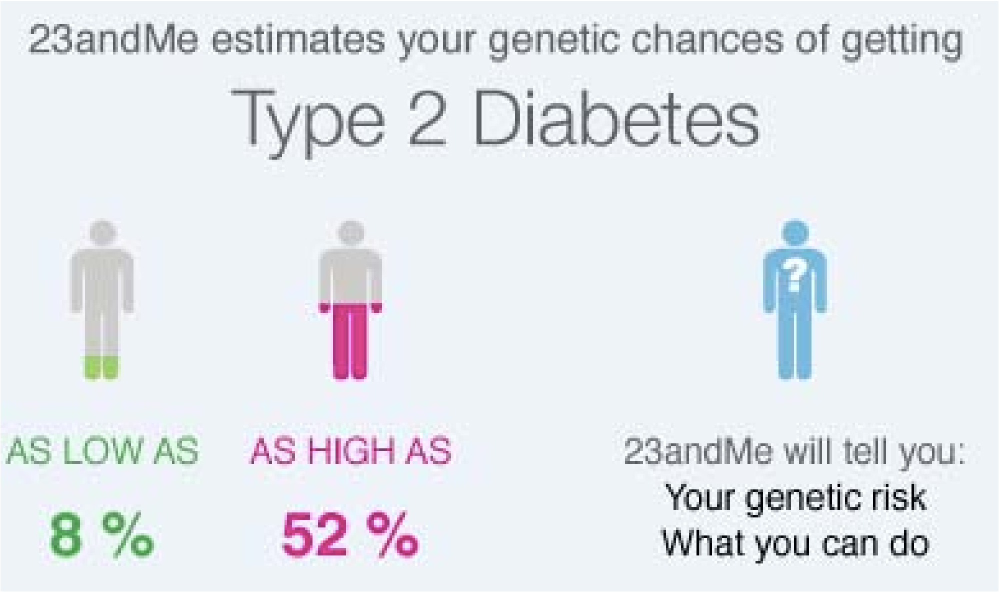
Challenges in genomic privacy

Sriram Sankararaman

Department of Computer Science
Department of Human Genetics
UCLA

Genomic data

Consumer genomics



23andMe estimates your genetic chances of getting

Type 2 Diabetes

AS LOW AS

8 %

AS HIGH AS

52 %

23andMe will tell you:
Your genetic risk
What you can do

GWAS

Biobanks

Geisinger mycode | 100,000+ PARTICIPANTS

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

biobank uk

Outline

Consumer genomics

GWAS

Clinical genomics and biobanks

Consumer genomics

AncestryDNA, 23andMe

~ 1million genotyped customers

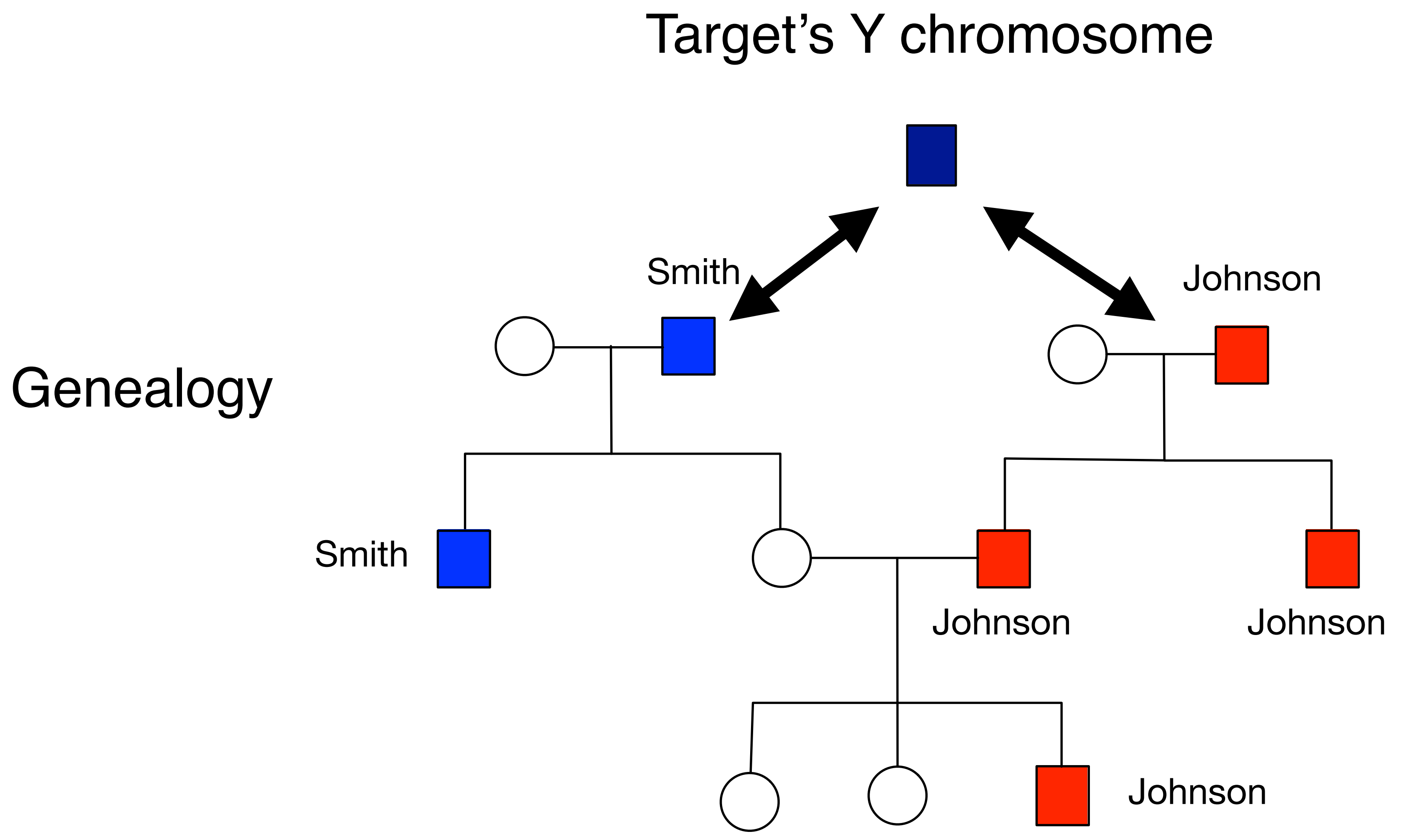
FamilyTreeDNA

~100K individuals

Identity tracing

Genealogical triangulation

De-identified genome to surname



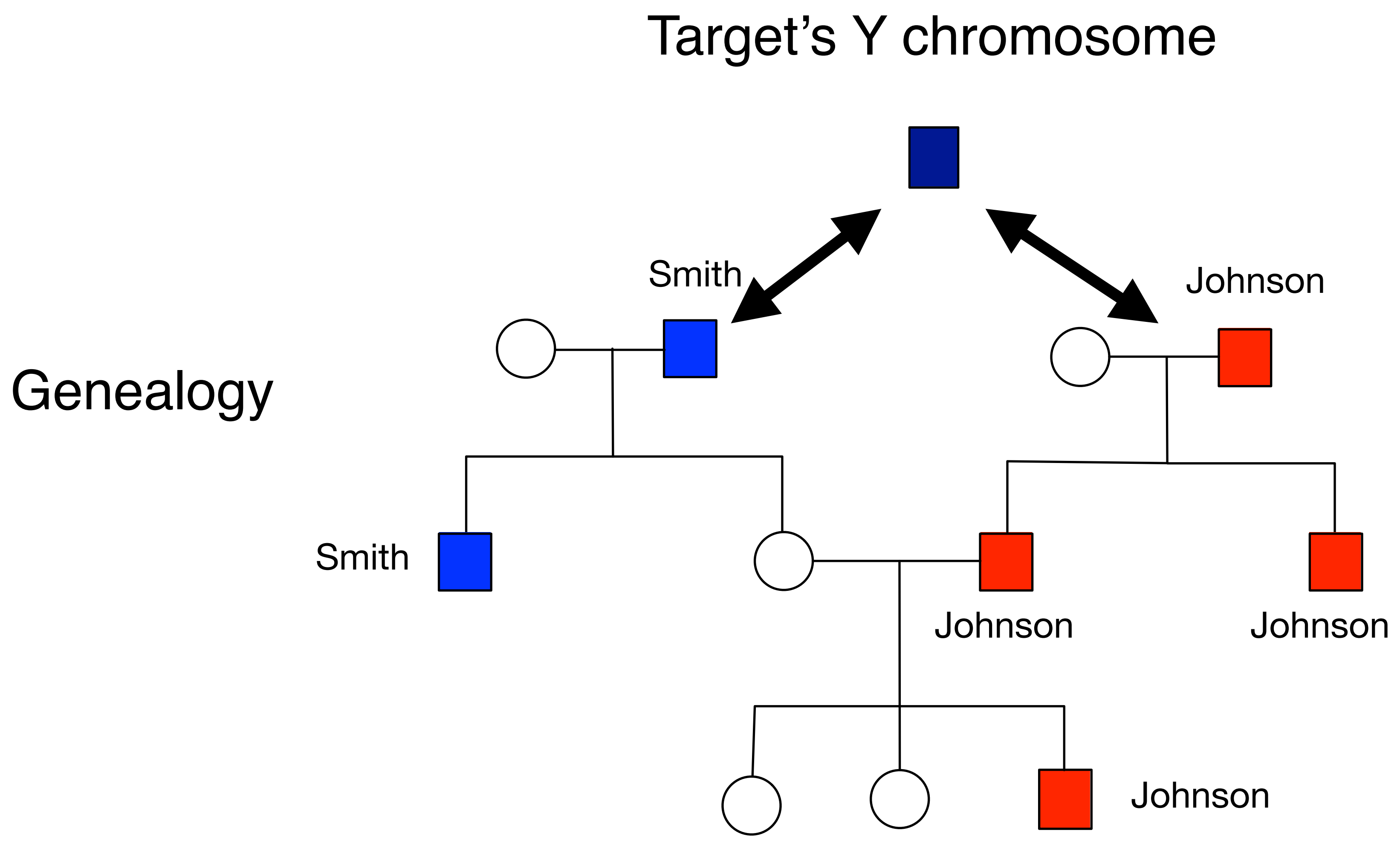
Gitschier et al. 2009

Identity tracing

Genealogical triangulation

Given query and potential match i , compute

$$t_i = argmax_t \Pr(\text{Mismatches}_i | \text{Generations} = t, \theta)$$



Gymrek et al. 2013

Identity tracing

Genealogical triangulation

Genealogical databases (~135K records, 39,000 unique records)

~12% of males can be correctly identified (83% unknown, 5% false positive rate)

Many surnames shared by less than <400,000 individuals (as informative as zip code)

Combine with year of birth and state of residency, gives median list of 12 individuals.

Gymrek et al. 2013

Questions

What happens with whole genome sequence vs Y chromosomes ?

Designing private queries

Outline

Consumer genomics

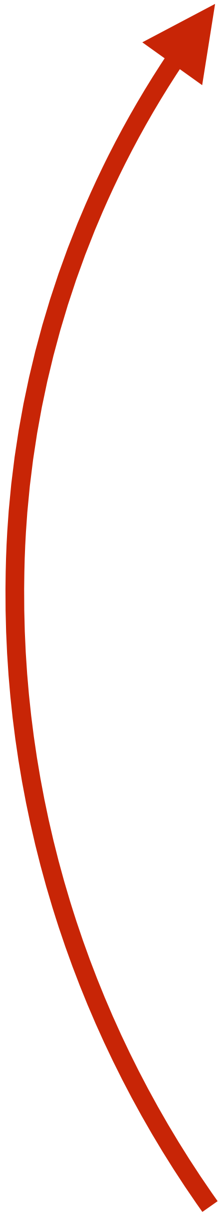
GWAS

Clinical genomics and biobanks

Genome-wide Association Studies (GWAS)

						SNPs						Phenotype
Individual	A	C	G	A	A	C	G	G	T	A	A	1
	C	C	G	G	T	C	G	G	T	C	T	1
	C	C	T	A	T	G	A	A	A	A	A	0
	A	T	G	A	A	G	G	G	T	A	T	0

GWAS pipeline



Data cleaning

- Remove outlier SNPs and individuals
- Impute missing data

Identify confounders

- Observed (Gender, Age)
- Unobserved (Ancestry): Needs to be inferred

Compute association statistics for a phenotype with each SNP j

$$Y \sim X_j + O + U$$

Compute model diagnostics

Replicate

GWAS have low power

Most SNP effects are weak

Strongest association for type-2 diabetes increase risk by ~20%

Large number of hypothesis tested (p-value < 5e-8 for statistical significance)

The missing heritability problem

For type-2 diabetes, current associations only explain ~12% of risk.

Sample size is key

Data sharing and meta-analyses

Opens up the possibility of re-identification attacks

Sample size is key

A two-tier access system for many repositories

Restricted access to individual-level genotype and phenotype data

Public access for summary statistics

Identification from summary statistics

SNP →

Cases ↓

0	1	1	0	0	0	1	0	0	1	0	1
0	1	1	1	0	0	0	0	1	0	1	0
0	1	1	1	0	0	1	0	0	1	0	0
1	1	0	0	0	0	1	1	0	0	0	1

0.25 1 0.75 0.5 0 .... 0.5 0.25 0.5

0 1 1 1 0 0 0 0 1 0 1 0 : Is this in the case ?

Controls ↓

0	1	0	0	0	0	1	0	0	1	1	0
0	1	0	1	0	0	0	0	1	1	0	0
0	1	0	1	1	1	1	1	0	1	0	1
0	1	0	0	1	0	1	1	0	0	0	1

0 1 0 0.5 0.5 .... 0.5 0.25 0.5

Homer et al, 2008 + many others

Identification also possible with other summary statistics

Marginal regression coefficients (Im et al. 2012) and their standard errors

Questions

Marginal regression coefficients

What is the lower bound in this setting if the summary statistics are distorted ?

What if the goal is to accurately predict the phenotype (and not the membership)?

Questions

DP in High-dimensional setting

How do we perform meta-analysis on DP estimates?

Two goals of GWAS : prediction vs discovering new associations (hypothesis testing)

Is one easier than the other ?

How do we build DP GWAS pipelines ?

Outline

Consumer genomics

GWAS

Clinical genomics and biobanks

Federated genomic datasets

Resides across multiple centers

Global Alliance for Genomics and Health (GA4GH)

- ~300 institutions with software interface that connects across these institutions

The Beacon project

Web service that allows multiple datasets that are registered in the GA4GH to be queried

Query: Do you have any genomes with an A at position 104,444 on chromosome 1 ?

Answer : Yes or No

Allows researchers to explore datasets for alleles of interest before they decide to apply for access

The Beacon project

Tracing attack on the beacons (Shringarpure et al. 2015)

Analogous to Homer et al. 2008

Questions

Potential for application of DP given the relatively small number of queries (compared to the larger number of SNPs released in GWAS)

Local versus global models of DP (each entity wants to participate without being open to breaches)

Other *-omics data

Gene expression

Schadt et al. 2012

Microbiome data

Franzosa et al. 2015

Electronic medical records

Loukides et al. 2010