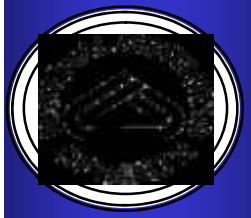
U.S. Department of Health and Human Services



National Institutes of Health

National Heart, Lung, and Blood Institute

OVERVIEW OF INTERNATIONAL AND NATIONAL LARGE POPULATION STUDIES

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

NATIONAL INSTITUTES OF HEALTH

NATIONAL HEART, LUNG, AND BLOOD INSTITUTE

Teri A. Manolio, M.D., Ph.D.
Director, Epidemiology and Biometry Program
Division of Epidemiology and Clinical Applications

ISSUES TO BE ADDRESSED

- Ongoing and planned large cohort studies
- Optimal design of prospective cohort studies
- Optimal design of case-control studies
- Challenges of phenotype definition and disease-based endpoints in epidemiologic studies
- Use of existing cohorts vs. establishing new cohorts for study of genetic variants and environment

ONGOING AND PLANNED LARGE COHORT STUDIES

International

- Biobank Japan
- deCODE Genetics
- Estonian Genome Project
- Public Population Project in Genetics (P3G): CARTaGENE, GenomeEUtwin, EGP, CIGMR
- UK Biobank

U.S.

- Marshfield Personalized Medicine Project
- National Children's Study
- Variety of clinical samples at NWU, Duke, etc.

INTERNATIONAL AND NATIONAL COHORT STUDIES

Study	Country	Size	Age	Outcomes	Access
Biobank Japan	Japan	300,000	20+	47 common diseases	Limited to Japan
deCODE Genetics	Iceland	~200,000	All	50 common diseases	With colla-boration
Estonian Genome Project	Estonia	100,000 – 1,000,000		Common diseases	With colla-boration
UK Biobank	UK	500,000	40-65	Common diseases	With colla-boration

INTERNATIONAL AND NATIONAL COHORT STUDIES

Study	Country	Size	Age	Outcomes	Access
CARTa GENE	Canada	~50,000	25-74	Common diseases	With colla-boration
GenomeEU twin	7 European countries	800,000 twin pairs		Stature, obesity, migraine, CHD, stroke, longevity	With colla-boration
Marshfield PMP	US	40,000	18+	Multiple; ADRs	
National Children's Study	US	100,000 infants	0-21	Multiple; chem/ env exposures, development	

BIOBANK JAPAN

- To clarify on large the causes of disease and medication side effects in relation to genetic variants; ultimately to develop new drugs and diagnostics ("personalized medicine")
- Samples and patient data will be collected by network of collaborating organizations and private universities, including Nihon University, Juntendo University and Tokushukai group ("3rd largest hospital group in world")
- Project hoped to stimulate development of legislation to protect personal information
- Begun 2003, 90,000 samples collected to date (120,000 disease cases; each pt has 1.3 diseases)
- Distribution of DNA/serum to Japanese researchers begun

ESTONIAN GENOME PROJECT

- To find links between genes, environmental factors and common diseases and apply information to increase efficiency of health care
- Up to 1M persons, though seeming to scale down to 100,000
- Begun 10/2002; 10,000 recruited in pilot as of 2004
- Written informed consent; 60-90-minute questionnaire including genealogy; ht, wt, BP, HR; 50 ml blood sample
- Personalized information to be made available to participants ("gene donors") and physicians
- Non-profit EBP Foundation in public-private partnership with EGeen Inc., exclusive commercial licensee of data base
- Contracts terminated by mutual agreement in 12/2004

MARSHFIELD PERSONALIZED MEDICINE PROJECT

- To translate genetic data into specific knowledge about disease that is clinically relevant and will enhance patient care
- Utilizes Marshfield Epidemiologic Study Area
  - Marshfield Clinic system of care
  - Long-standing electronic medical record
  - Active programs in genomics and clinical research
- Up to 40,000 persons ages 18+
- Begun 9/2002; 17,000 recruited with response rate ~ 45%
- Written informed consent, 30-minute visit with questionnaires, DNA extraction, blood and serum stored
- Data encrypted, no one with access to identifiable clinical information will also have access to genetic information

DECODE GENETICS

- Biopharmaceutical company applying its discoveries in human genetics to development of drugs for common diseases
- Utilizes unique resources of Icelandic nation
  - Relatively isolated population, founder effects
  - Genealogic database extending to settlement ~ 900 A.D.
  - Small number of referral hospitals, good record systems
- Currently 110,000 Icelanders, 25,000 non-Icelanders
- Begun 1998; controversy over proposal for "opt-out" consent to link health records, health database ultimately abandoned
- Written informed consent for all genetic studies, third-party encryption

FAMILIAL AGGREGATION OF A BROADLY DEFINED PHENOTYPE OF ATRIAL FIBRILLATION IN ICELAND

David O. Arnar MD PhD, Sverrir Thorvaldsson MS, Teri Manolio, MD PhD, Kristleifur Kristjansson MD, Augustine Kong PhD, Gudmundur Thorgeirsson MD PhD, Hakon Hakonarson MD PhD, Kari Stefansson MD PhD

Landspitali University Hospital, Reykajvik, Iceland, deCODE Genetics Inc, Reykjavik, Iceland, National Institutes of Health (NHLBI), Bethesda, MD, USA

Presented at American Society for Human Genetics, Toronto, 2004.

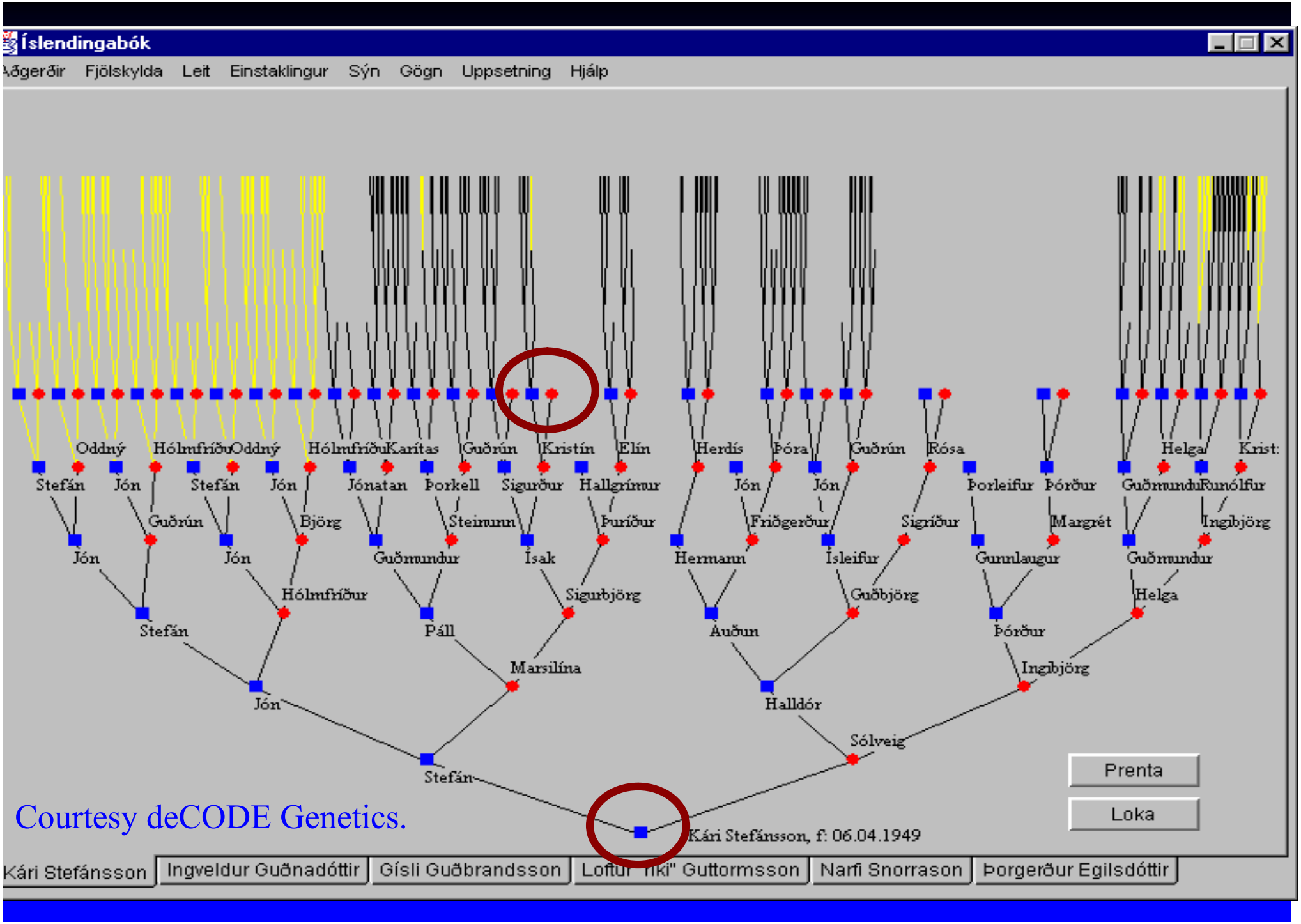
UNIQUENESS OF ICELANDIC POPULATION FOR GENETIC STUDIES

- Iceland founded in 9th century by settlers of mixed Northern European descent
- Population ~ 285,000
- Careful genealogic records

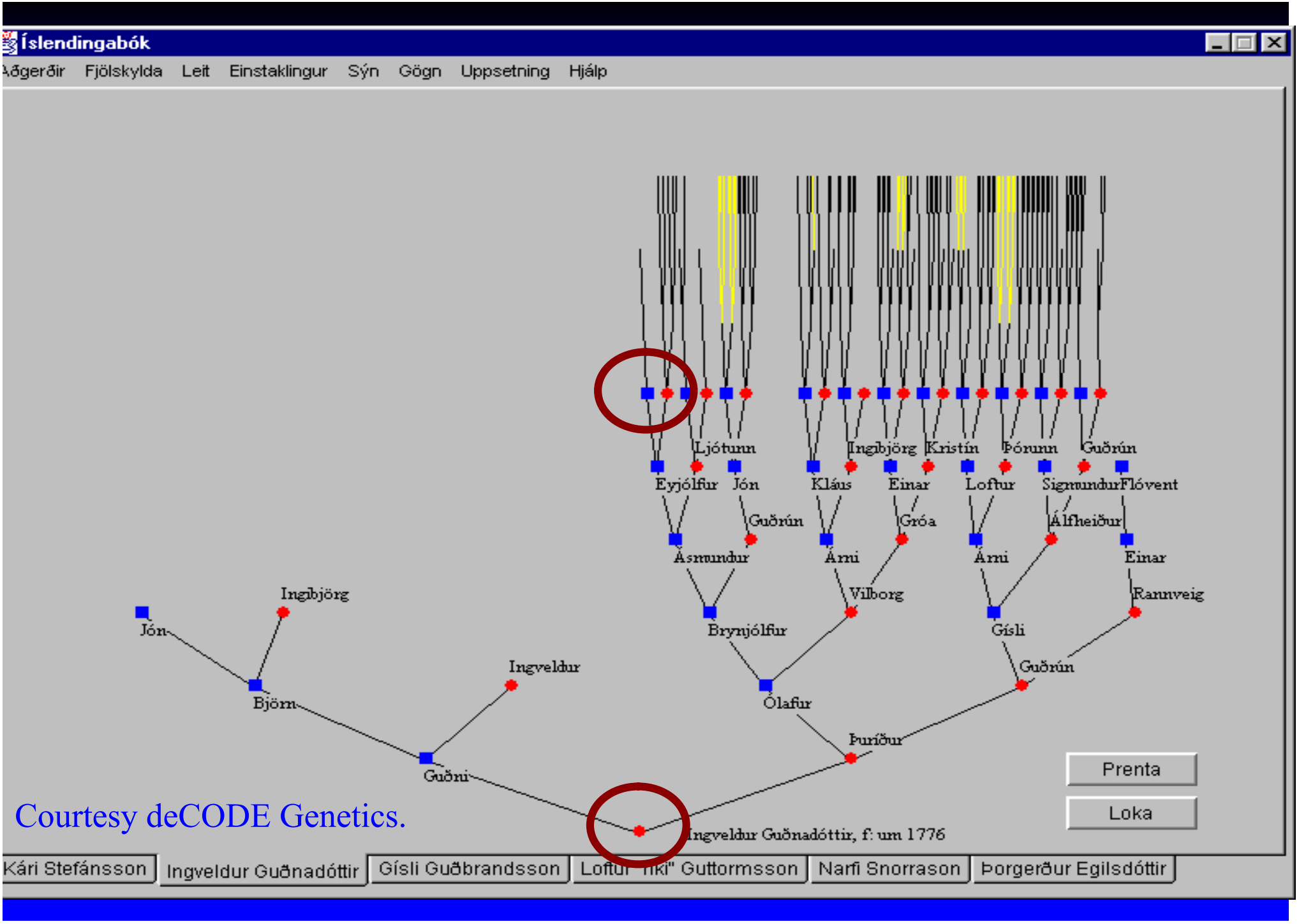
N. European descent → Same genetic background

Isolation for 11 centuries → Fewer variants

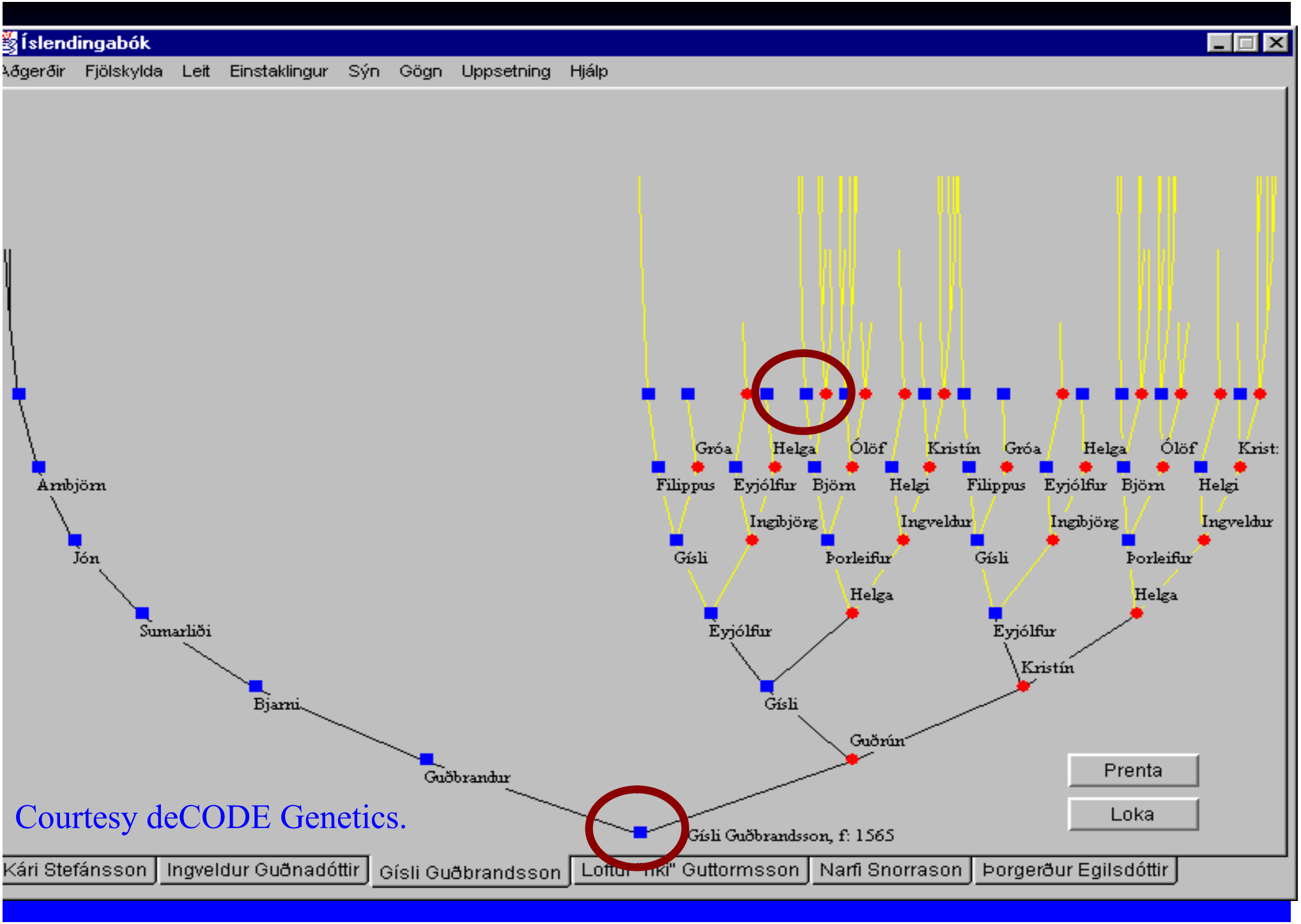
Courtesy deCODE Genetics.



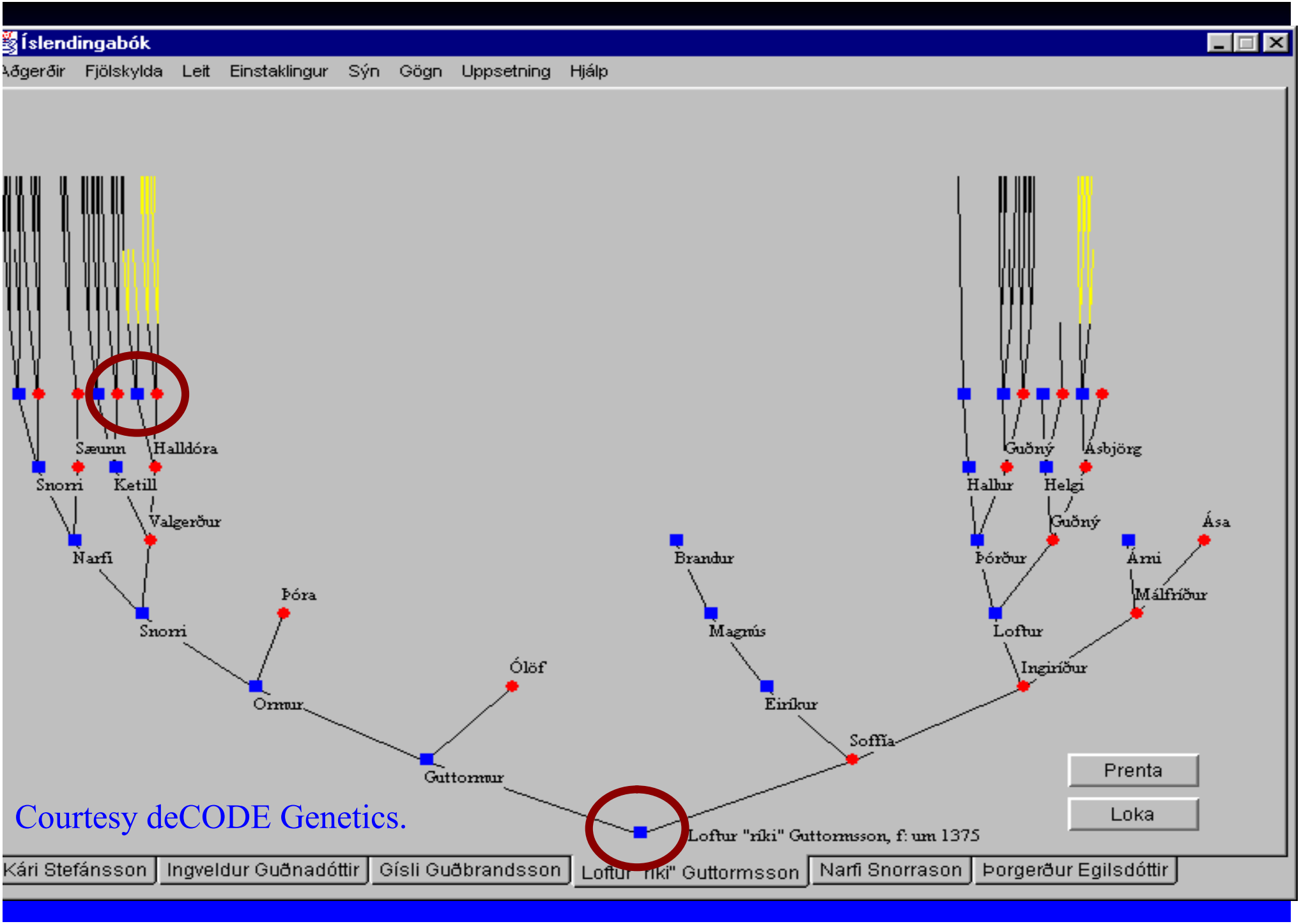
Courtesy deCODE Genetics.



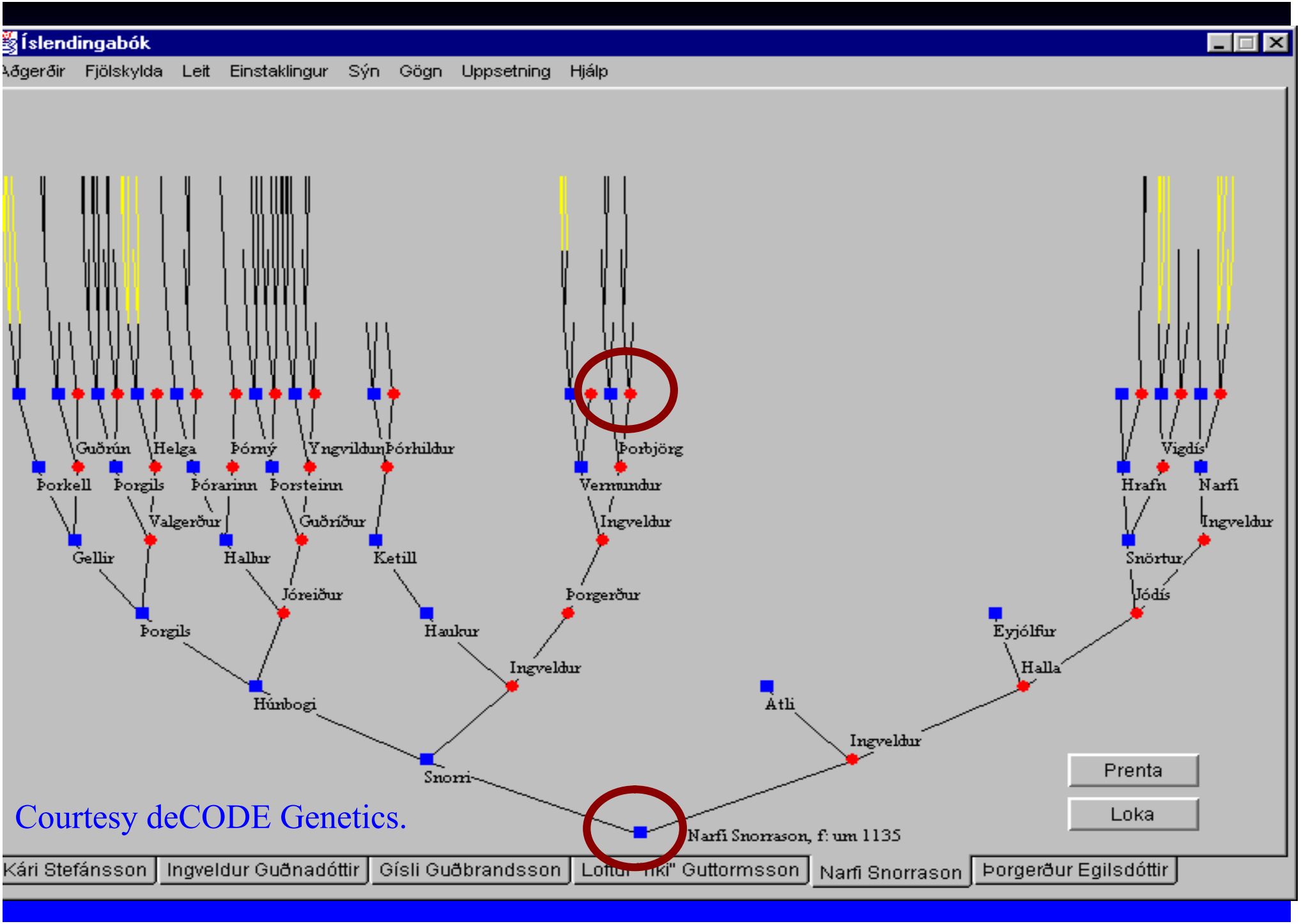
Courtesy deCODE Genetics.



Courtesy deCODE Genetics.



Courtesy deCODE Genetics.



Courtesy deCODE Genetics.

Íslendingabók

Aðgerðir Fjölskylda Leit Einstaklingur Sýn Gögn Uppsetning Hjálp

Skoða einstakling

Faðir	Móðir
Skallagrímur Kveldúlfsson	Bera Yngvarsdóttir
F: (863)	F: (870)
D:	D:
Landnámsmaður.	
Heimildir: Landnáma	Heimildir: Landnáma

Egill Skallagrímsson

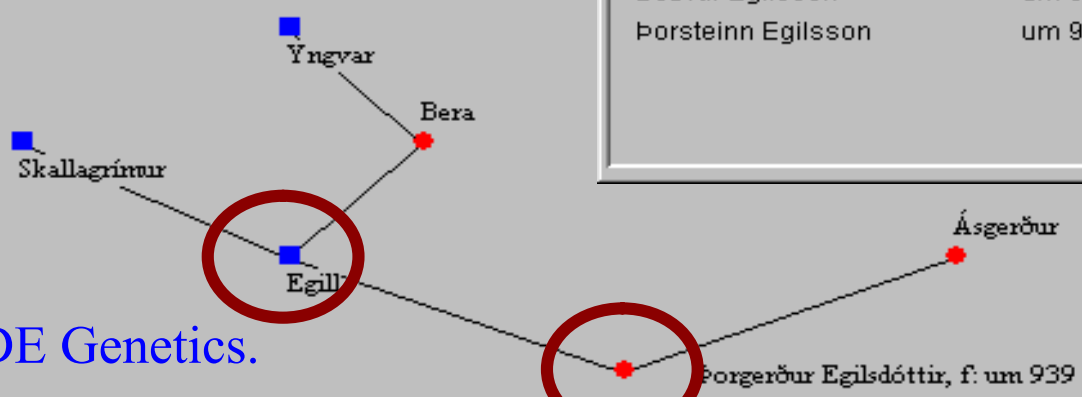
Fæddur: (910) 77875

Heimildir: Landnáma

Skáld. Bjó á Borg.

Makar og börn

Ásgerður Bjarnardóttir	(915)	
Þorgerður Egilsdóttir	um 939	
Bera Egilsdóttir	um 940	
Gunnar Egilsson	(942)	
Böðvar Egilsson	um 943 - um 957	Drukknaði.
Þorsteinn Egilsson	um 945	".. bjó að Borg. Hann a



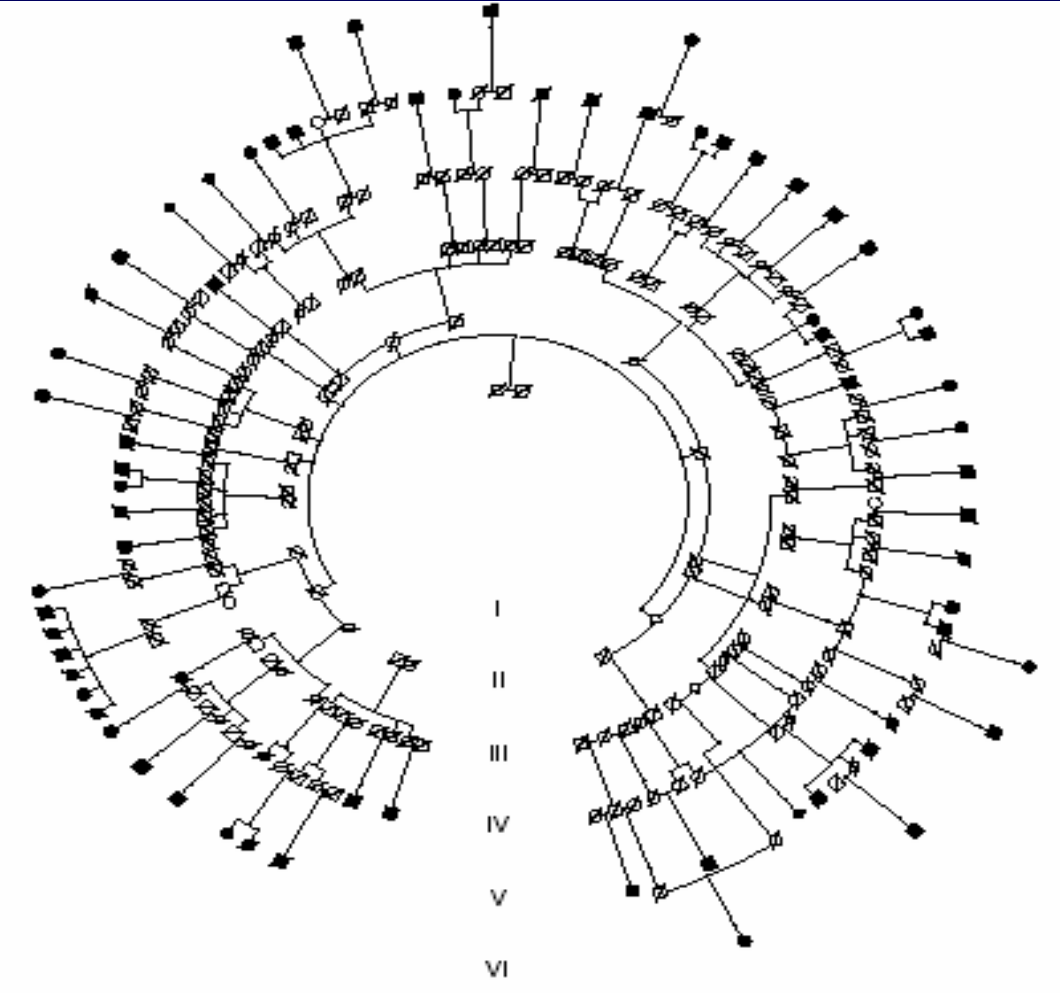
Prenta

Loka

Kári Stefánsson | Ingveldur Guðnadóttir | Gísli Guðbrandsson | Lofti "Ríki" Guttormsson | Narfi Snorrason | Þorgerður Egilsdóttir

Courtesy deCODE Genetics.

LARGE REPRESENTATIVE PEDIGREE SHOWING 69 PATIENTS WITH ATRIAL FIBRILLATION



Arnar *et al*, ASHG 2004.

STRENGTH OF EXTENSIVE GENEALOGIES

- Common diseases do not show Mendelian inheritance patterns
- Affected siblings infrequent in common diseases, but many patients may have more distant relatives with same disease

Degree of Relatives	Risk Ratio [95% CI]	P-Value
1	1.77 [1.67,1.88]	< 0.001
2	1.36 [1.27,1.44]	< 0.001
3	1.18 [1.14,1.23]	< 0.001
4	1.10 [1.06,1.13]	< 0.001
5	1.05 [1.02,1.07]	< 0.001

Arnar *et al*, ASHG 2004.

DISEASES WITH GENES MAPPED, VARIANTS IDENTIFIED, DRUG TARGETS IN TESTING

- Age-related macular degeneration
- Alzheimer´s disease
- Anxiety
- Asthma
- Atopy
- Benign Prostatic Hyperplasia
- Chronic Obstructive Pulmonary Disorder
- Essential tremor
- Familial combined hyperlipidemia
- Hypertension
- Longevity
- Migraine
- Myocardial infarction
- Non-insulin dependent diabetes
- Obesity
- Osteoarthritis
- Osteoporosis
- Parkinson´s disease
- Peripheral artery occlusive disorder
- Pre-eclampsia
- Prostate cancer
- Psoriasis
- Rheumatoid arthritis
- Schizophrenia
- Stroke

Courtesy deCODE Genetics.

POPULATION IMPACT OF RISK-RELATED GENETIC VARIANTS

"Genes are merely risk factors passed on from parents to children....

- Determine prevalence of variants in diverse groups
- Examine associations identified in family studies, assess magnitude and independence

  common risk factors are not strong

  strong risk factors are not common
- Define associations with variety of phenotypes
- Identify factors, particularly environmental factors, modifying genotype-phenotype relationships

"EPIDEMIOLOGIC ARCHITECTURE" OF RECENTLY IDENTIFIED GENETIC VARIANTS

Variant	q_A (%)	Risk	Other Phenotypes, Associations	Modifiers
NRG1 haplo core	7.5	2.2	--	--
PDE4D haplo G0	8.8	2.0	stroke risk factors, subtypes	--
BMP2 haplo C	1.4-1.9	1.8-4.4	low BMD, fractures, sites	menopausal status (?)

tefansson H et al, *Am J Hum Genet* 2002;71:877-892; Grettarsdottir S et al, *Nat Genet* 003;35:131-138; Strykarsdottir U et al; PLoS Biology; 2003;1(3):1-10.

NEED FOR LARGE COHORT STUDY OF GENES AND ENVIRONMENT

Identifying and reducing disease risk depends on unbiased determination of:

- quantitative contributions of environmental and genetic factors
- interactions among them
- complex interplay among disorders sharing common risk factors (such as heart disease, hypertension, and diabetes)

Replication of associations and estimation of their magnitude, consistency, and temporality best obtained through prospective, population-based cohort studies

Collins FS, *Nature* 2004; 429:475-477.

POPULATION-BASED COHORT STUDIES

- Definition: prospective investigation of representative sample of population followed for development of specified endpoints
- Purpose: to identify risk factors predisposing to development of disease in the general population, particularly risk factors:

  affected by disease, treatment, lifestyle changes
  subject to imperfect or biased recall
  with hypothesized early pathogenic effect

- Complement other epidemiologic study designs:

  surveillance studies
  cross-sectional surveys
  case-control studies
  clinical epidemiology studies

MAJOR NHLBI COHORT STUDIES

Study	N	Age	Entry	Minorities
Framingham Cohort	5,209	28-62	1948-50	--
Framingham Offspring	5,124	20-74	1971-75	--
Framingham Gen3	~ 4,000	20-60	2002-04	--
Honolulu Heart Program	8,006m	46-68	1965-68	100% JA
CARDIA	5,115	18-30	1985-86	52% AA
ARIC	15,787	45-64	1985-87	27% AA
CHS	5,888	65-100	1989-90	16% AA
Strong Heart Study	4,549	45-74	1989-91	100% AI
Women's Health Initiative	161,809w	50-79	1993-98	18% multiple
MESA	6,749	45-84	2000-02	28% AA, 22% HA, 12% CA
Jackson Heart Study	5,308	35-84	2000-04	100% AA
Hispanic Cohort	16,000	35-84	2006-10	100% HA

PROS AND CONS OF COHORT STUDIES

DISADVANTAGES

- They are expensive.
- They take a long time.
- They are very broad-based.

ADVANTAGES

- They provide risk information obtainable through no other means.
- They are understandable to the public and media.
- They identify modifiable risk factors for potential preventive interventions.

CHARACTERISTICS OF IDEAL COHORT STUDY

- Size matters
- Representative sample that can be generalized back to source population; randomly sampled with high response rate
- Diverse in geography, socioeconomic status, race/ethnicity
- Extensive, standardized, reproducible characterization of exposures, risk factors and disease status at entry
- Repeated interim measures to assess change in exposures, disease status; add new exposure measures
- Comprehensive, standardized assessment of outcomes

TYPES OF BIAS IN EPIDEMIOLOGIC RESEARCH

Bias: "Any effect at any stage of investigation or inference tending to produce results that depart systematically from the true value (to be distinguished from random error)."

--Last's *Dictionary of Epidemiology*, 1983

Selection

- Non-respondent
- Prevalence-incidence
- Admission rate
- Detection signal
- Membership
- Lead-time

Observer/Interviewer

- Diagnostic suspicion
- Exposure suspicion
- Recall
- Family information
- Ascertainment
- Reporting

Sackett DL *J Chr Dis* 1979; 32:51-63.

BASIC ASSUMPTIONS FOR BIAS-FREE CASE-CONTROL STUDY

- Cases are representative of all persons who develop the disease/condition
- Controls are representative of the general “healthy” population who do not develop the disease
- Collection of risk factor and exposure information is the same for cases and controls

PROS AND CONS OF CASE-CONTROL STUDIES

ADVANTAGES

- May be the only way to study rare diseases or those of long latency
- Existing records can occasionally be used if risk factor data collected independent of disease status
- Can study multiple etiologic factors simultaneously
- May be less time-consuming and expensive
- If assumptions met, inferences are reliable

PROS AND CONS OF CASE-CONTROL STUDIES

DISADVANTAGES

- Relies on recall or records for information on past exposures; validation can be difficult or impossible
- Selection of appropriate comparison group may be difficult
- Multiple biases may give spurious evidence of association between risk factor and disease
- Usually cannot study rare exposures
- Temporal relationship between exposure and disease can be difficult to determine

"BUT," THEY SAY, "*THIS* IS GENETICS!"

(you dumb epidemiologist)

"*THIS* IS DIFFERENT!"

- Genes are measured the same way in cases and controls
- Information on key exposure is easy to validate
- No recall or reporting involved
- Temporal relationship between genes and disease is clear

"BUT," I SAY,

- Bias-free ascertainment of cases and controls is still major concern; cases in most clinical series unlikely to be representative
- Assessment of risk modifiers or gene-environment interactions is likely to be incomplete or flawed

CASE-CONTROL STUDIES AND RARE DISEASES

- For a disease with incidence of 8 cases per 1,000 among unexposed, cohort study would require 3,889 exposed and 3,889 unexposed persons to detect two-fold increase in risk
- Case-control study would require 188 cases and 188 controls, assuming 30% exposure
- For disease with incidence of 2 cases per 1,000 among unexposed, would need 15,700 exposed and 15,700 unexposed to detect two-fold risk
- Case-control study would *still require only 188 cases and 188 controls*

Schlessman JJ. *Case-Control Studies*, 1982.

WHAT TO DO?

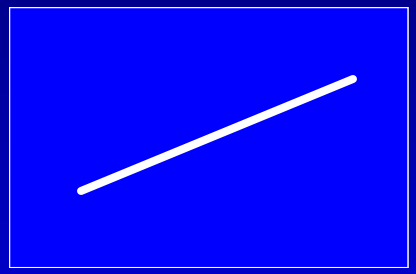
- "Nesting" a case-control study within a prospective cohort probably provides the best of both worlds
- Large proportion of cohort members who do not develop disease provide little incremental information
- If exposure information can be collected and stored for later measurement, can wait for cases to accrue and then measure exposures in limited sample of non-cases
  - stored biologic samples
  - stored images
- Can be expanded to "case-cohort" concept with representative sample of cohort, regardless of disease status, used for multiple comparisons

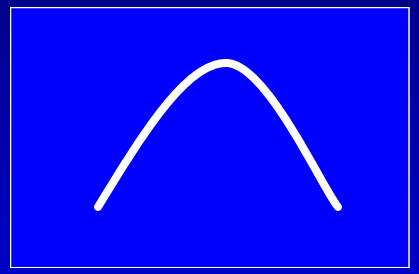
DISADVANTAGES OF FOCUS ON "DISEASE CASES" OR CLINICAL ENDPOINTS

- Clinical endpoint determination requires:
  - recognition of symptoms by study subject
  - relatively rapid access to sources of medical care
  - proper diagnosis by treating physician
- All involve potential biases, particularly in economically challenged countries without organized health care system
- Reliance solely on clinical endpoints can bias risk relationships due to under-detection, biased ascertainment, misclassification of cases

DISADVANTAGES OF DISEASE-BASED STUDIES

- Dichotomous outcomes almost always less powerful than continuous ones assuming one understands nature of continuous relationship





- Rely on crossing some threshold of definition or recognition that may be less relevant to overall health and well-being
- Lesser degrees of abnormality or dysfunction important in understanding pathophysiology and etiology
- Quantitative traits such as BP and BMI should lend themselves well to genetic association studies

QUANTITATIVE TRAITS IN NHLBI COHORT STUDIES

Study	BP	BMI	Chol	µAlb	Dep'n	Cog'n	PFT
Framingham Cohort	X	X	X	X	X	X	X
Framingham Offspring	X	X	X	X	X	X	X
Framingham Gen3	X	X	X	X			X
Honolulu Heart	X	X	X		X	X	X
CARDIA	X	X	X	X			X
ARIC	X	X	X	X	X	X	X
CHS	X	X	X	X	X	X	X
Strong Heart	X	X	X	X	X		X
WHI	X	X	X		X	X	
MESA	X	X	X	X	X		
Jackson Heart	X	X	X	X	X	X	X
Hispanic Cohort	X	X	X	X	X		X

SUBCLINICAL DISEASE AS A PHENOTYPE FOR GENETIC RESEARCH

Subclinical disease: disease detected non-invasively before it has produced signs and symptoms (*?????* or *kline*, bed or couch)

- Subclinical measures examine early stages of disease, are relatively free of biases related to severity, diagnostic suspicion, or completeness of medical investigation
- Subclinical disease unlikely to have directly affected health behavior, such as lifestyle modification or medication use
- Continuous nature of most subclinical measures enhances power to detect risk associations over discrete measures
- Subclinical measures permit epidemiologic investigation of disease risk to focus on biology of disease rather than on vagaries in its diagnosis

ADVANTAGES OF NEW COHORT

- Design: based on needs of study rather than convenience; get it right from the start
- State of art: use up-to-date technology, address current health concerns
- Consistent protocol: avoid lowest common denominator
- Poolability/survivorship: easier to pool on genetics than environment?
- Consent: more straightforward, up-to-date, avoiding complexity of many changes over time

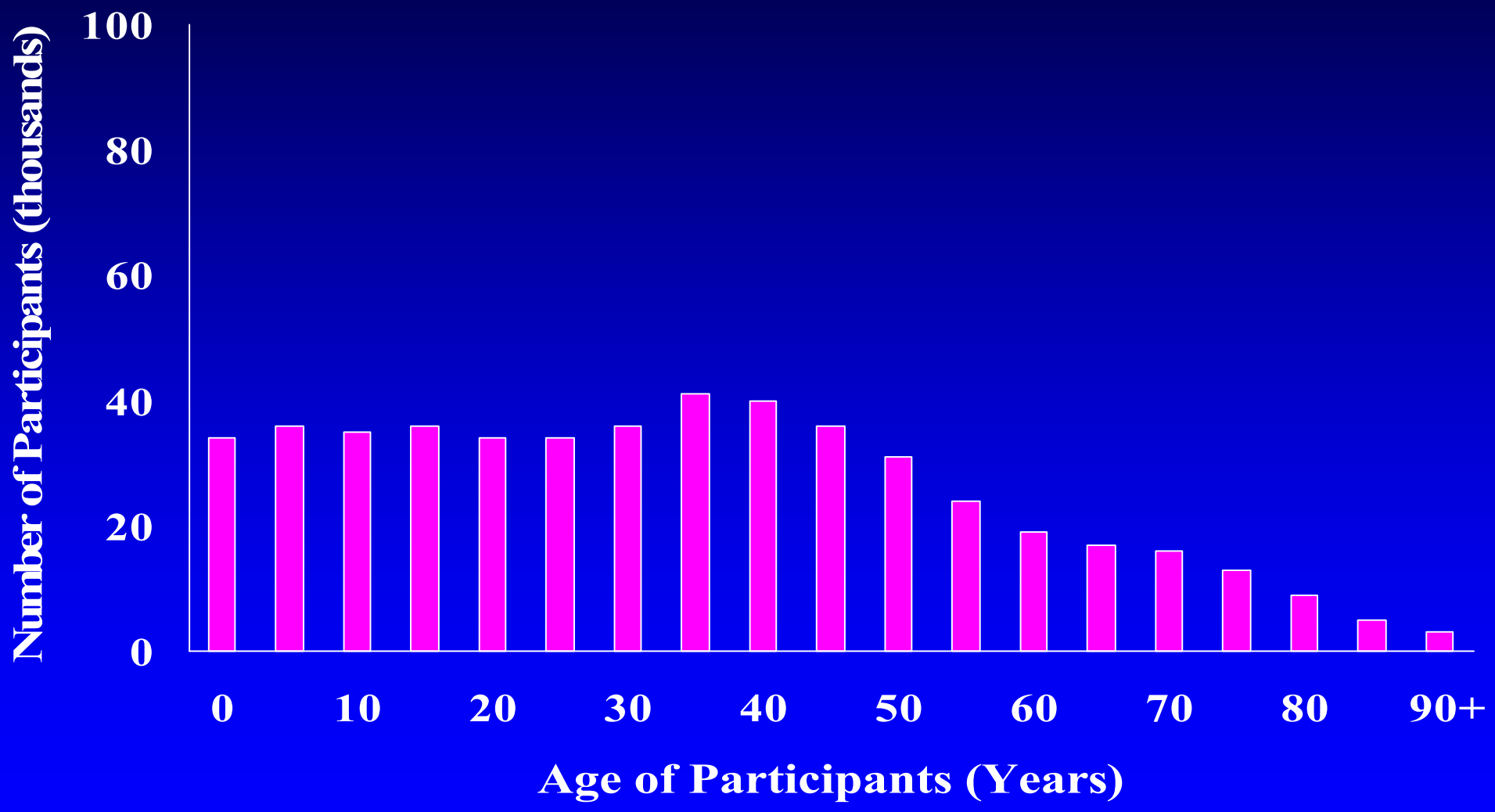
ADVANTAGES OF NEW COHORT (2)

- Multiple outcomes: built in from start
- Free and open access: establish up front; consider separating functions that store and distribute from those that collect and analyze
- Biologic specimens: fresh, high-quality, suitable for proteomics or RNA analysis
- Diversity
- Younger ages: most existing cohorts middle age or older

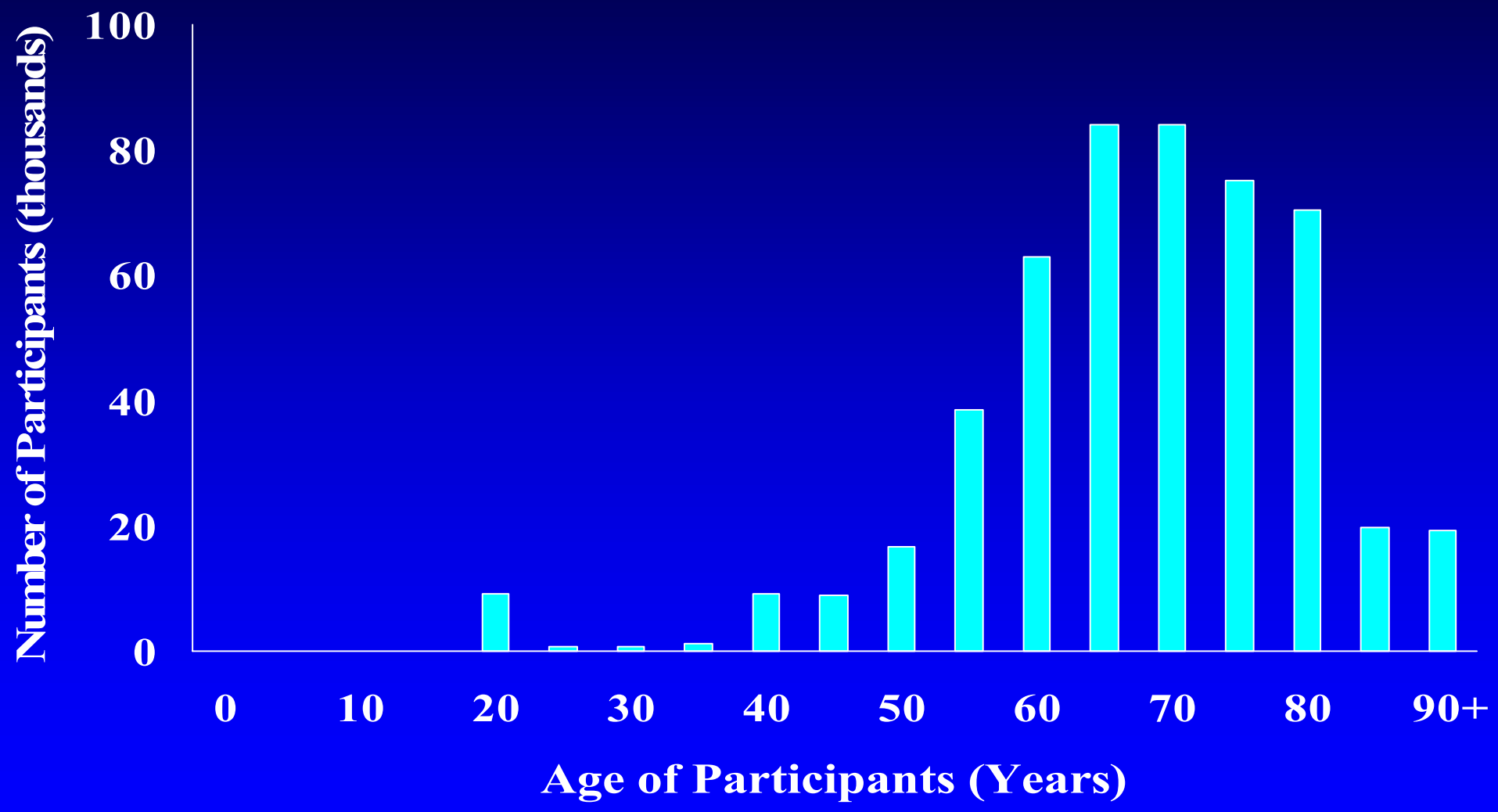
ADVANTAGES OF EXISTING COHORTS

- Saves time/money: usefully supplement in cost-effective way, leverage existing investment
- Experience and expertise: already shown can collect high quality data
- Recruitment: may have higher response rate
- Community responsiveness: relationships with communities already established
- IRB and institution-specific requirements: time-consuming, iterative process, already worked out
- Valuable ongoing work: don't be too quick to abandon

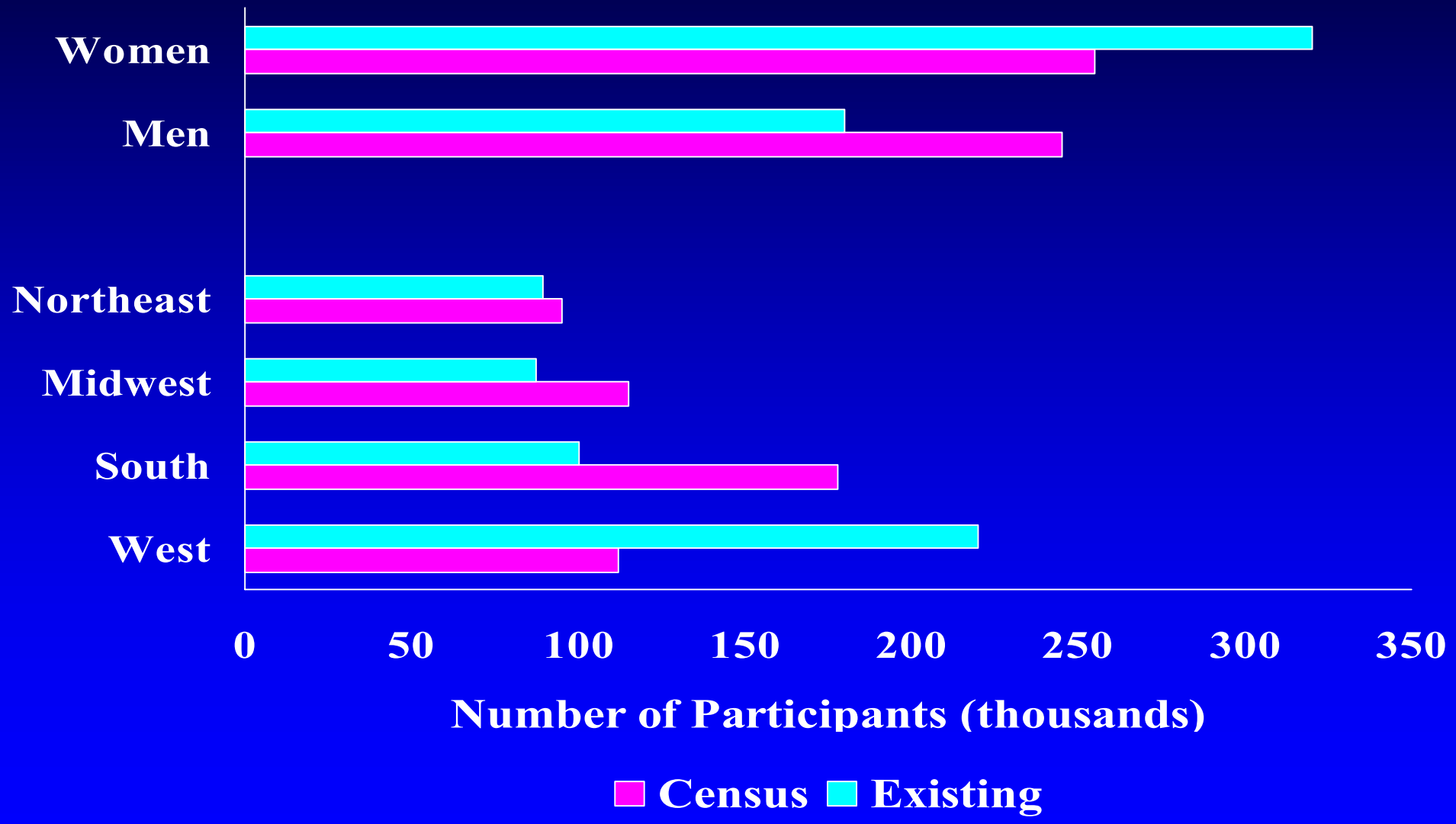
ESTIMATED AGE DISTRIBUTION OF REPRESENTATIVE US COHORT (2000 CENSUS)



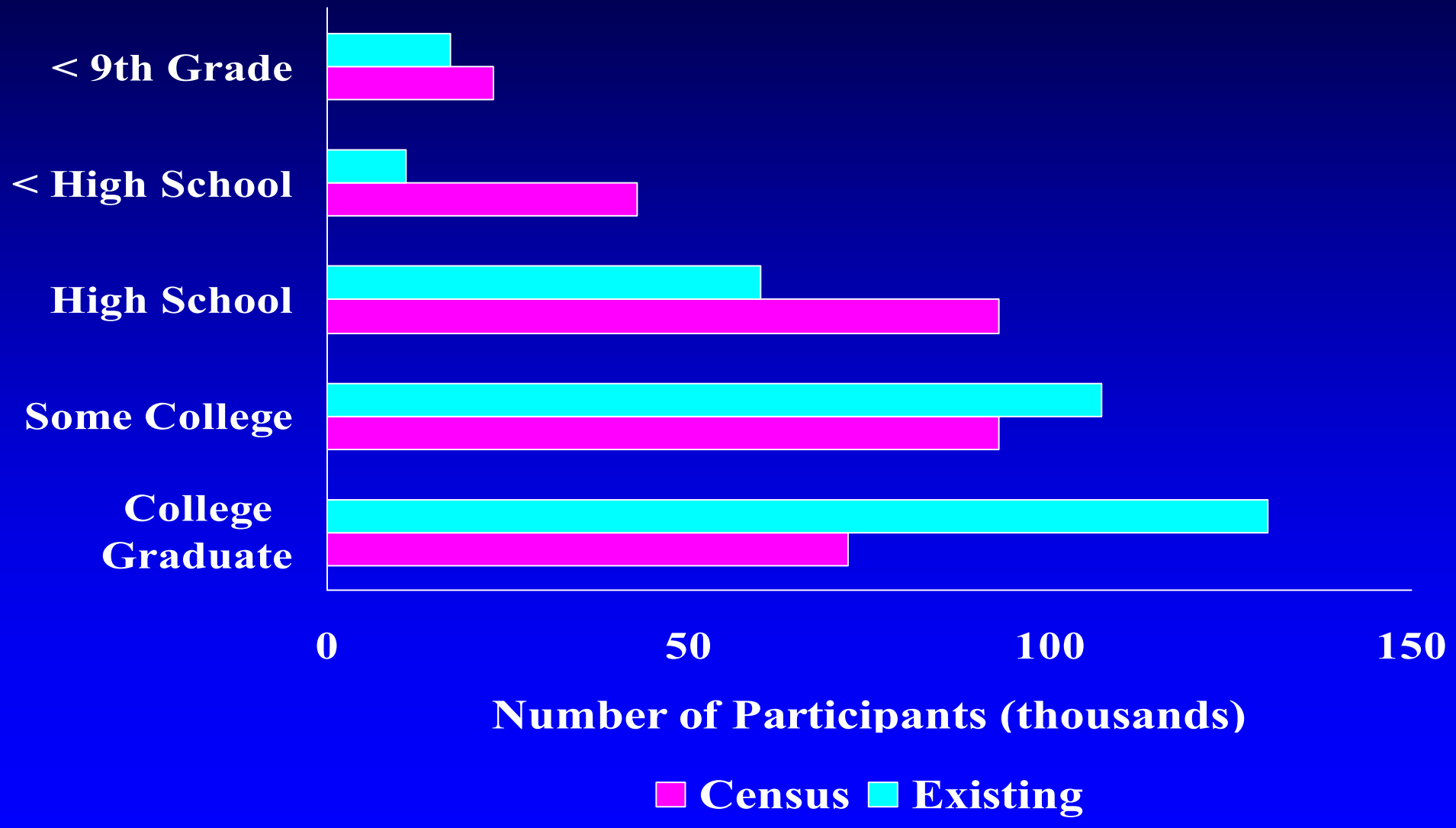
ESTIMATED AGE DISTRIBUTION OF EXISTING NIH-FUNDED COHORTS

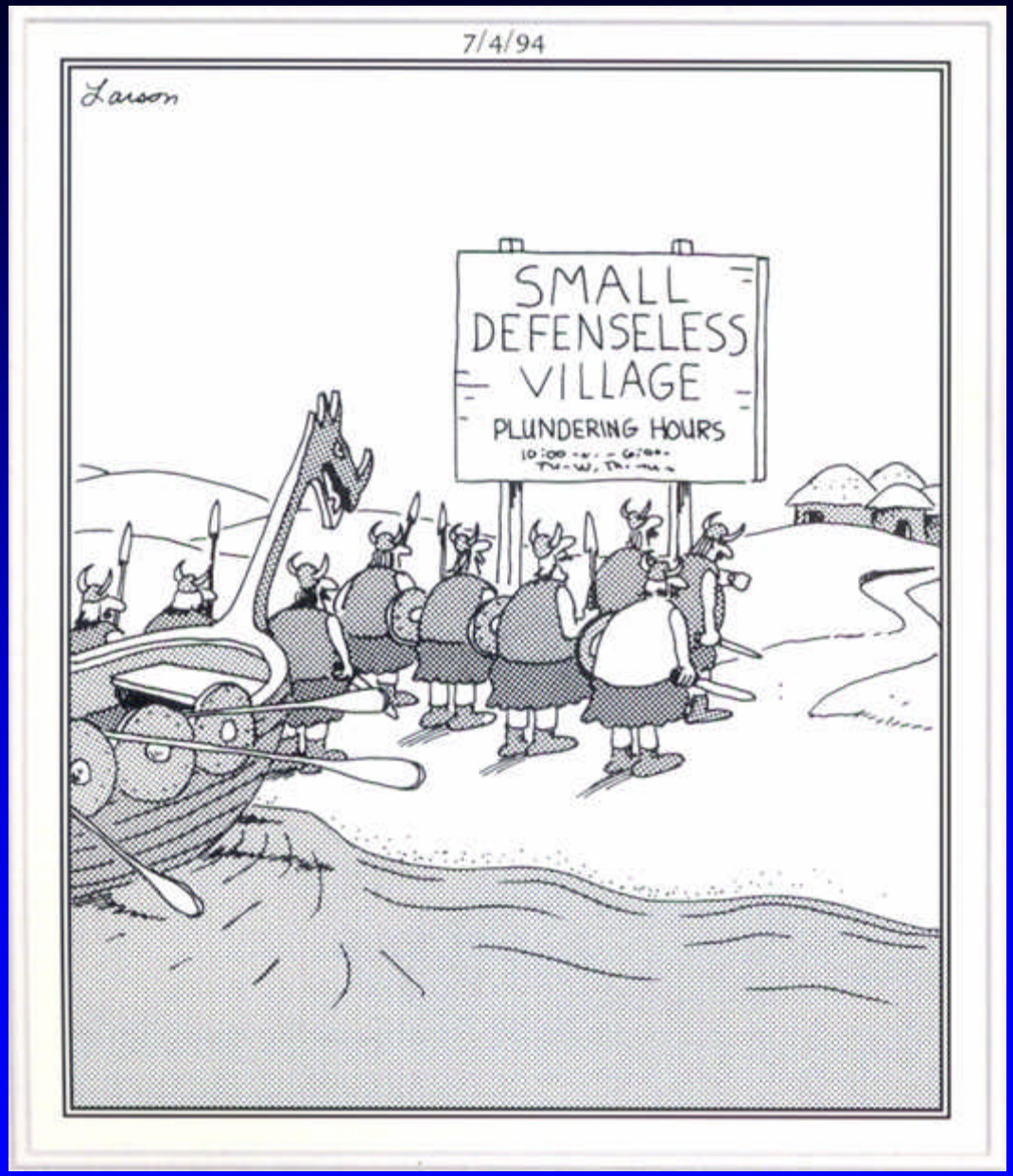


PROJECTED SEX AND REGIONAL DISTRIBUTION OF EXISTING COHORTS AND US CENSUS



PROJECTED EDUCATION DISTRIBUTION OF EXISTING COHORTS AND US CENSUS (Age ≥ 25)





Larson, G. *The Complete Far Side*. 2003.

COHORT STUDIES OF CANCER

Study	N	Age	Entry	Minorities
Adventist Health Study II	125,000	30+	2002	N/A
Cancer Prevention Study II	123,000	50-74	1992-93	N/A
California Teachers' Cohort	133,000w	30+	1995	87% White
Multiethnic Cohort Study	214,000	45-75	1993-96	Asian, Latino, Black, Hawaiian
Nurses' Health Study	121,700w	30-55	1976	97% White
Physicians' Health Study I and II	19,200m	40-84	1982; 1997	92% White
PLCO Screening Trial	150,000	55-74	1993	89% White
Southern Community Cohort Study	90,000	40-79	2001-	67% Black
Women's Health Initiative	161,800w	50-79	1993-98	18% multiple

DESIRABLE CHARACTERISTICS OF LARGE US COHORT STUDY

- Large sample size
- Full representation of minority groups
- Broad range of ages
- Broad range of genetic backgrounds and environmental exposures
- Family-based recruitment for at least part of the cohort to control for population stratification
- Broad array of clinical and laboratory data, regular follow up for events, additional exposure assessment

After Collins FS, *Nature* 2004; 429:475-477.

DESIRABLE CHARACTERISTICS OF LARGE US COHORT STUDY (continued)

- Technologically advanced dietary, lifestyle, and environmental exposure data
- Collection and storage of biological specimens
- Sophisticated data management system
- Access to materials and data by all researchers
- Goals should not be "hypothesis-limited"
- Comprehensive community engagement from the outset
- State of the art (?dynamic) consent to allow multiple uses of data and regular feedback to participants

After Collins FS, *Nature* 2004; 429:475-477.