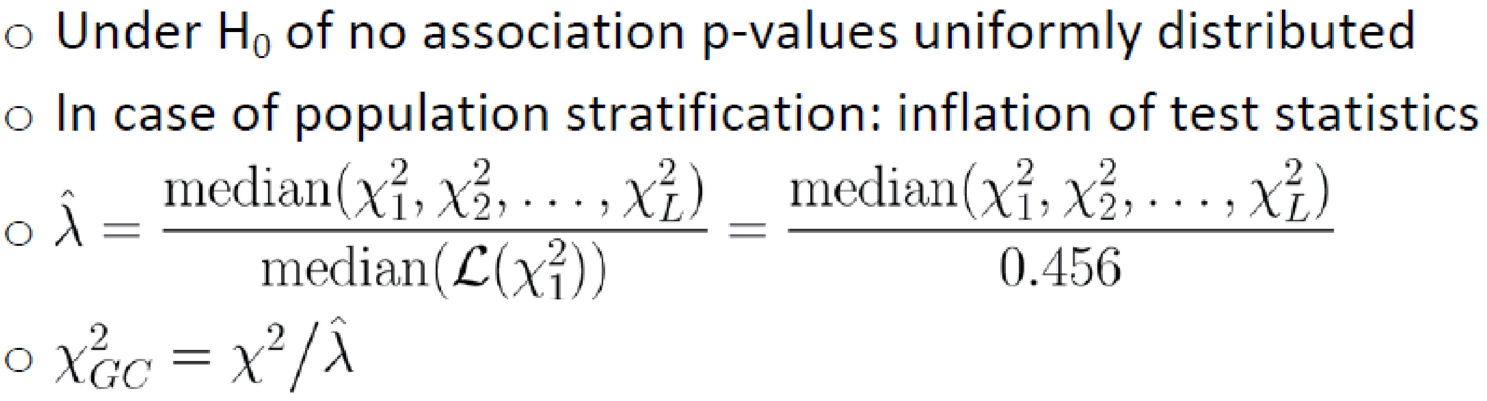


What are typical methods to deal with population stratification?

- Methods to deal with spurious associations generated by population structure generally require a number (at least >100) of widely spaced null SNPs that have been genotyped in cases and controls in addition to the candidate SNPs.
- These methods large group into:
  - Genomic control methods
  - Structured association methdos
  - Principal component-based methods

What is genomic control?

- In Genomic Control (GC), a 1-df association test statistic is computed at each of the null SNPs, and a parameter λ is calculated as the empirical median divided by its expectation under the chi-squared 1-df distribution.
- Then the association test is applied at the candidate SNPs, and if λ > 1 the test statistics are divided by λ.



- ○ Under H_0 of no association p-values uniformly distributed
- ○ In case of population stratification: inflation of test statistics
- ○ $\hat{\lambda} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{\text{median}(\mathcal{L}(\chi_1^2))} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{0.456}$
- ○ $\chi_{GC}^2 = \chi^2 / \hat{\lambda}$

What is genomic control?

- The motivation for GC is that, as we expect few if any of the null SNPs to be associated with the phenotype, a value of λ > 1 is likely to be due to the effect of population stratification, and dividing by λ cancels this effect for the candidate SNPs.
- GC performs well under many scenarios, but can be conservative in extreme settings (and anti-conservative if insufficient null SNPs are used).
- There is an analogous procedure for a general (2 df) test; The method can also be applied to other testing approaches.

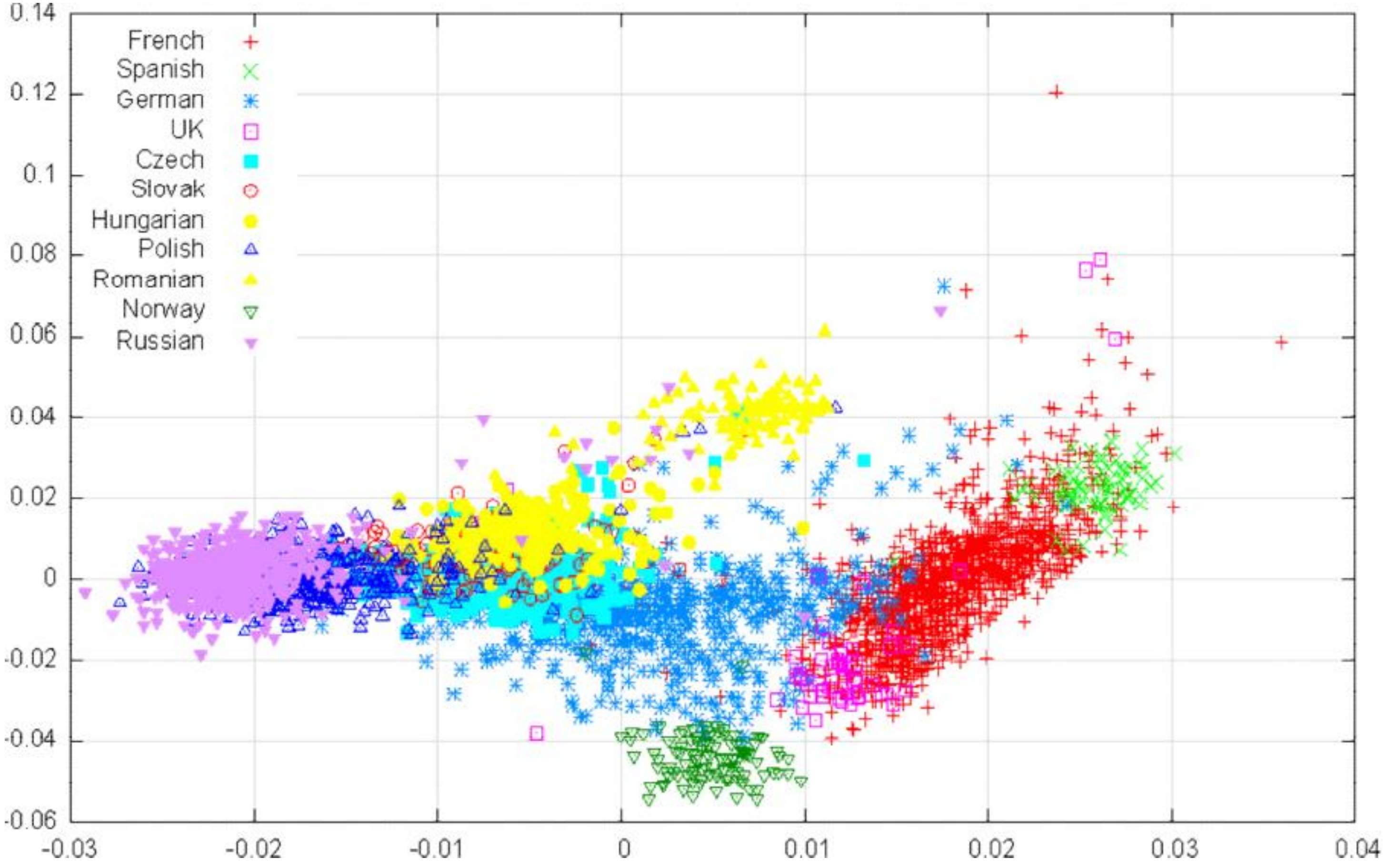
What is a structured association method?

- Structured association (SA) approaches are based on the idea of attributing the genomes of study individuals to hypothetical subpopulations, and testing for association that is conditional on this subpopulation allocation.
- Several clustering algorithms exist to estimate the number of subpopulations.
- These approaches (such as Bayesian clustering approaches) are computationally demanding, and because the notion of subpopulation is a theoretical construct that only imperfectly reflects reality, the question of the correct number of subpopulations can never be fully resolved….

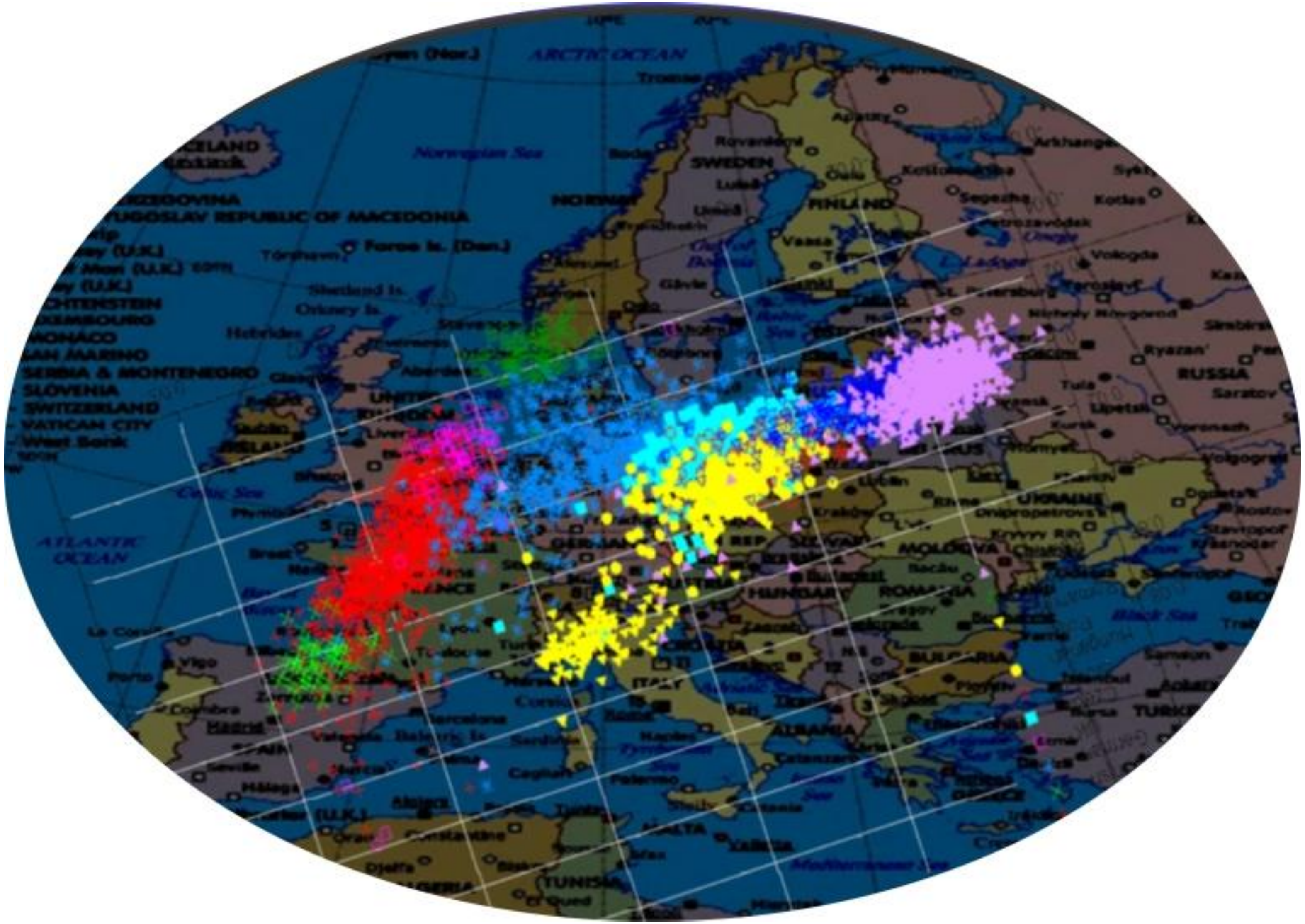
What is principal components analysis?

- When many null markers are available, principal components analysis provides a fast and effective way to diagnose population structure.
- Principal components are linear combinations of the original “variables” (here SNPs) that optimized in such a way that as much of the variation in the data is retained.

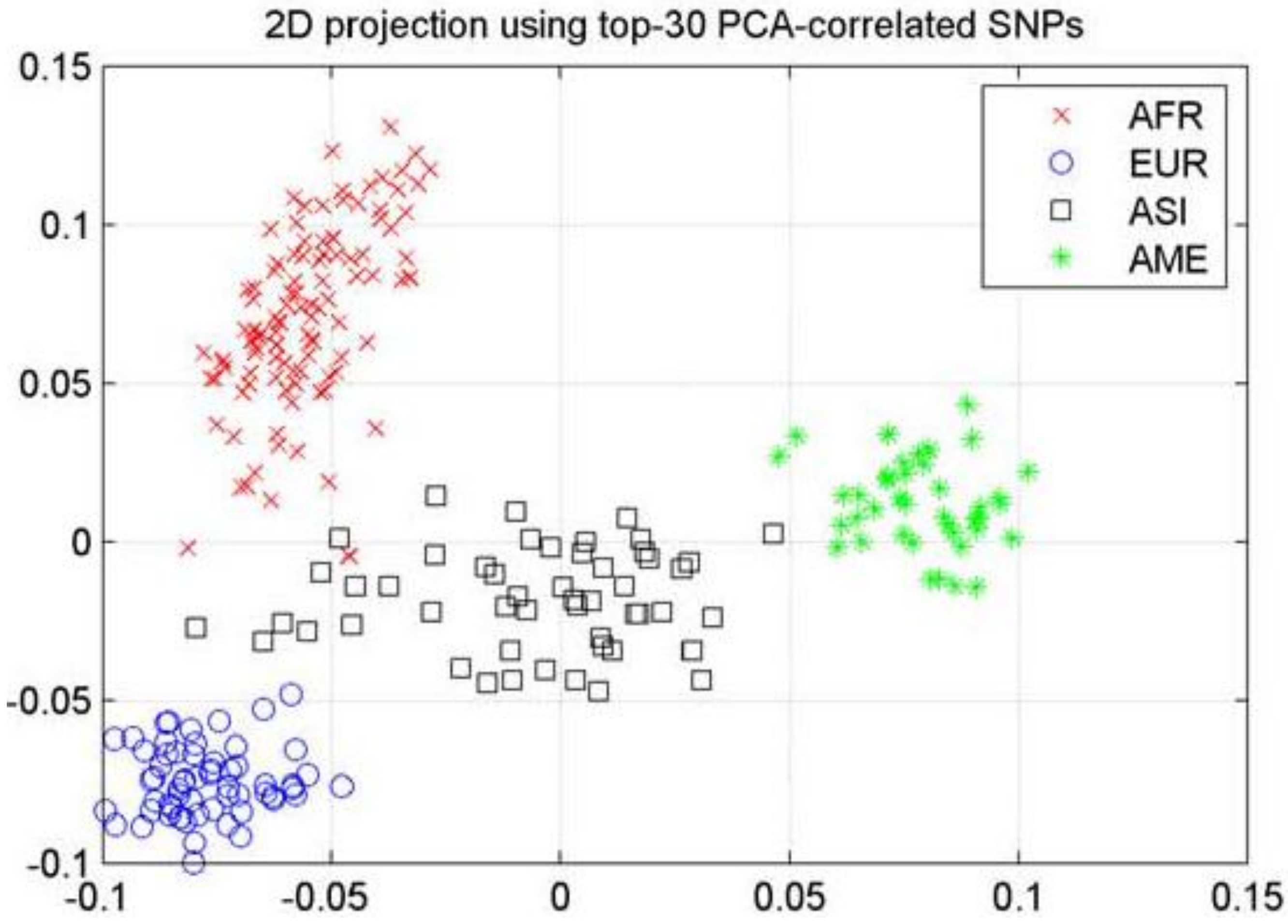
- In European data, the first 2 principal components "nicely" reflect the N-S and E-W axes !



Y-axis: PC2 (6% of variance); X-axis: PC1 (26% of variance)


ARCTIC OCEAN
Norwegian Sea
ICELAND
SWEDEN
FINLAND
NORWAY
Faroe Is. (Den.)
Shetland Is.
Orkney Is.
Hebrides
RUSSIA
Saratov
Ryazan'
Tula
Lipetsk
Voronezh
Arkhangel'sk
Vologda
UKRAINE
Dnipropetrovs'k
MOLDOVA
ROMANIA
BULGARIA
HUNGARY
ALBANIA
ITALY
TUNISIA
TURKEY
Black Sea
ATLANTIC OCEAN
Celtic Sea
Brest
Toulouse
Barcelona
Balearic Is.
Valencia
Madrid
Coimbra
Porto
Vigo
Sevilla
Málaga
Murcia
Algiers
Constantine
Djelfa
Corsica
Sardinia
Naples
Palermo
Catanzaro
MALTA
Lublin
Kraków
Bacău
Odesa
Simferopol'
Rostov
Volgograd
Samsun
Izmir
Vaasa
Gävle
Umeå
Luleå
Tromsø
Bodø
Rovaniemi
Oulu
Kotlas
Segezha
Petrozavodsk
Onega
L. Ladoga
Gulf of Bothnia
Baltic Sea
Aberdeen
Glasgow
Tórshavn
Trondheim
Oslo
Stavanger

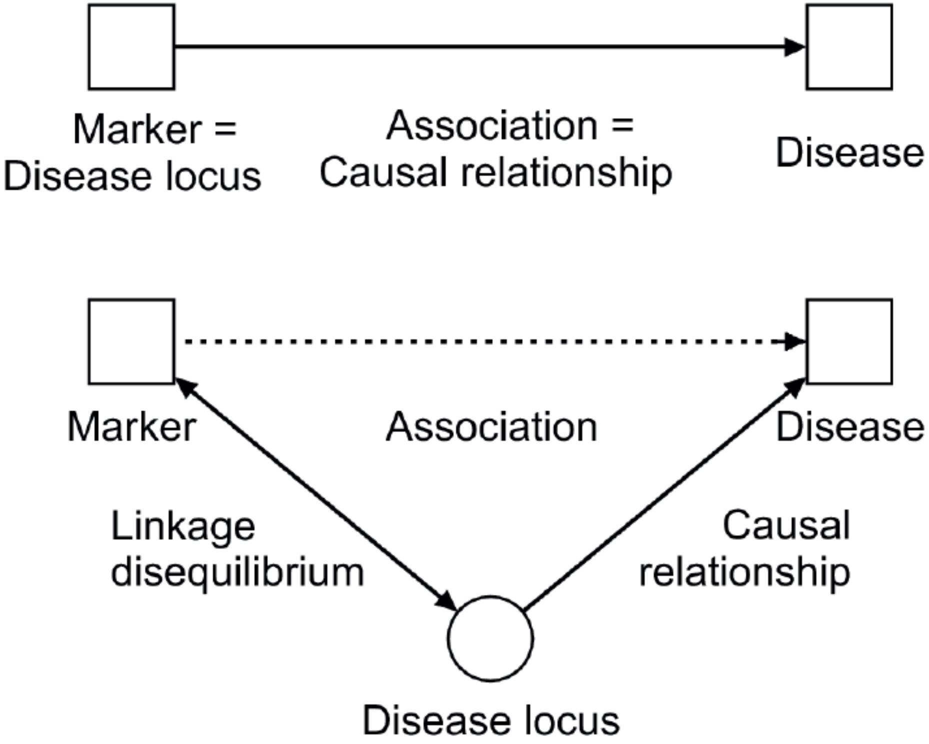
- Does the same hold on a “global” (world) level?



(Paschau 2007)

4 Tests of association

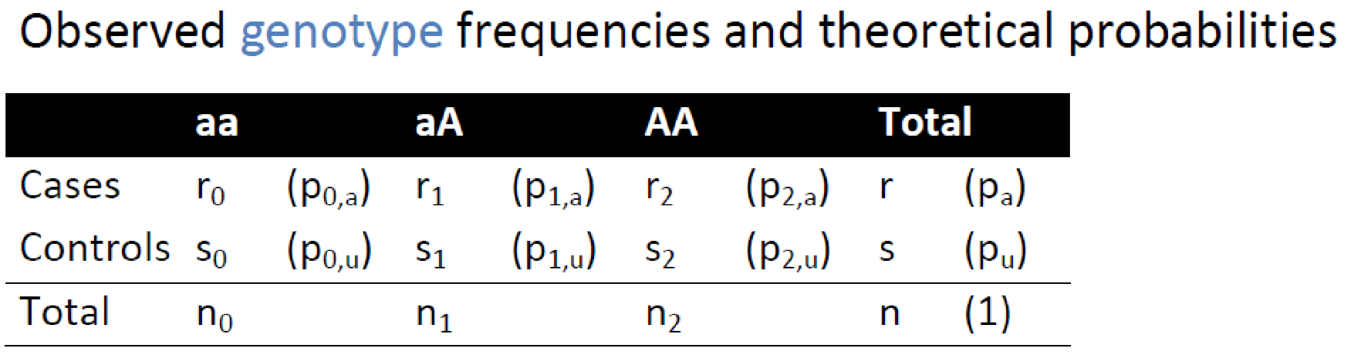
What is the causal model underlying genetic association?



(Ziegler and Van Steen 2010)

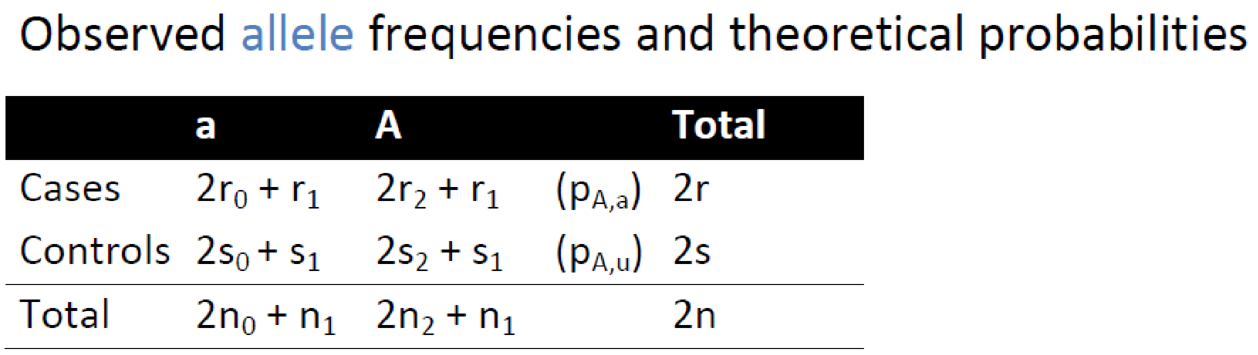
4.a Single SNP

What are common association tests (dichotomous traits)?



Observed genotype frequencies and theoretical probabilities

	aa		aA		AA		Total	
Cases	r_0	$(p_{0,a})$	r_1	$(p_{1,a})$	r_2	$(p_{2,a})$	r	(p_a)
Controls	s_0	$(p_{0,u})$	s_1	$(p_{1,u})$	s_2	$(p_{2,u})$	s	(p_u)
Total	n_0		n_1		n_2		n	(1)



Observed allele frequencies and theoretical probabilities

	a	A		Total
Cases	$2r_0 + r_1$	$2r_2 + r_1$	$(p_{A,a})$	$2r$
Controls	$2s_0 + s_1$	$2s_2 + s_1$	$(p_{A,u})$	$2s$
Total	$2n_0 + n_1$	$2n_2 + n_1$		$2n$

(Ziegler and Van Steen 2010)

What are common association tests (dichotomous traits)?

Standard allele test:

- χ^2 test of independence
- Equivalent to

$$\chi_A^2 = 2n \cdot \frac{[(2r_0 + r_1)(2s_2 + s_1) - (2r_2 + r_1)(2s_0 + s_1)]^2}{2r \cdot 2s \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)}$$

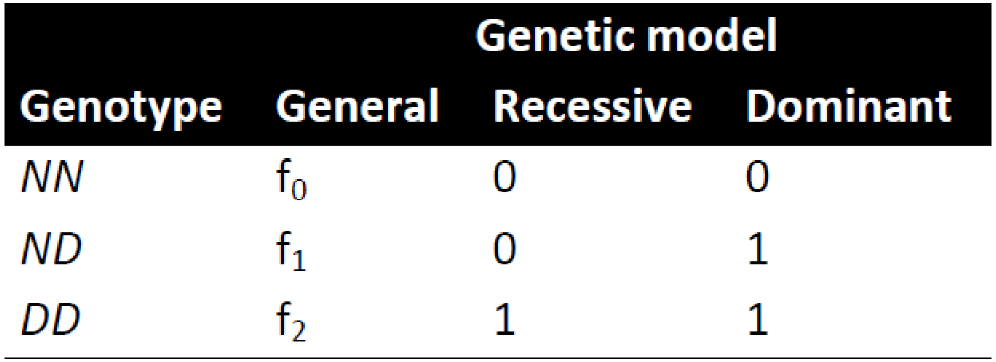
- Asymptotically χ^2 with 1 degree of freedom (d.f.)

Standard genotype test:

- χ^2 test of independence
- Asymptotically χ^2 with 2 d.f.

(Ziegler and Van Steen 2010)

What are common association tests (dichotomous traits)?



	Genetic model		
Genotype	General	Recessive	Dominant
NN	f_0	0	0
ND	f_1	0	1
DD	f_2	1	1

Penetrances for simple Mendelian inheritance patterns

- **Trait T:** coded phenotype
- **Penetrance:** P(T|Genotype)
- **Complete penetrance**: P(T|DD) = 1 (simplified definition)

What are common association tests (dichotomous traits)?

	Dominant		Heterozygote		Recessvie	
	aa	aA or AA	aa or AA	aA	aa or aA	AA
Cases	r_0	$r_1 + r_2$	$r_0 + r_2$	r_1	$r_0 + r_1$	r_2
Controls	s_0	$s_1 + s_2$	$s_0 + s_2$	s_1	$s_0 + s_1$	s_2
Total	n_0	n_2	$n_0 + n_2$	n_1	$n_0 + n_1$	n_2

- $\chi^2_{dom} = n \cdot \dfrac{\big(r_0(s_1+s_2) - (r_1+r_2)s_0\big)^2}{r \cdot s \cdot n_0 \cdot (n_1+n_2)}$
- $\chi^2_{het} = n \cdot \dfrac{\big(r_1(s_0+s_2) - (r_0+r_2)s_1\big)^2}{r \cdot s \cdot n_1 \cdot (n_0+n_2)}$
- $\chi^2_{rec} = n \cdot \dfrac{\big((r_0+r_1)s_2 - r_2(s_0+s_1)\big)^2}{r \cdot s \cdot (n_0+n_1) \cdot n_2}$

What are common association tests (dichotomous traits)?

- The Cochran-Armitage trend test measures a linear trend in proportions weighted by general measure of exposure dosage: variable x in regression model =#alleles

$$\chi^2_{trend} = \frac{n}{rs} \cdot \frac{\left(2r_2 s - 2r s_2 + r_1 s - s_1 r\right)^2}{2n_2 n + (2n_2 + n_1)(n_0 - n_2)}$$

- Max test: computes maximum over standardized tests for different genetic models, providing a global test

Which test should be used in applications?

- Trend test if no biological hypothesis
- Trend test optimal if additive genetic model
- Dom test optimal if dominant genetic model
- Rec test optimal if recessive genetic model
- Trend test identical to allele test if HWE exactly fulfilled
- Asymptotic version of Max test alternative to trend test

Sasieni 1997 Biometrics, Zou 2006 Ann Hum Genet, Guedj et al. 2008 Ann Hum Genet
Hothorn & Hothorn 2009 Biom J

How are genetic effects measured?

	G = 1	G = 0	Total
Cases	x_1	y_1	n_1
Controls	x_0	y_0	n_0

Case control study:

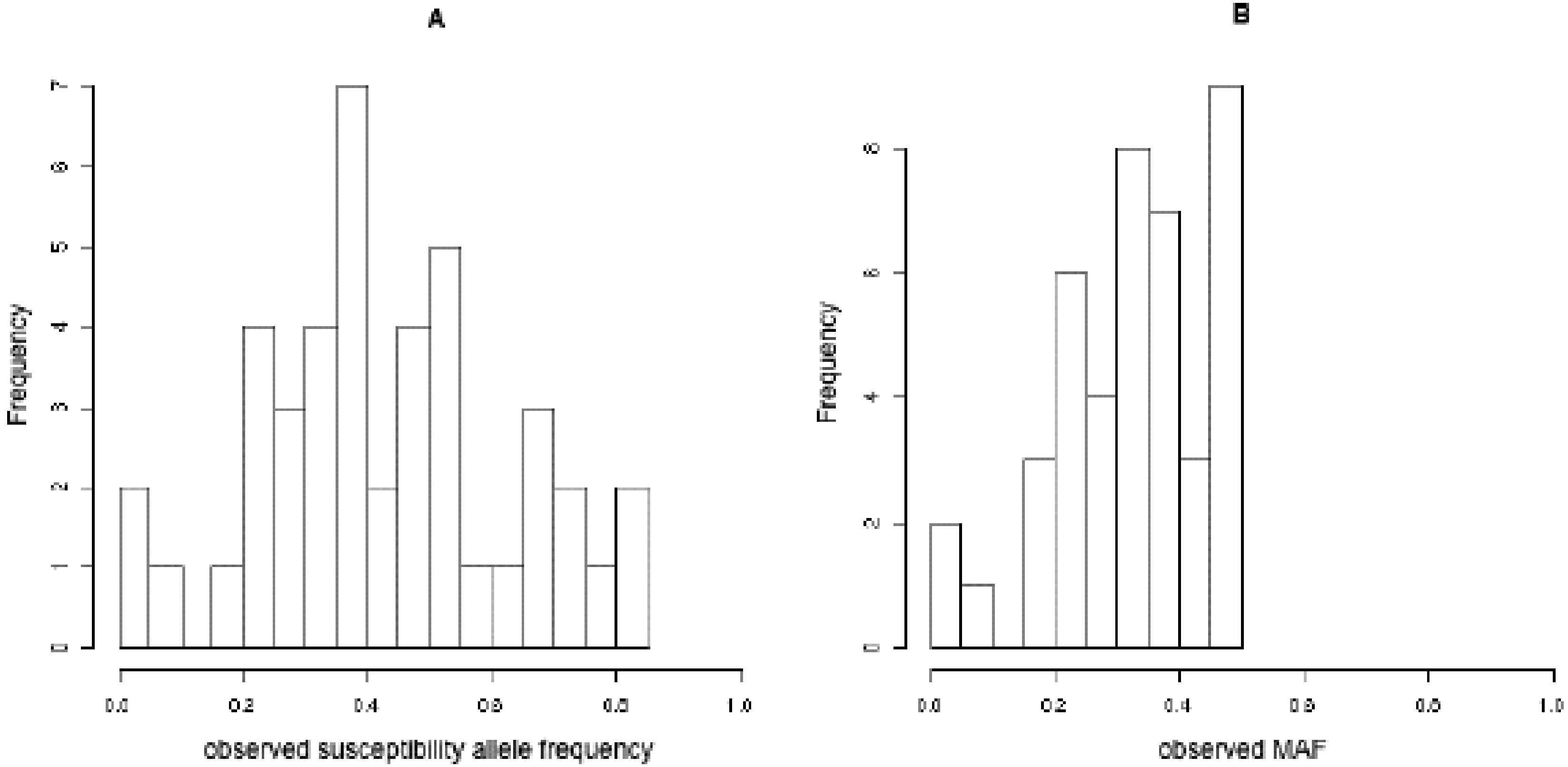
- Odds ratio: $\widehat{OR}_G = \frac{x_1 y_0}{y_1 x_0}$
- Attributable risk in variant carriers (… in the exposed):

$$\mathrm{AR}_G = \frac{P(\mathrm{aff}|G=1) - P(\mathrm{aff}|G=0)}{P(\mathrm{aff}|G=1)} = \frac{p_1 - p_0}{p_1} = \frac{RR-1}{RR} \approx \frac{RR-1}{RR}$$

RR being

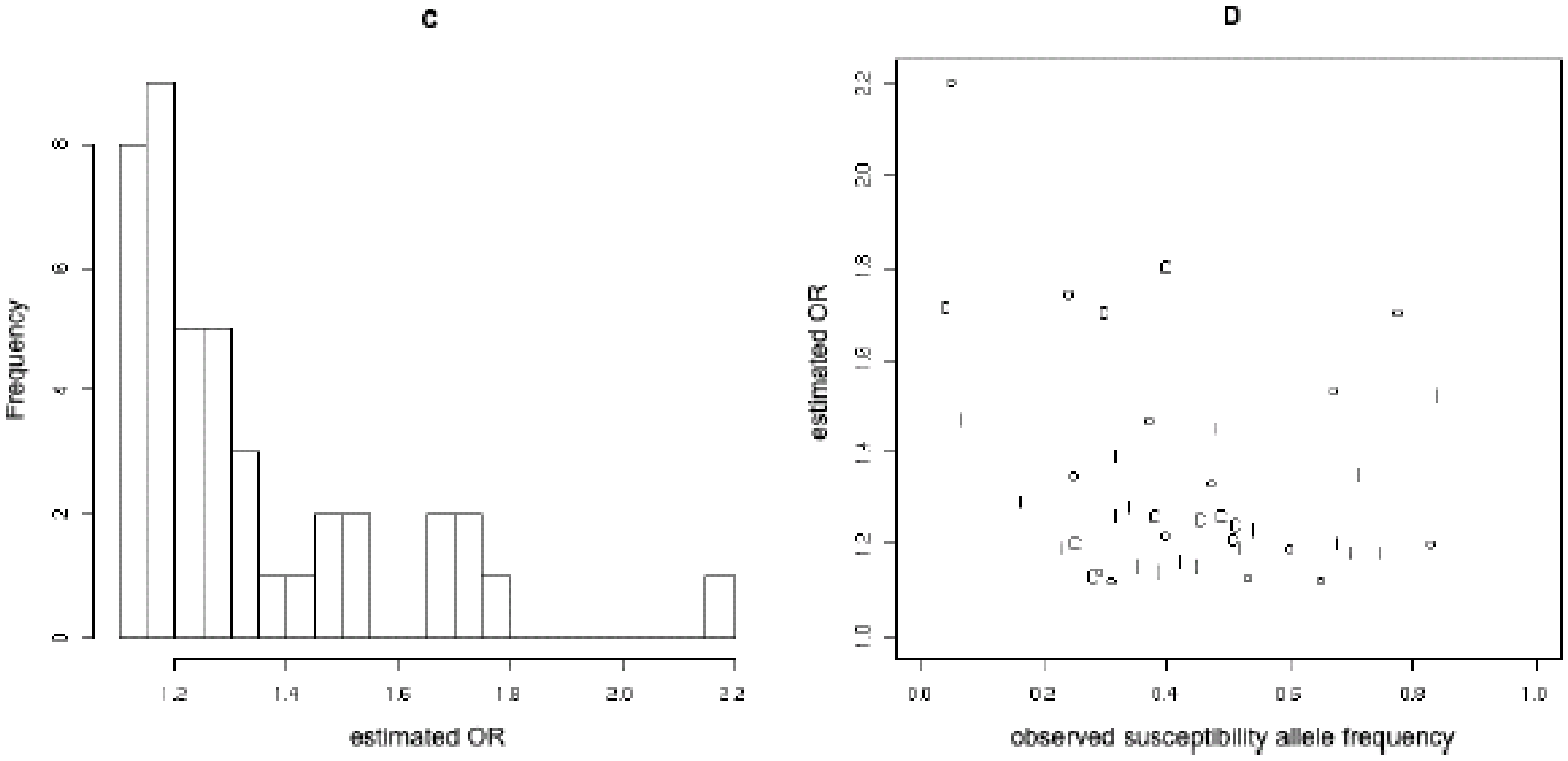
$$\frac{P(\mathrm{aff}|G=1)}{P(\mathrm{aff}|G=0)} = \frac{p_1}{p_0}$$

Which odds ratios (measures of effect) can we expect?



(A and B) Histograms of susceptibility allele frequency and MAF, respectively, at confirmed susceptibility loci. . (Iles 2008)

Which odds ratios (measures of effect) can we expect?



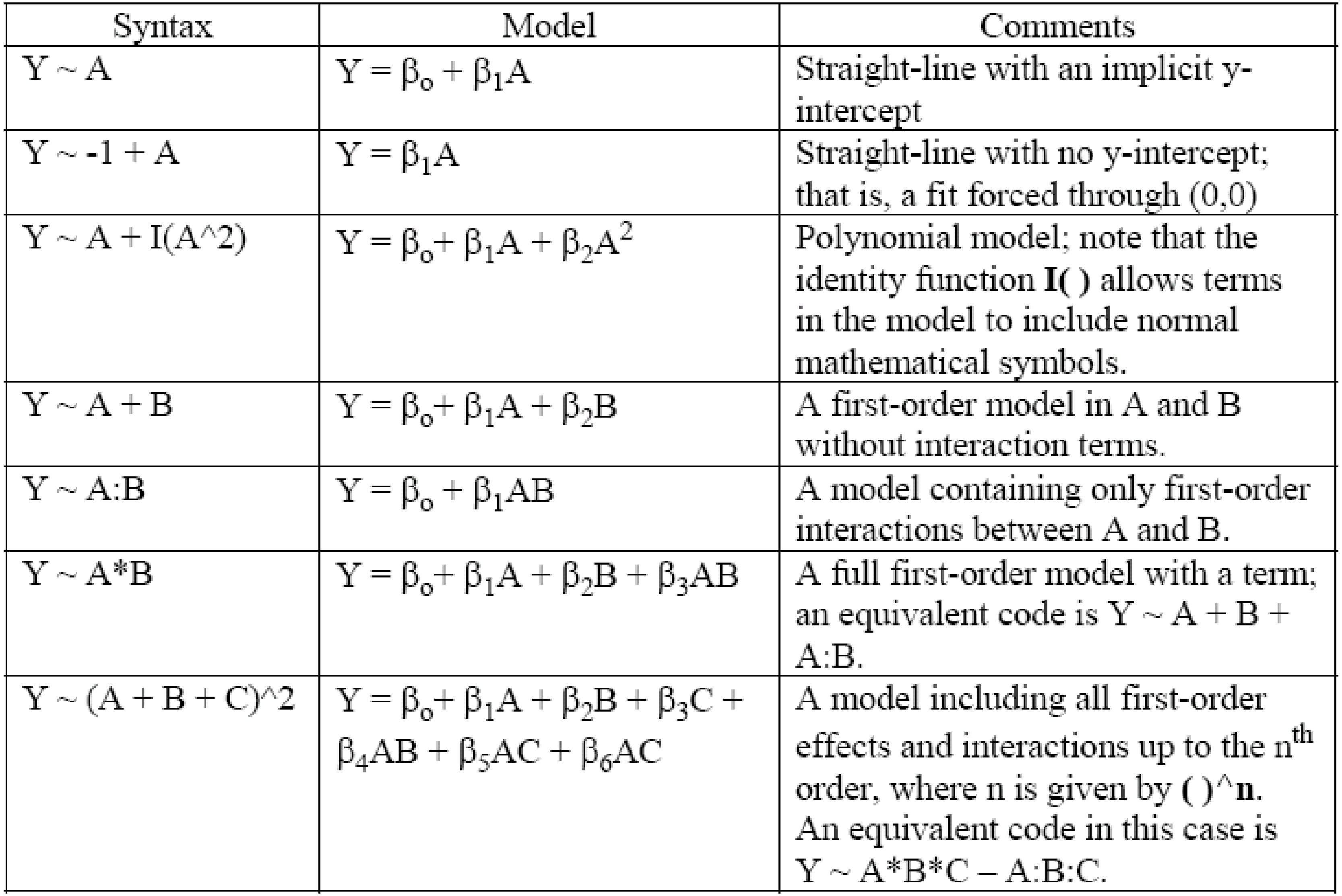
doi:10.1371/journal.pgen.0040033.g001

(C) Histogram of estimated ORs (estimate of genetic effect size) at confirmed susceptibility loci. (Iles 2008)

4.b Repeated single SNP tests

The regression framework

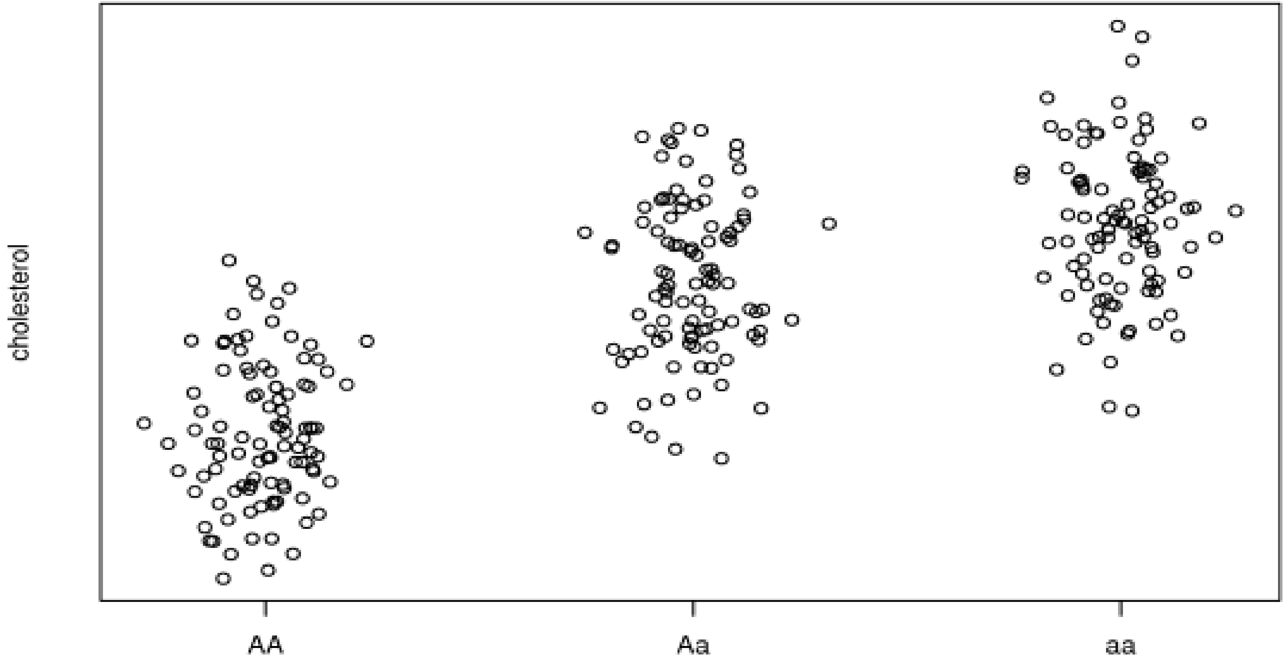
- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, X_1, …, X_m.
- When m=1 it is called simple regression but when m > 1 it is called multiple regression or sometimes multivariate regression.
- When there is more than one Y, then it is called multivariate multiple regression
- The basic syntax for doing regression in R is lm(Y~model) to fit linear models and glm() to fit generalized linear models (e.g. logistic regression models in the “dichotomous trait” setting before). Next slide: syntax !



Syntax	Model	Comments
Y ~ A	$Y = \beta_o + \beta_1 A$	Straight-line with an implicit y-intercept
Y ~ -1 + A	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
Y ~ A + I(A^2)	$Y = \beta_o + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function **I( )** allows terms in the model to include normal mathematical symbols.
Y ~ A + B	$Y = \beta_o + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
Y ~ A:B	$Y = \beta_o + \beta_1 AB$	A model containing only first-order interactions between A and B.
Y ~ A*B	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is Y ~ A + B + A:B.
Y ~ (A + B + C)^2	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 AC$	A model including all first-order effects and interactions up to the nth order, where n is given by **( )^n**. An equivalent code in this case is Y ~ A*B*C – A:B:C.

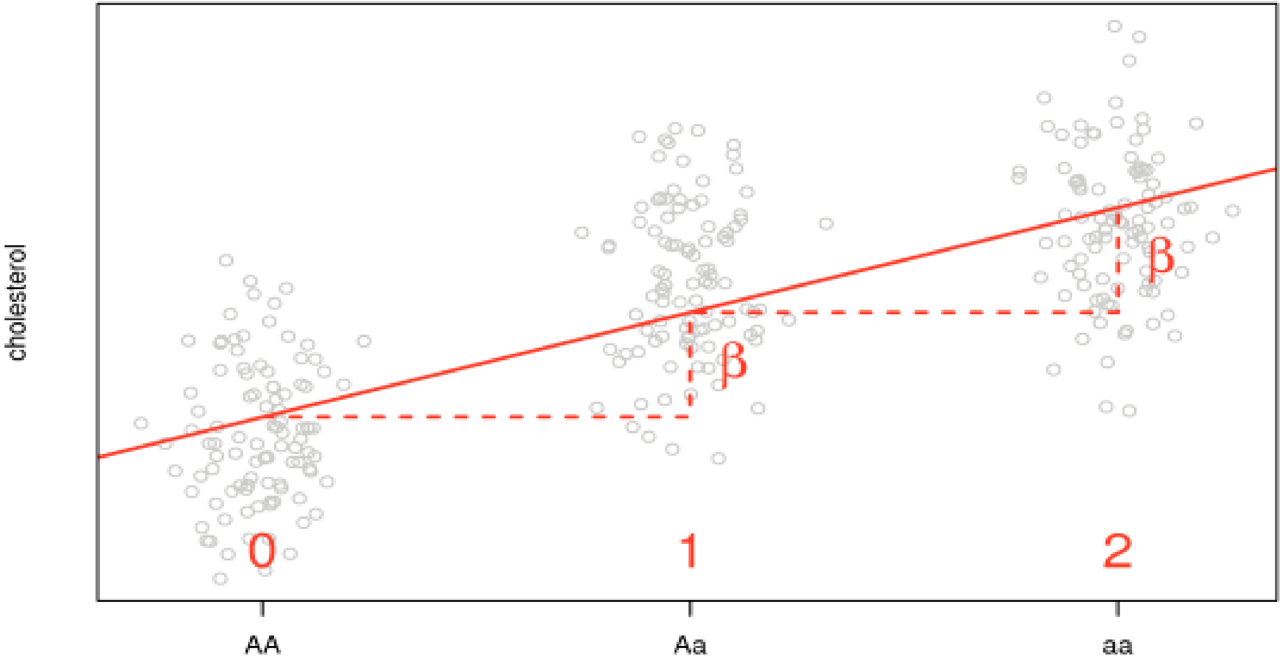
Use of `lm()` in genetics

Some data; cholesterol levels plotted by genotype (single SNP)



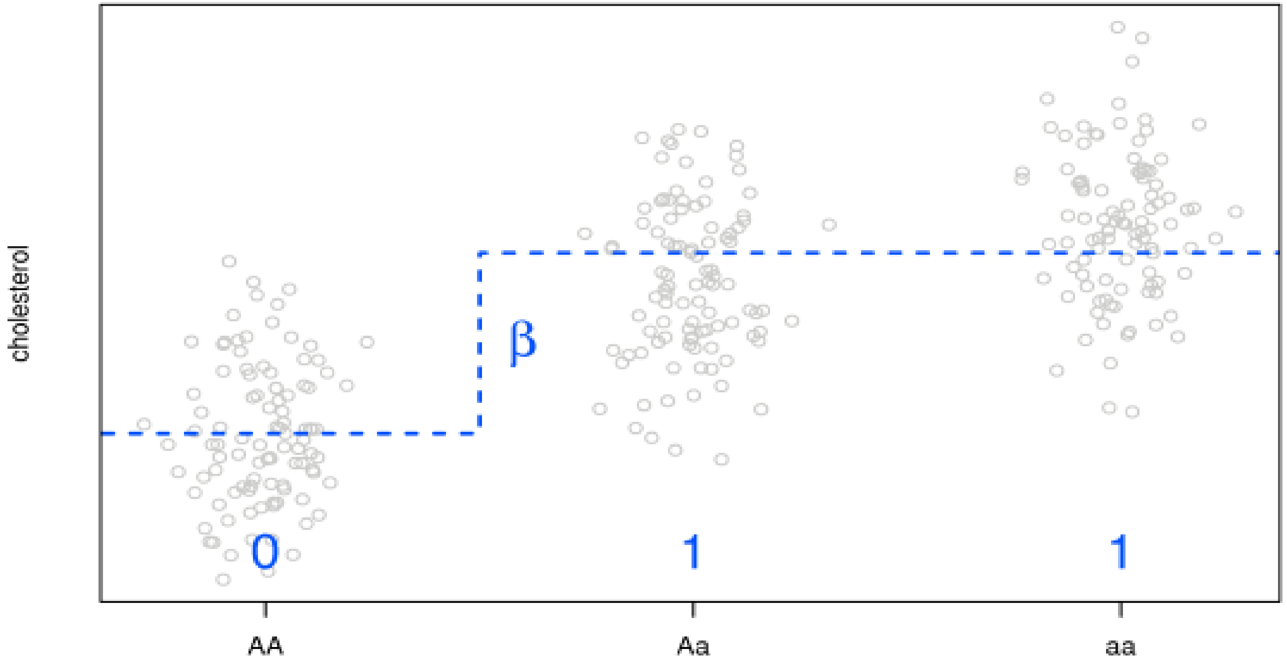
Use of `lm()` in genetics

Additive model (the most commonly used)



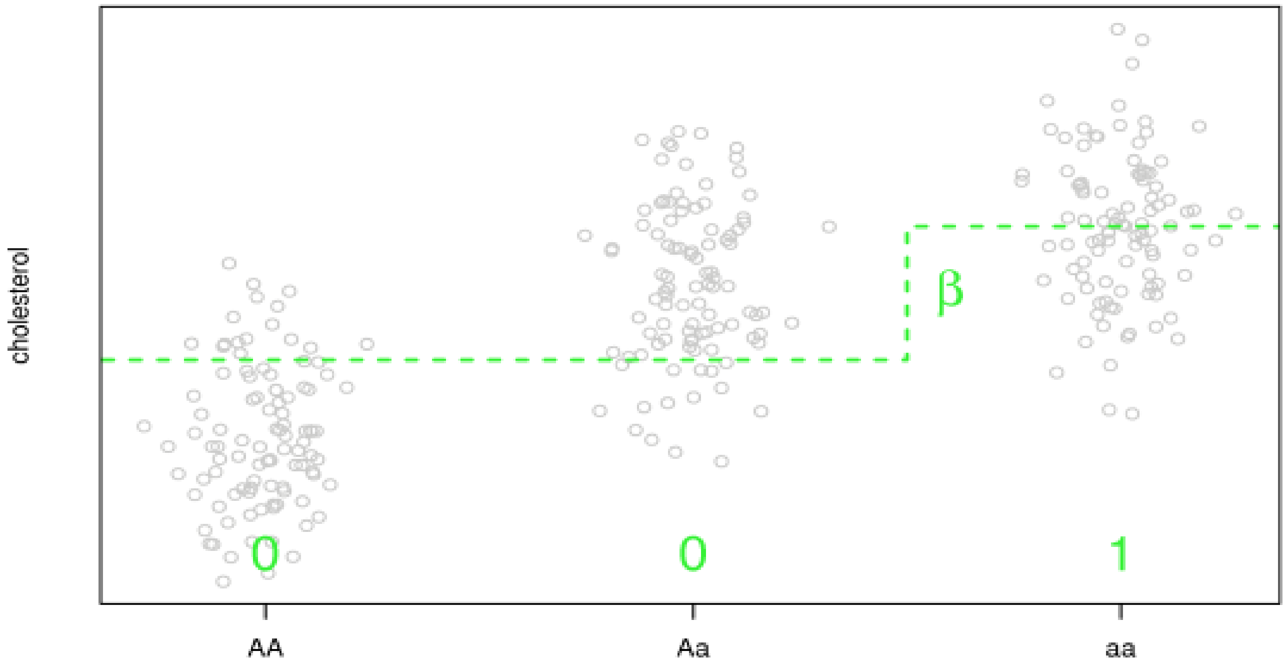
Use of lm() in genetics

Dominant model (best fit to this data)



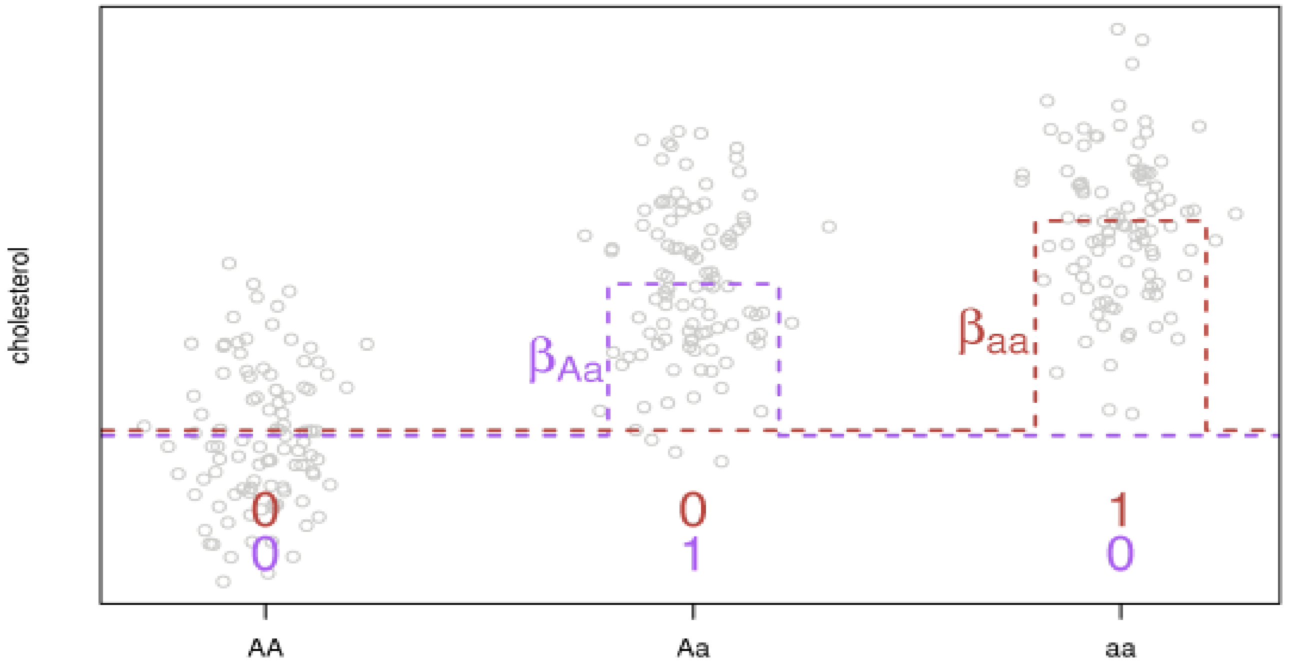
Use of `lm()` in genetics

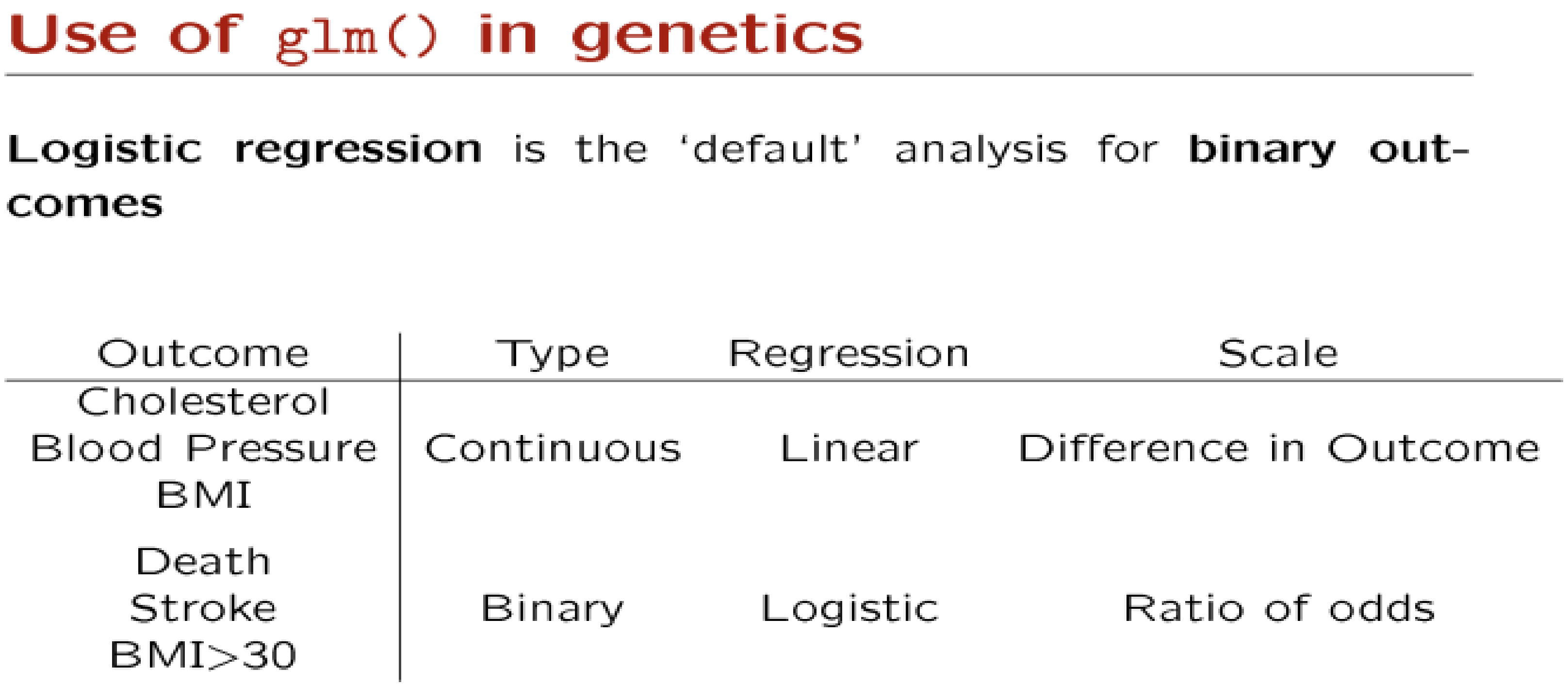
Recessive model (least stable for rare `aa`)



Use of `lm()` in genetics

2 parameter model (robust but can be overkill)





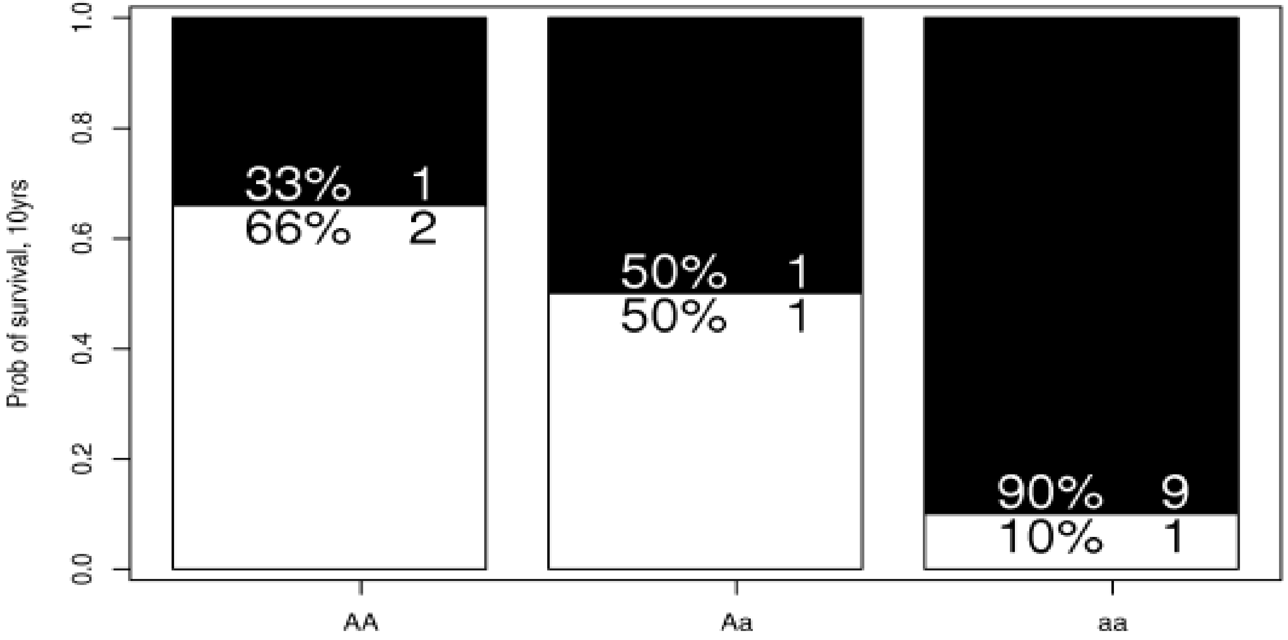
Use of `glm()` in genetics

Logistic regression is the 'default' analysis for **binary outcomes**

Outcome	Type	Regression	Scale
Cholesterol Blood Pressure BMI	Continuous	Linear	Difference in Outcome
Death Stroke BMI>30	Binary	Logistic	Ratio of odds

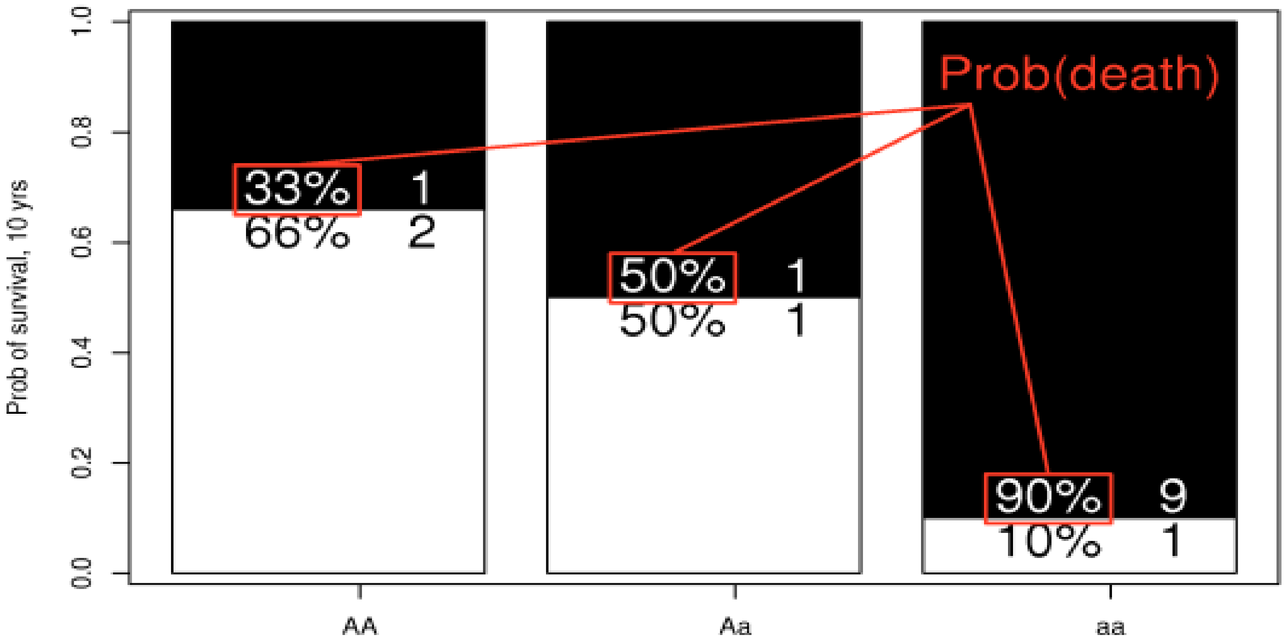
Use of `glm()` in genetics

Odds are a [gambling-friendly] measure of chance;



Use of `glm()` in genetics

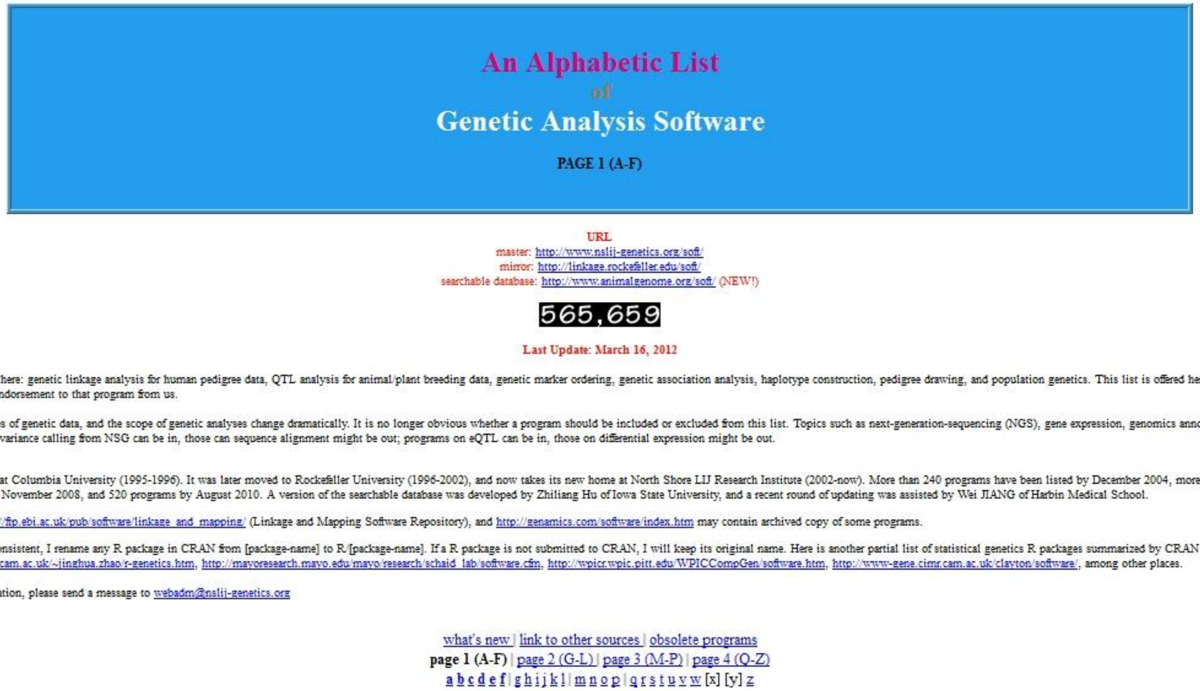
Odds are a [gambling-friendly] measure of chance;



Can screening for 1000nds of SNPs be performed automatically in R?

- **GenAbel** is designed for the efficient storage and handling of GWAS data with fast analysis tools for quality control, association with **binary and quantitative traits**, as well as tools for visualizing results.
- *pbatR* provides a GUI to the powerful PBAT software which performs family and population based family and population based studies. The software has been implemented to take advantage of parallel processing, which vastly reduces the computational time required for GWAS.
- *SNPassoc* provides another package for carrying out GWAS analysis. It offers descriptive statistics of the data (including patterns of missing data!) and tests for Hardy-Weinberg equilibrium. Single-point analyses with binary or quantitative traits are implemented via generalized linear models, and multiple SNPs can be analyzed for haplotypic associations or epistasis.

Is there one tool that fits it all? NO



An Alphabetic List
of
Genetic Analysis Software

PAGE 1 (A-F)

URL
master: http://www.nslij-genetics.org/soft/
mirror: http://linkage.rockefeller.edu/soft/
searchable database: http://www.animalgenome.org/soft/ (NEW!)

565,659

Last Update: March 16, 2012

Computer software on the following topics are included here: genetic linkage analysis for human pedigree data, QTL analysis for animal/plant breeding data, genetic marker ordering, genetic association analysis, haplotype construction, pedigree drawing, and population genetics. This list is offered here as a service to the gene mapping community. The inclusion of a program should not be interpreted as an endorsement to that program from us.

In the last few years, new technology produces new types of genetic data, and the scope of genetic analyses change dramatically. It is no longer obvious whether a program should be included or excluded from this list. Topics such as next-generation-sequencing (NGS), gene expression, genomics annotation, etc. can all be relevant to a genetic study, yet be specialized topics by themselves. Though programs on variance calling from NSG can be in, those can sequence alignment might be out; programs on eQTL can be in, those on differential expression might be out.

This page was created by Dr. Wentian Li, when he was at Columbia University (1995-1996). It was later moved to Rockefeller University (1996-2002), and now takes its new home at North Shore LIJ Research Institute (2002-now). More than 240 programs have been listed by December 2004, more than 350 programs by August 2005, close to 400 programs by December 2006, close to 480 programs by November 2008, and 520 programs by August 2010. A version of the searchable database was developed by Zhiliang Hu of Iowa State University, and a recent round of updating was assisted by Wei JIANG of Harbin Medical School.

Some earlier software can be downloaded from EBI: ftp://ftp.ebi.ac.uk/pub/software/linkage_and_mapping/ (Linkage and Mapping Software Repository), and http://genamics.com/software/index.htm may contain archived copy of some programs.

More and more packages are now written in R. To be consistent, I rename any R package in CRAN from [package-name] to R/[package-name]. If a R package is not submitted to CRAN, I will keep its original name. Here is another partial list of statistical genetics R packages summarized by CRAN (http://cran.r-project.org/web/views/Genetics.html). Yet more R packages can be found in: http://www.mrc-epid.cam.ac.uk/~jinghua.zhao/r-genetics.htm, http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm, http://wpicr.wpic.pitt.edu/WPICCompGen/software.htm, http://www-gene.cimr.cam.ac.uk/clayton/software/, among other places.

If you have new programs to add or any updated information, please send a message to webadm@nslij-genetics.org

what's new | link to other sources | obsolete programs
page 1 (A-F) | page 2 (G-L) | page 3 (M-P) | page 4 (Q-Z)
a b c d e f | g h i j k l | m n o p | q r s t u v w [x] [y] z

(http://linkage.rockefeller.edu/soft/)

Other analytic methods

- Recursive Partitioning (CART; Breiman 1984, Foulkes 2005)
- Random Forests (Pavolov 1997)
- Combinatorial Partitioning (Nelson 2001)
- **Multifactor-Dimensionality Reduction (Ritchie 2001)**
- Permutation-Based Procedures (Trimming/Weighting; Hoh 2000)
- Multivariate Adaptive Regression Splines (Friedman 1991)
- Boosting (Schapire 1990)
- Support Vector Machines (Vapnik 2000)
- Neural Networks (Friedman & Tukey 1974, Friedman & Stuetzle 1981)
- Bayesian Pathway Modeling (Conti 2003, Cortessis & Thomas 2004)
- Clique-Finding (Mushlin 2006)

What is a multiple testing correction?

- Simultaneously test *m* null hypotheses, one for each SNP *j*

  H_{0j}: no association between SNP *j* and the trait
- Every statistical test comes with an inherent false positive, or type I error rate—which is equal to the threshold set for statistical significance, generally 0.05.
- However, this is just the error rate for one test. When more than one test is run, the overall type I error rate is much greater than 5%.

What is a multiple testing correction?

- Suppose 100 statistical tests are run when (1) there are no real effects and (2) these tests are independent, then the probability that no false positives occur in 100 tests is 0.95^{100} = 0.006. So the probability that at least one false positive occurs is 1-0.006=0.994 or 99.4%
- There is not a single measure to quantify false positives (Hochberg et al 1987).
- Several multiple testing corrections have been developed and curtailed to a genome-wide association context, when deemed necessary: *Bonferroni* (highly conservative) [divide each single SNP-based p-value by the nr of tests before comparing to the nominal sign level 0.05] vs *permutation-based* (highly computational demanding) [keep the LD structure, but swap the trait labels among the subjects]

4.c Replication

nature genetics

Freely associating

Editorial: Once and Again—Issues Surrounding Replication in Genetic Association Studies

J. Hirschhorn

May

PERSPECTIVE

The Future of Association Studies: Gene-Based Analysis and Replication

Benjamin M. Neale[1] and Pak C. Sham[1,2]

Am J Hum Genet July

Editorial

Replication Publication

Mark Patterson

Statistical false positive or true disease pathway?

John A Todd

Nat Genet July 2006

What does replication mean?

- Replicating the genotype-phenotype association is the “gold standard” for “proving” an association is genuine
- Most loci underlying complex diseases will not be of large effect. It is unlikely that a single study will unequivocally establish an association without the need for replication
- SNPs most likely to replicate:
    - Showing modest to strong statistical significance
    - Having common minor allele frequency
    - Exhibiting modest to strong genetic effect size

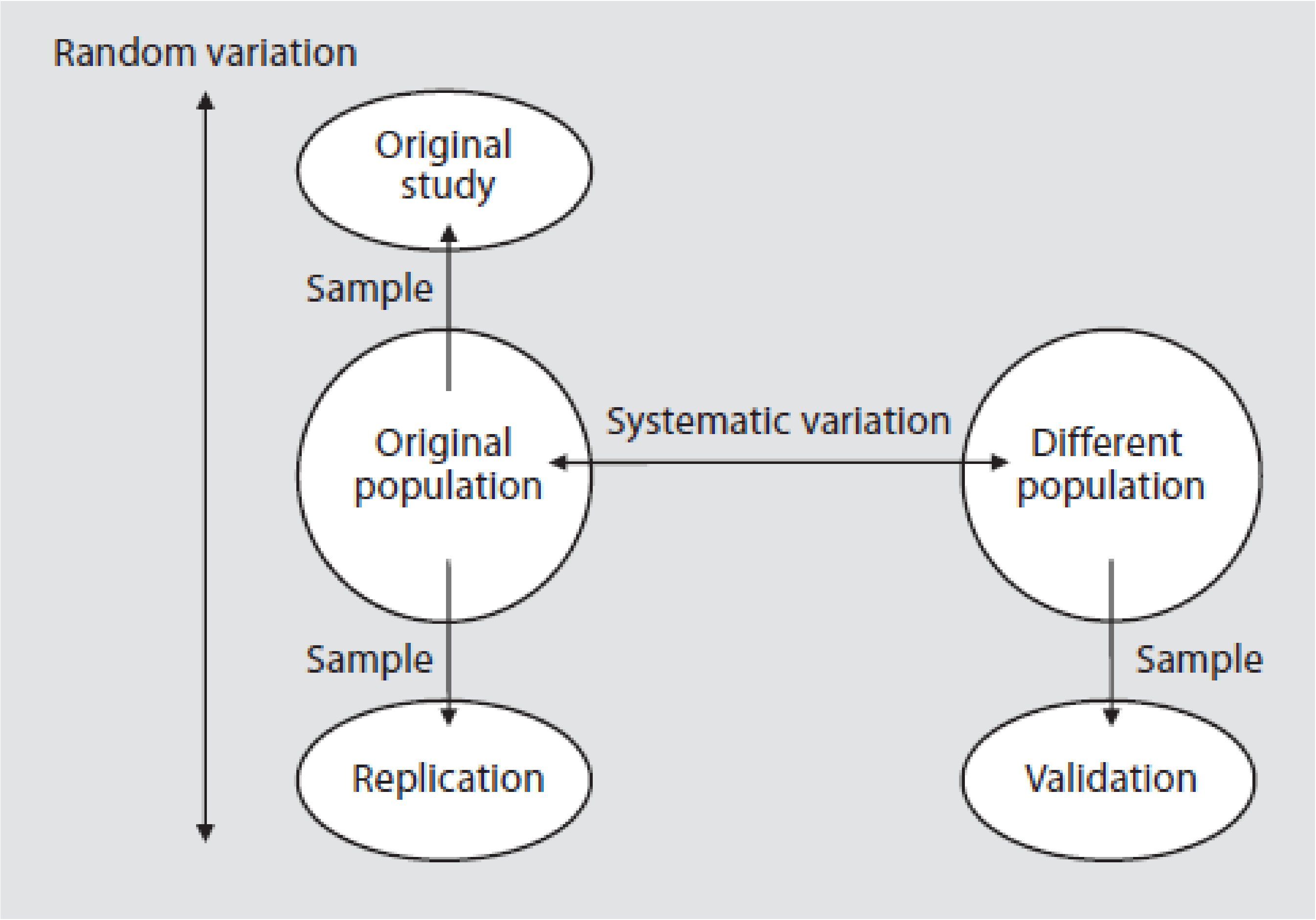
- Note: Multi-stage design analysis results should not be seen as “evidence for replication” ...

Guidelines for replication studies

- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal should be in the same direction
- Joint analysis should lead to a lower *p*-value than the original report
- Well-designed negative studies are valuable

➔ **check the NHGRI Catalog of GWA studies**
www.genome.gov/gwastudies/

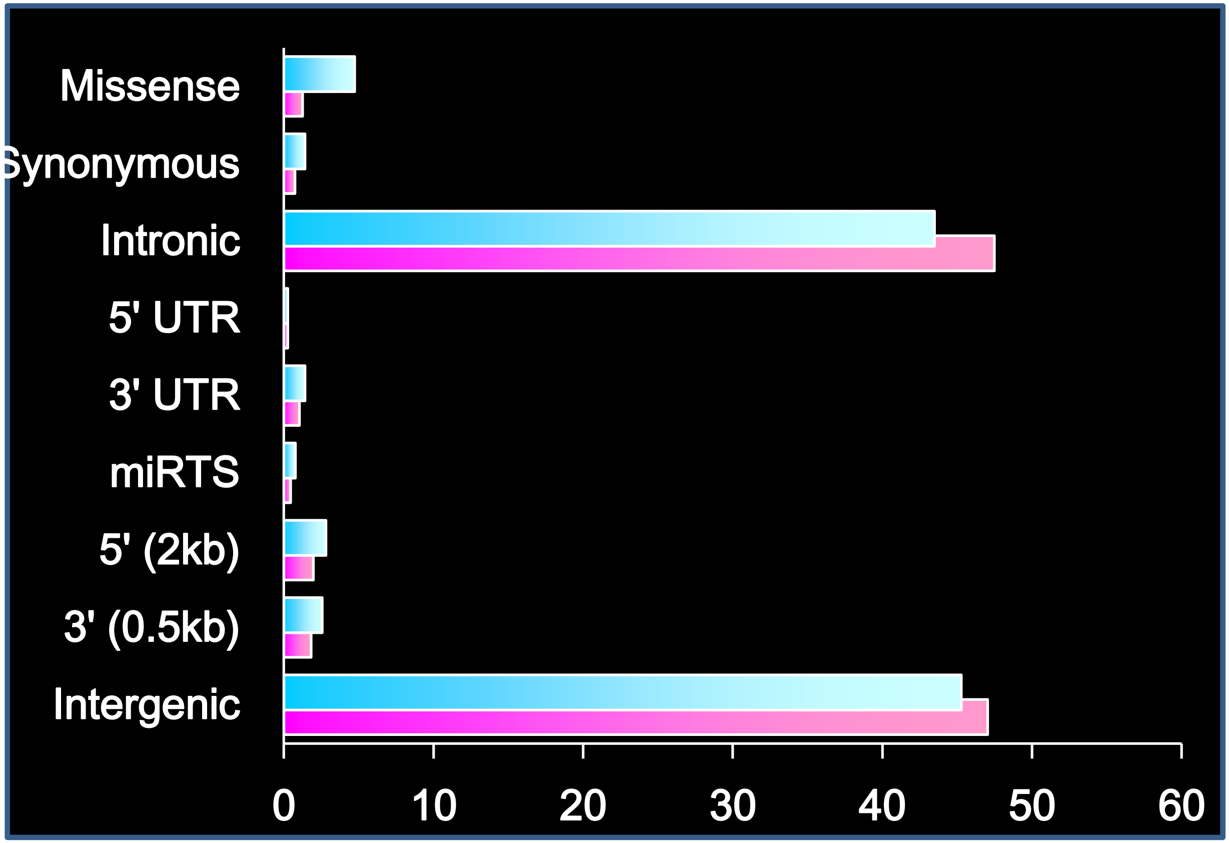
What does validation mean?



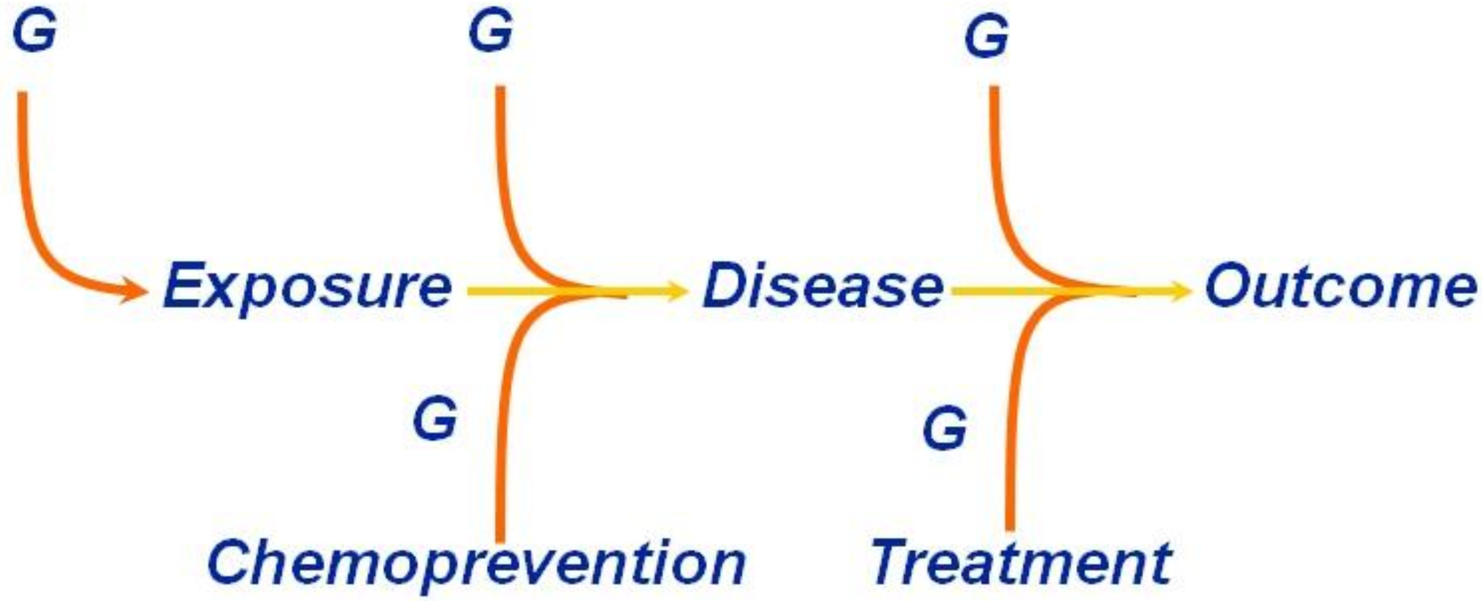
(Igl et al. 2009)

5 Interpretation and follow-up

What have GWA studies learnt us about functionality? (Manolio 2010)



What have GWA studies learnt us about functionality? (Rebbeck et al 2004)



Are there criteria for assessing the functional significance of a variant?

Criterion	Strong Support	Moderate Support	Neutral Information	Evidence Against
Nucleotide Sequence	Variant disrupts a known functional motif	missense change, disrupts putative functional motif	-	Non-functional change
Evolutionary Conservation	Strong conservation across species, multigene family	Some conservation across species or multigene family	Not known	No conservation
Population Genetics	Strong deviations from expected frequencies	Some deviations from expected frequencies	Not known	No deviations from expected frequencies
Experimental	Consistent evidence in human target tissue	Some evidence	No data available	No functional effect
Exposures	Variant affects relevant metabolism in target tissue	Variant affects metabolism	No data available	Variant does not affect metabolism
Epidemiology	Consistent and reproducible reports	Reports without replication	No data available	No association

"The more we find, the more we see, the more we come to learn.

The more that we explore, the more we shall return."

Sir Tim Rice, *Aida*, 2000

Main References:

- Ziegler A and Van Steen K 2010: IBS short course on "Genome-Wide Association Studies"
- Balding D 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics, 7, 781-791.
- Kruglyak L 2008. The road to genomewide association studies. Nature Reviews Genetics 9: 314-
- Wang et al 2005. Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics 6: 109-
- Peltonen L and McKusick VA 2001. Dissecting human disease in the postgenomic era. Science 291, 1224-1229
- Li 2007. Three lectures on case-control genetic association analysis. Briefings in bioinformatics 9: 1-13.
- Rebbeck et al 2004. Assessing the function of genetic variants in candidate gene association studies 5: 589-
- Robinson 2010. Common Disease, Multiple Rare (and Distant) Variants. PLoS Biology 8(1): e1000293