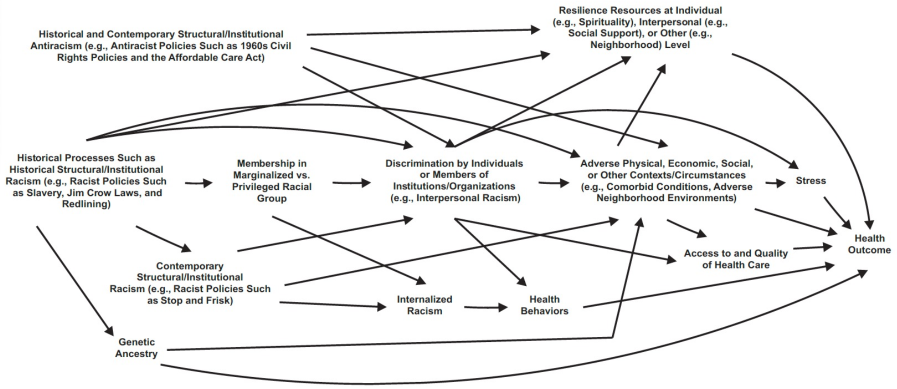
NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics

Discovering the causes of cancer and informing the means for prevention

Genetic and Environmental Factors in Cancer Epidemiology Cohorts and Consortia: Opportunities and Challenges

Peter Kraft

Director, Trans-Divisional Research Program
Division of Cancer Epidemiology and Genetics
National Cancer Institute



Historical and Contemporary Structural/Institutional Antiracism (e.g., Antiracist Policies Such as 1960s Civil Rights Policies and the Affordable Care Act)

Resilience Resources at Individual (e.g., Spirituality), Interpersonal (e.g., Social Support), or Other (e.g., Neighborhood) Level

Historical Processes Such as Historical Structural/Institutional Racism (e.g., Racist Policies Such as Slavery, Jim Crow Laws, and Redlining)

Membership in Marginalized vs. Privileged Racial Group

Discrimination by Individuals or Members of Institutions/Organizations (e.g., Interpersonal Racism)

Adverse Physical, Economic, Social, or Other Contexts/Circumstances (e.g., Comorbid Conditions, Adverse Neighborhood Environments)

Stress

Health Outcome

Access to and Quality of Health Care

Contemporary Structural/Institutional Racism (e.g., Racist Policies Such as Stop and Frisk)

Internalized Racism

Health Behaviors

Genetic Ancestry

“Gene-environment interaction” acknowledges the complex web of causality: interconnectedness, context

We can acknowledge complexity while studying the marginal impact of individual factors.

Howe (2022) Am J Epidemiol

Why study genes and environment?

- Leverage assumed effect modifiers to increase power
- Provide insights into biological mechanism
- Improve risk prediction and prognostic models

Kraft and Hunter (2010); Garcia-Closas et al. (2010)

Why study genes and environment?

- Leverage assumed effect modifiers to increase power
- Provide insights into biological mechanism
- Improve risk prediction and prognostic models

Kraft and Hunter (2010); Garcia-Closas et al. (2010)

Genome-wide GxE: what have we learned?

Trait	Exposure	Sample Size	Novel Loci	PMID
Pulmonary function	Smoking	50,000	3	
Blood pressure	Smoking	600,000	8	
Colorectal cancer	Diet	70,000	2	38749303
Colorectal cancer	Aspirin	70,000	2	38809988
Colorectal cancer	Folate	70,000	1	37640106
Colorectal cancer	Diabetes	70,000	2	37365285
Colorectal cancer	BMI	70,000	1	37249599
Colorectal cancer	Smoking	70,000	3	
Breast cancer	7 risk factors	150,000	2	37559094

For comparison, the number of loci identified via marginal tests:
blood pressure, 136; colorectal cancer, 205; breast cancer, 250.

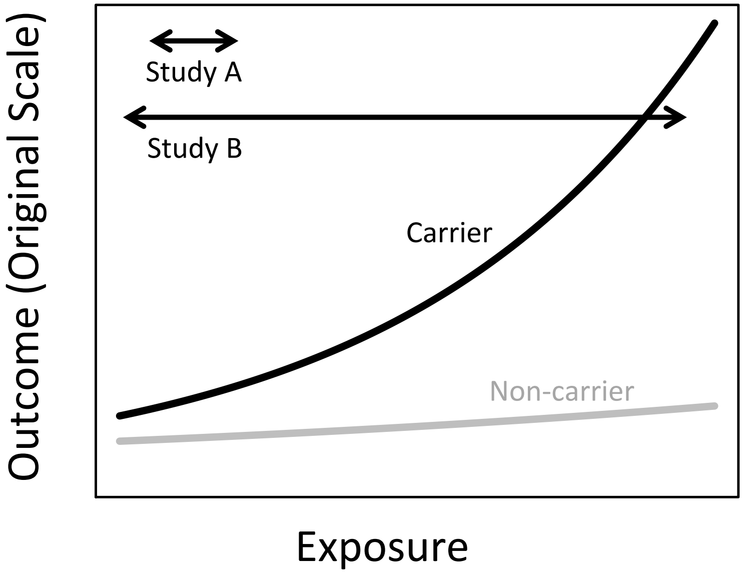
Colon cancer: 30,000 cases, 40,000 controls, breast cancer: 70,000 cases, 80,000 controls

Why so few loci?

- Limited sample sizes
- Measurement error
- Limited diversity

Why so few loci?

- Limited sample sizes
- Measurement error
- Limited diversity



Increasing exposure range can increase power to detect GxE interactions

Kraft and Aschard (2013) Eur J Epidemiol

Slide courtesy of L Mechanic

FTO, Physical Activity and BMI

Kilpelainen et al. (2011). PLoS Medicine. 8(11). e1001116

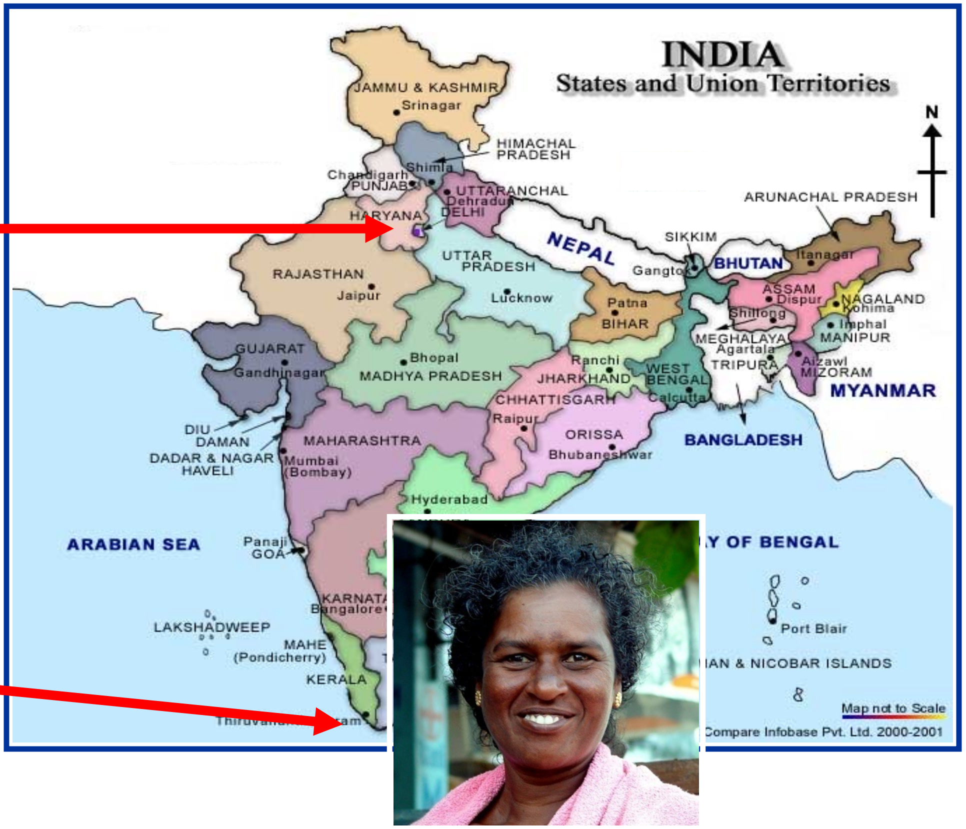
- Meta-analysis of 218,166 European-ancestry subjects
- Risk of Obesity (BMI ≥ 30 vs. BMI < 25 kg/m^2) for *FTO* rs9939609

	OR (95% CI)
rs9939609: Inactive	1.30 (1.24-1.36)
rs9939609: Active	1.22 (1.19-1.25)
Interaction	0.92 (0.88-0.97)
	P-value = 0.0010

Slides courtesy of N Chatterjee

India health study

New Delhi



Trivandrum

Participant characteristics by region

Characteristic	New Delhi	Trivandrum
Total (n=1,313)	n=619	n=694
Age, years (mean, SD)	47.4 ± 10.0	48.8 ± 9.2
Household monthly income, %		
<5,000 rupees	7.1	71.9
>10,000 rupees	76.7	3.1
Household items, %		
Car	25	7
Refrigerator	87	58
Washing machine	79	14
Total physical activity, MET-hr/wk	42.5 ± 43.8	147.3 ± 85.2
Vigorous physical activity, MET-hr/wk	0.6 ± 6.8	26.2 ± 51.4
Sitting, hr/day	10.4 ± 2.0	5.0 ± 2.3
Centrally obese, %	82.1	60.2

Moore (2011) Obesity

Association of *FTO* rs3751812 with waist circumference

Characteristic	N	Effect size per T allele (95% CI)	P_{trend}	Interaction by PA
Overall	1,209	+1.61 cm (0.67, 2.55)	0.0008	
New Delhi				
Overall	578	+2.53 cm (1.08, 3.97)	0.0006	
By PA				
≤ 91 MET-hrs/wk	517	**+2.36 cm (0.82, 3.89)**	0.003	
92-151 MET-hrs/wk	32	**+6.39 cm (1.94, 10.85)**	0.005	
152-217 MET-hrs/wk	24	**-0.95 cm (-7.33, 5.42)**	0.77	
218+ MET-hrs/wk	5	**N/A**	N/A	
Trivandrum				
Overall	574	+0.87 cm (-0.35, 2.08)	0.16	
By PA				
≤ 91 MET-hrs/wk	170	**+3.50 cm (0.90, 6.10)**	0.008	
92-151 MET-hrs/wk	132	**+1.13 cm (-1.08, 3.33)**	0.32	
152-217 MET-hrs/wk	141	**+1.04 cm (-1.63, 3.70)**	0.45	
218+ MET-hrs/wk	131	**-2.32 cm (-4.82, 0.18)**	0.07	

Moore (2011) Obesity

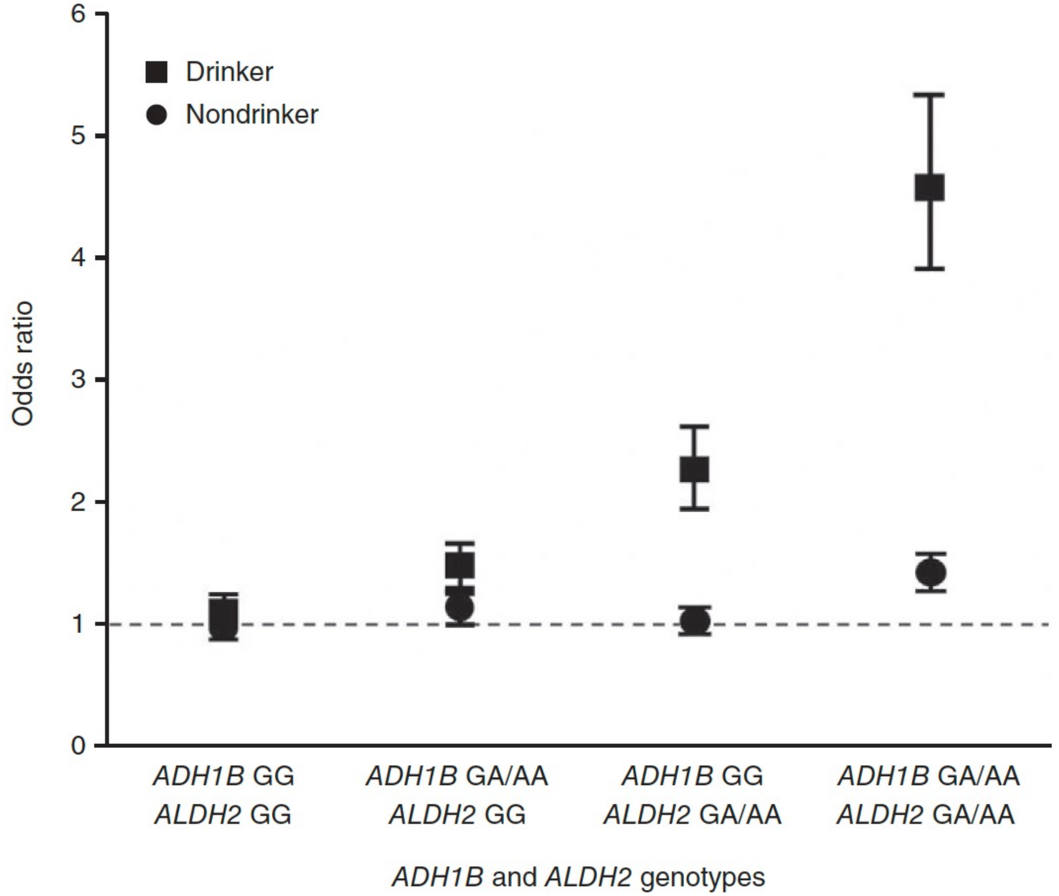


Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

Interaction between alcohol intake, variants in alcohol metabolism genes, and esophageal cancer risk in Chinese GWAS participants.

Cannot be studied in many other populations due to rarity of the rs11066015 A allele.

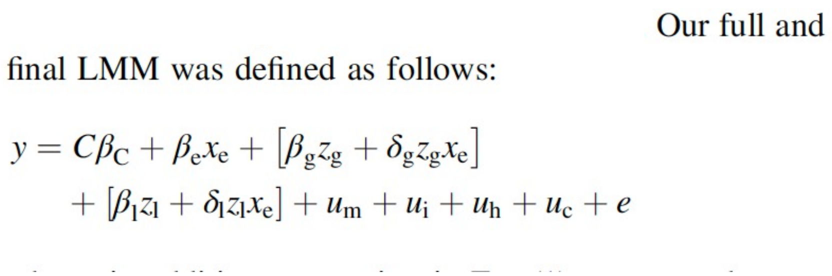
Wu (2012) Nat Genet

European Journal of Human Genetics (2020) 28:656–668
https://doi.org/10.1038/s41431-019-0545-8

ARTICLE

Mixed-model admixture mapping identifies smoking-dependent loci of lung function in African Americans

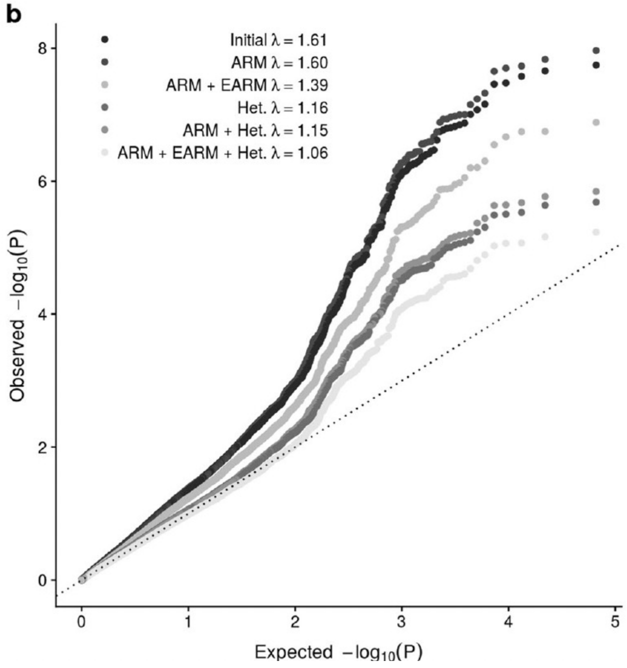
Andrey Ziyatdinov[1] · Margaret M. Parker[2] · Amaury Vaysse[3] · Terri H. Beaty[4] · Peter Kraft[1] · Michael H. Cho[2,5] · Hugues Aschard[1,3]



Our full and final LMM was defined as follows:

$$y = C\beta_C + \beta_e x_e + \left[\beta_g z_g + \delta_g z_g x_e\right] + \left[\beta_l z_l + \delta_l z_l x_e\right] + u_m + u_i + u_h + u_c + e$$

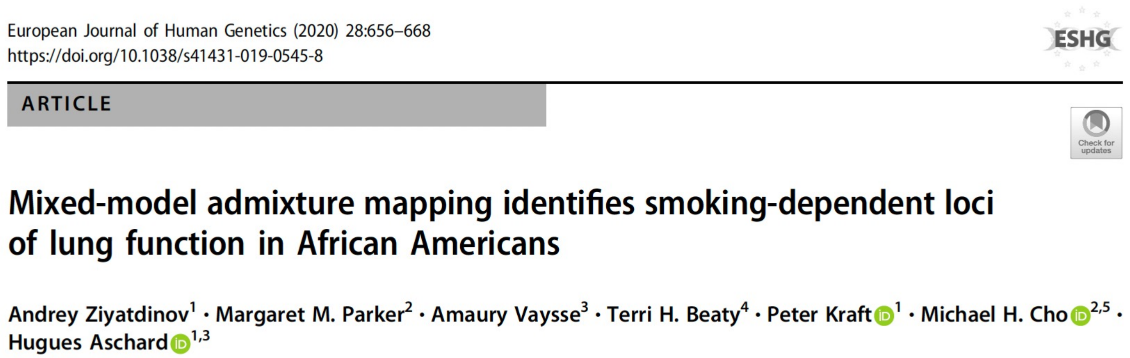
Model tests for local ancestry haplotypic effects, allowing for effect differences by E, while adjusting for fixed effects of E, global genetic similarity, and random effects for genetic similarity and heterogeneity in variance across exposures and study site.



QQ plots for local ancestry x E interaction tests.

Care is needed lest main effects bleed over to interaction estimates.

Ziyatdinov (2020) EJHG



European Journal of Human Genetics (2020) 28:656–668
https://doi.org/10.1038/s41431-019-0545-8

ARTICLE

Mixed-model admixture mapping identifies smoking-dependent loci of lung function in African Americans

Andrey Ziyatdinov[1] · Margaret M. Parker[2] · Amaury Vaysse[3] · Terri H. Beaty[4] · Peter Kraft[1] · Michael H. Cho[2,5] · Hugues Aschard[1,3]

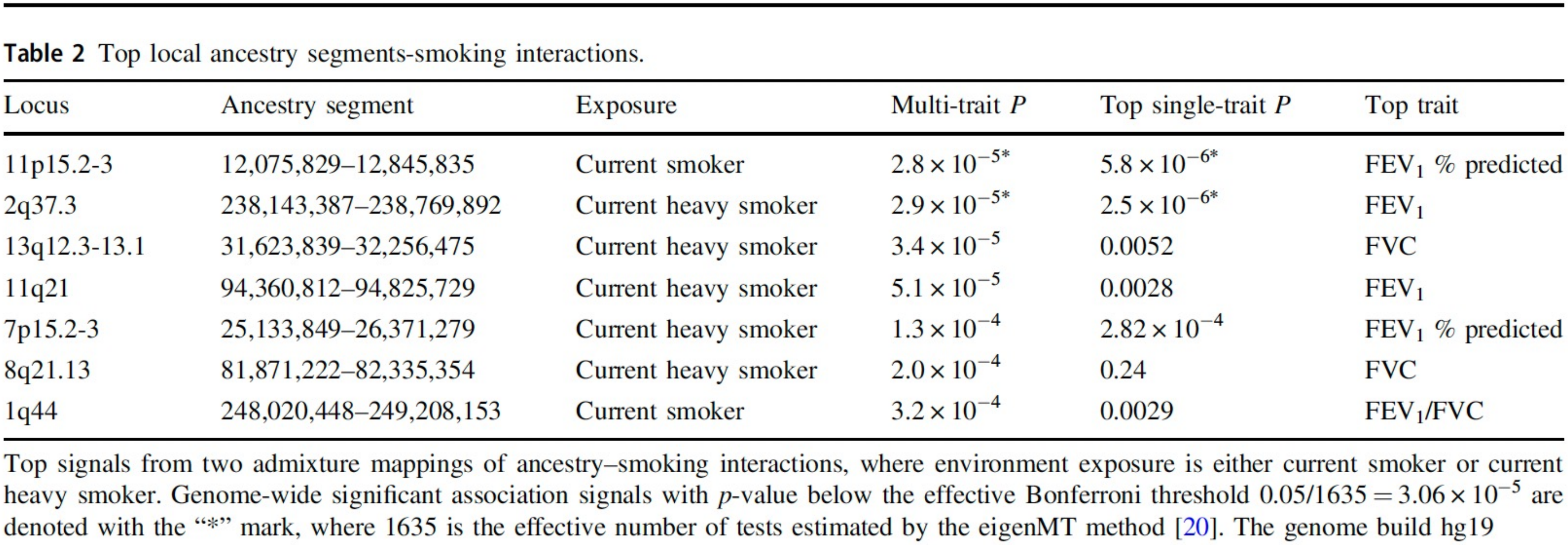


Table 2 Top local ancestry segments-smoking interactions.

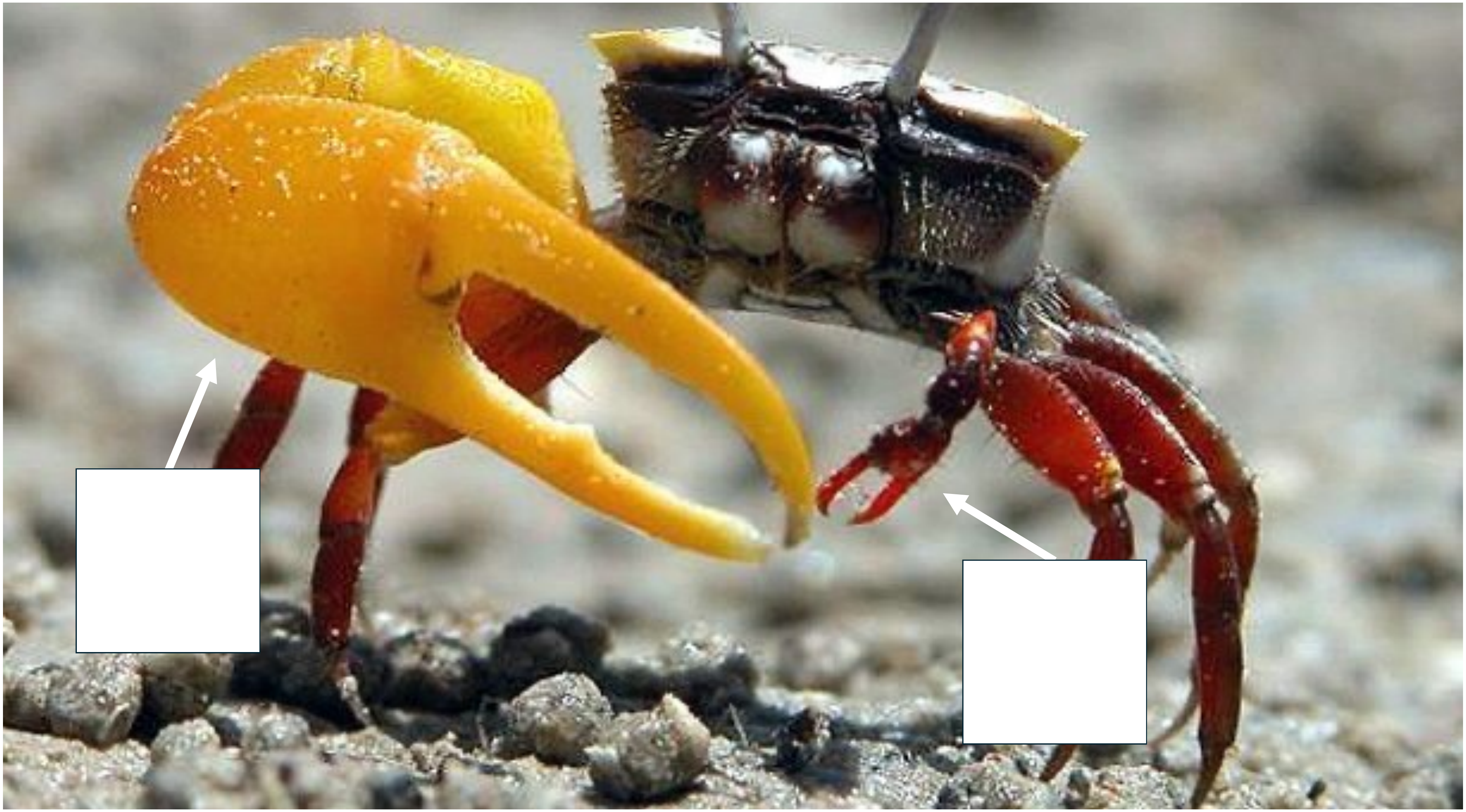
Locus	Ancestry segment	Exposure	Multi-trait P	Top single-trait P	Top trait
11p15.2-3	12,075,829–12,845,835	Current smoker	$2.8 \times 10^{-5*}$	$5.8 \times 10^{-6*}$	FEV_1 % predicted
2q37.3	238,143,387–238,769,892	Current heavy smoker	$2.9 \times 10^{-5*}$	$2.5 \times 10^{-6*}$	FEV_1
13q12.3-13.1	31,623,839–32,256,475	Current heavy smoker	3.4×10^{-5}	0.0052	FVC
11q21	94,360,812–94,825,729	Current heavy smoker	5.1×10^{-5}	0.0028	FEV_1
7p15.2-3	25,133,849–26,371,279	Current heavy smoker	1.3×10^{-4}	2.82×10^{-4}	FEV_1 % predicted
8q21.13	81,871,222–82,335,354	Current heavy smoker	2.0×10^{-4}	0.24	FVC
1q44	248,020,448–249,208,153	Current smoker	3.2×10^{-4}	0.0029	FEV_1/FVC

Top signals from two admixture mappings of ancestry–smoking interactions, where environment exposure is either current smoker or current heavy smoker. Genome-wide significant association signals with p-value below the effective Bonferroni threshold $0.05/1635 = 3.06 \times 10^{-5}$ are denoted with the "*" mark, where 1635 is the effective number of tests estimated by the eigenMT method [20]. The genome build hg19

Ziyatdinov (2020) EJHG

Why so few loci?

- Limited sample sizes
- Measurement error
- Limited diversity



connect

for cancer prevention study

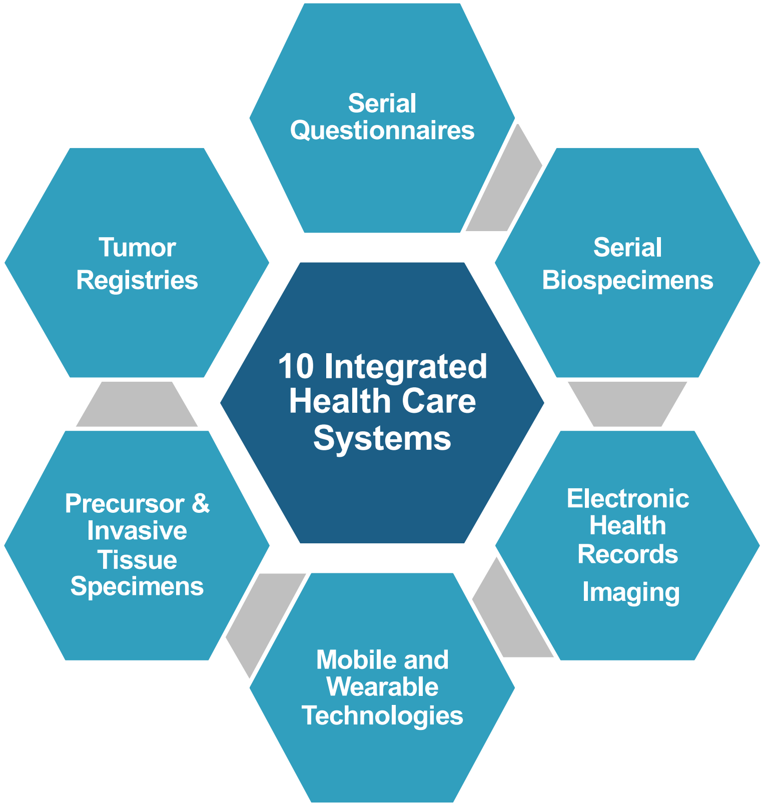
Connect today. Prevent cancer tomorrow.

Mia
Gaudet

https://www.cancer.gov/connect-prevention-study/
https://dceg.cancer.gov/research/who-we-study/cohorts/connect

Key Study Design Features of Connect

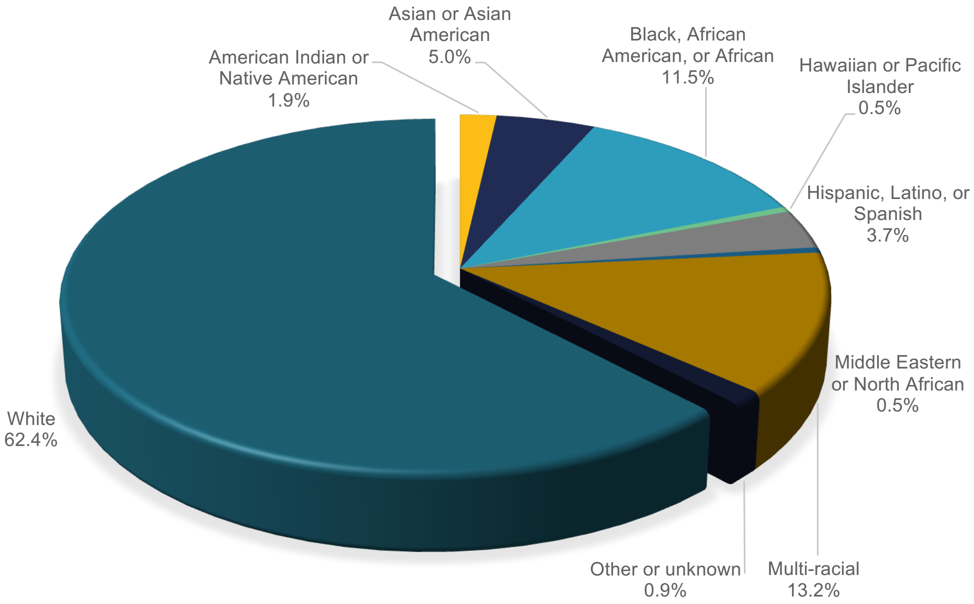
- 200,000 adults across the US
  - ✓ Aged 30-70 years
  - ✓ No history of cancer
  - ✓ Patients or members of partner health care systems
- Defined catchment population
- Survey and EHR data
- Comprehensive cancer and precancer outcomes
- Flexible infrastructure for enhancement studies


Serial Questionnaires
Tumor Registries
Serial Biospecimens
10 Integrated Health Care Systems
Precursor & Invasive Tissue Specimens
Electronic Health Records
Imaging
Mobile and Wearable Technologies

connect
for cancer prevention study

Demographic Distributions of ~30,000 Study Participants*

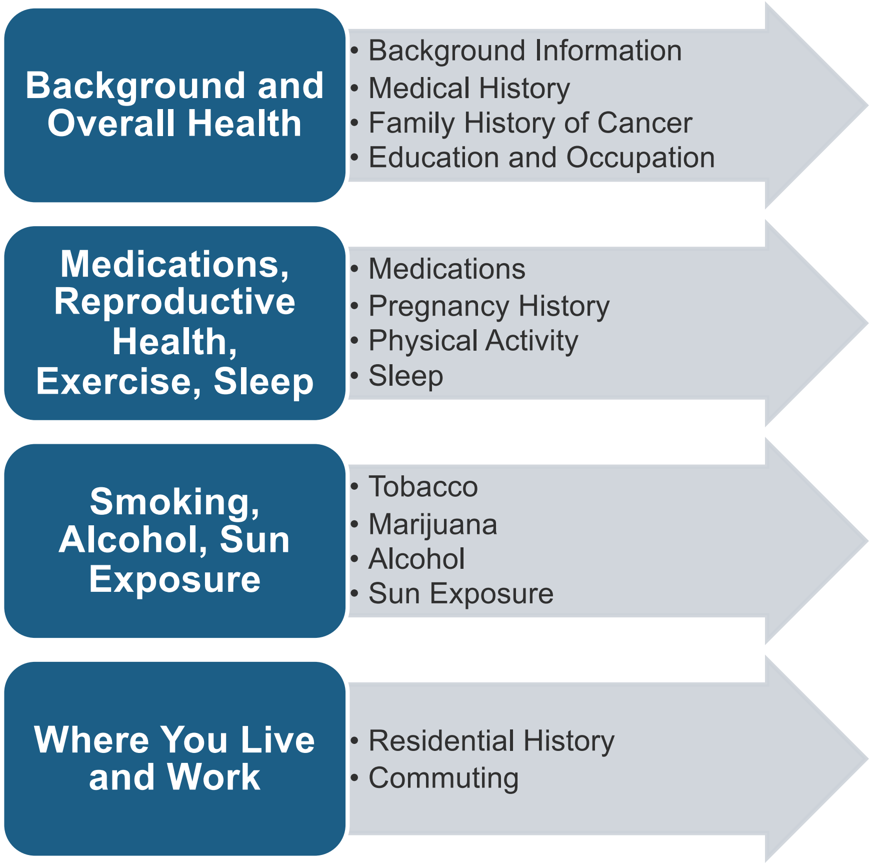
Demographic Factors	%
Males	31.7
Gender minorities	0.9
Sexual minorities	9.7
High school or less	9.1
Income, <$35,000	11.8



connect for cancer prevention study

*As of April '24. Currently over 50,000 enrolled.

Baseline Surveys

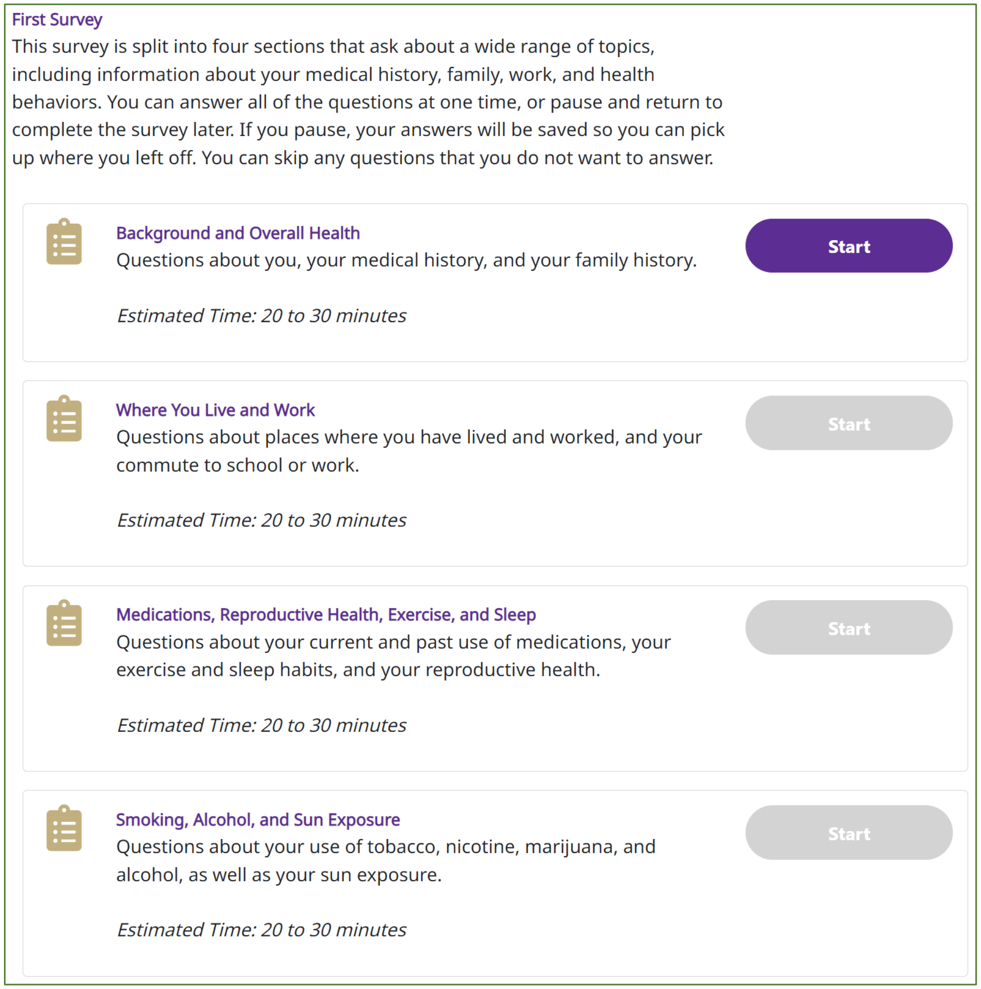


Background and Overall Health
- Background Information
- Medical History
- Family History of Cancer
- Education and Occupation

Medications, Reproductive Health, Exercise, Sleep
- Medications
- Pregnancy History
- Physical Activity
- Sleep

Smoking, Alcohol, Sun Exposure
- Tobacco
- Marijuana
- Alcohol
- Sun Exposure

Where You Live and Work
- Residential History
- Commuting



First Survey

This survey is split into four sections that ask about a wide range of topics, including information about your medical history, family, work, and health behaviors. You can answer all of the questions at one time, or pause and return to complete the survey later. If you pause, your answers will be saved so you can pick up where you left off. You can skip any questions that you do not want to answer.

Background and Overall Health

Questions about you, your medical history, and your family history.

Estimated Time: 20 to 30 minutes

Where You Live and Work

Questions about places where you have lived and worked, and your commute to school or work.

Estimated Time: 20 to 30 minutes

Medications, Reproductive Health, Exercise, and Sleep

Questions about your current and past use of medications, your exercise and sleep habits, and your reproductive health.

Estimated Time: 20 to 30 minutes

Smoking, Alcohol, and Sun Exposure

Questions about your use of tobacco, nicotine, marijuana, and alcohol, as well as your sun exposure.

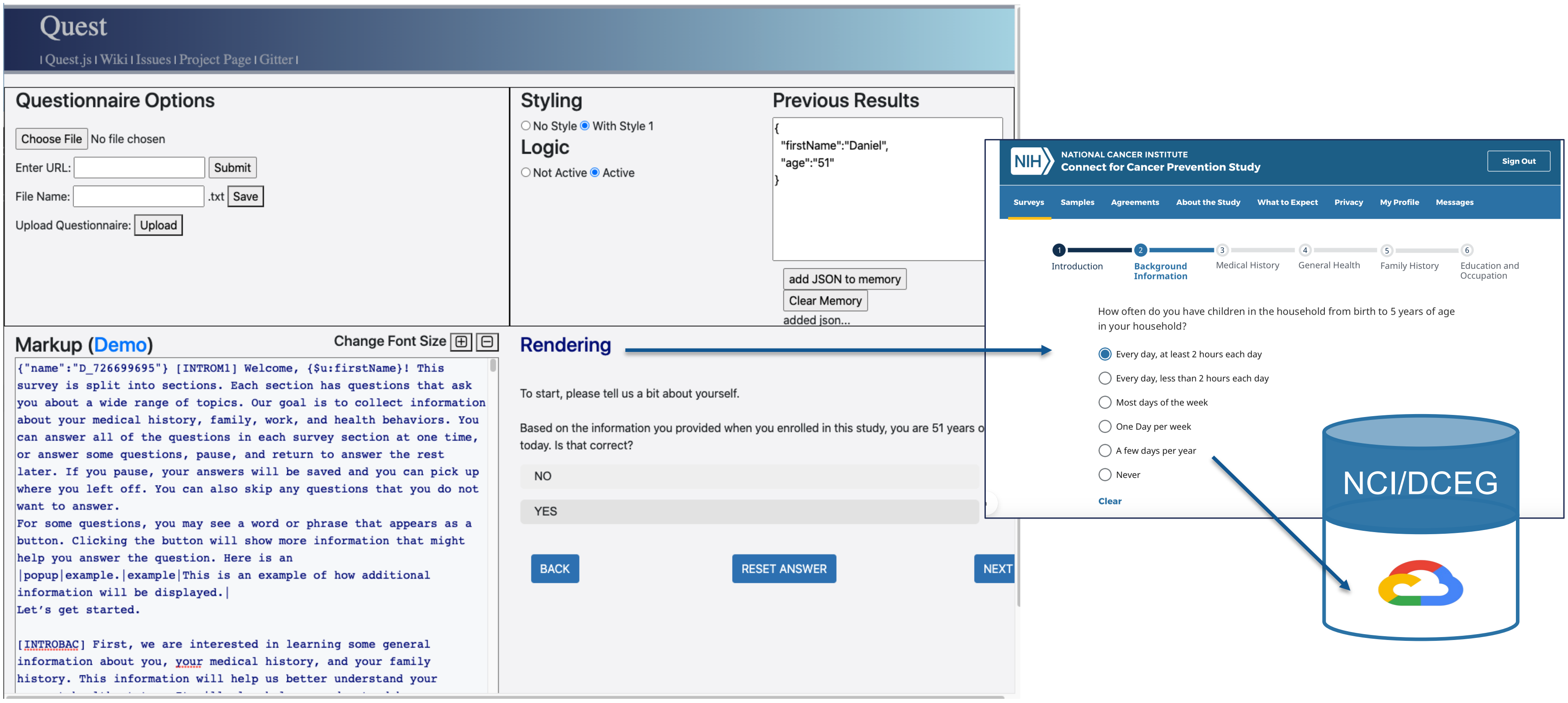
Estimated Time: 20 to 30 minutes

Content available on Connect GitHub

Surveys In Development

Cancer Screening History	Organ Inventory (born with/current), history of cancer screening tests
Cancer Diagnosis Surveys (17 sites)	Dx, symptoms, Patient-Provider interaction, medical history repeat assessment
Menstrual & Intimate Care Products	Vaginoplasty, powder, douching, vaginal cleansing products, menstrual products
Fecal Collection	Donation, bowel movements, meds, supplements, probiotics
Social Determinants of Health (SDOH)	Discrimination, police interaction, medical mistrust, social support, and financial, food, and housing insecurity
Hair Products	Dyes, relaxers, straighteners, perms, oils
Menstrual Experience Survey	Menstrual problems, endometriosis dx and treatment
Mothballs & Scented Products	Household exposure to p-DCB & Napthalene

Open-Source Code Available Now: Quest render surveys into progressive web applications



```
{
 "firstName":"Daniel",
 "age":"51"
}
```

```
{"name":"D_726699695"} [INTROM1] Welcome, {$u:firstName}! This
survey is split into sections. Each section has questions that ask
you about a wide range of topics. Our goal is to collect information
about your medical history, family, work, and health behaviors. You
can answer all of the questions in each survey section at one time,
or answer some questions, pause, and return to answer the rest
later. If you pause, your answers will be saved and you can pick up
where you left off. You can also skip any questions that you do not
want to answer.
For some questions, you may see a word or phrase that appears as a
button. Clicking the button will show more information that might
help you answer the question. Here is an
|popup|example.|example|This is an example of how additional
information will be displayed.|
Let's get started.

[INTROBAC] First, we are interested in learning some general
information about you, your medical history, and your family
history. This information will help us better understand your
```

To start, please tell us a bit about yourself.

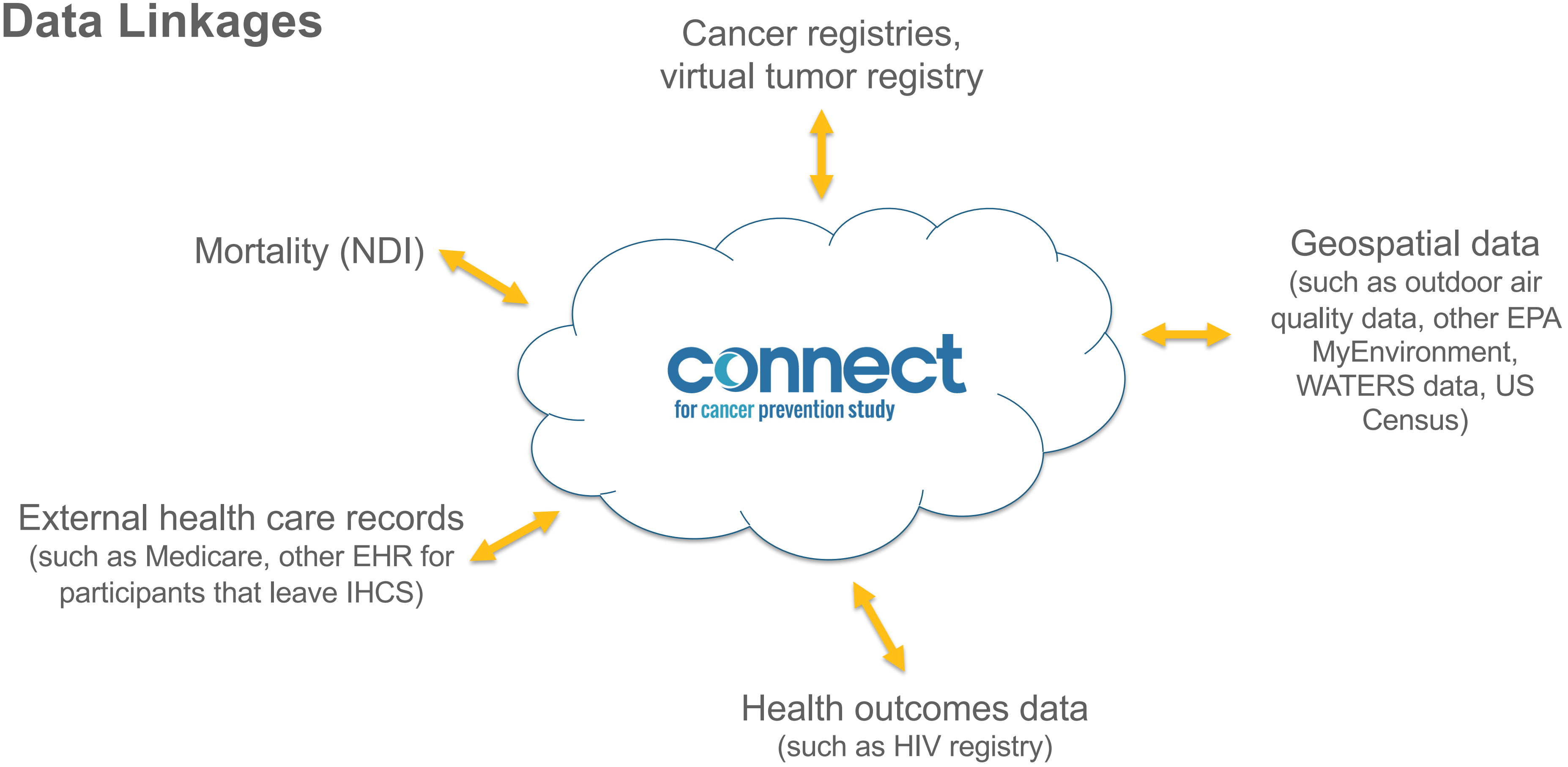
Based on the information you provided when you enrolled in this study, you are 51 years o today. Is that correct?

How often do you have children in the household from birth to 5 years of age in your household?

● Every day, at least 2 hours each day
○ Every day, less than 2 hours each day
○ Most days of the week
○ One Day per week
○ A few days per year
○ Never

Code available on GitHub episphere/connect.

Data Linkages



Cancer registries, virtual tumor registry

Mortality (NDI)

Geospatial data (such as outdoor air quality data, other EPA MyEnvironment, WATERS data, US Census)

External health care records (such as Medicare, other EHR for participants that leave IHCS)

Health outcomes data (such as HIV registry)

Connect Resource Access Principles

- Research resource for scientific community
- Broad data sharing policies
- Participant privacy and confidentiality
- F.A.I.R. data infrastructure

Target data release: 2026

Cross-study collaborations still necessary

- Assemble the large sample sizes
- Assess heterogeneity due to design or context
- Leverage existing resources and “Let 100 Flowers Bloom”

Challenges to cross-study collaborations

- Effective governance and data custodianship
- Data interoperability
- Limitations to real world data (sampling and measurement)

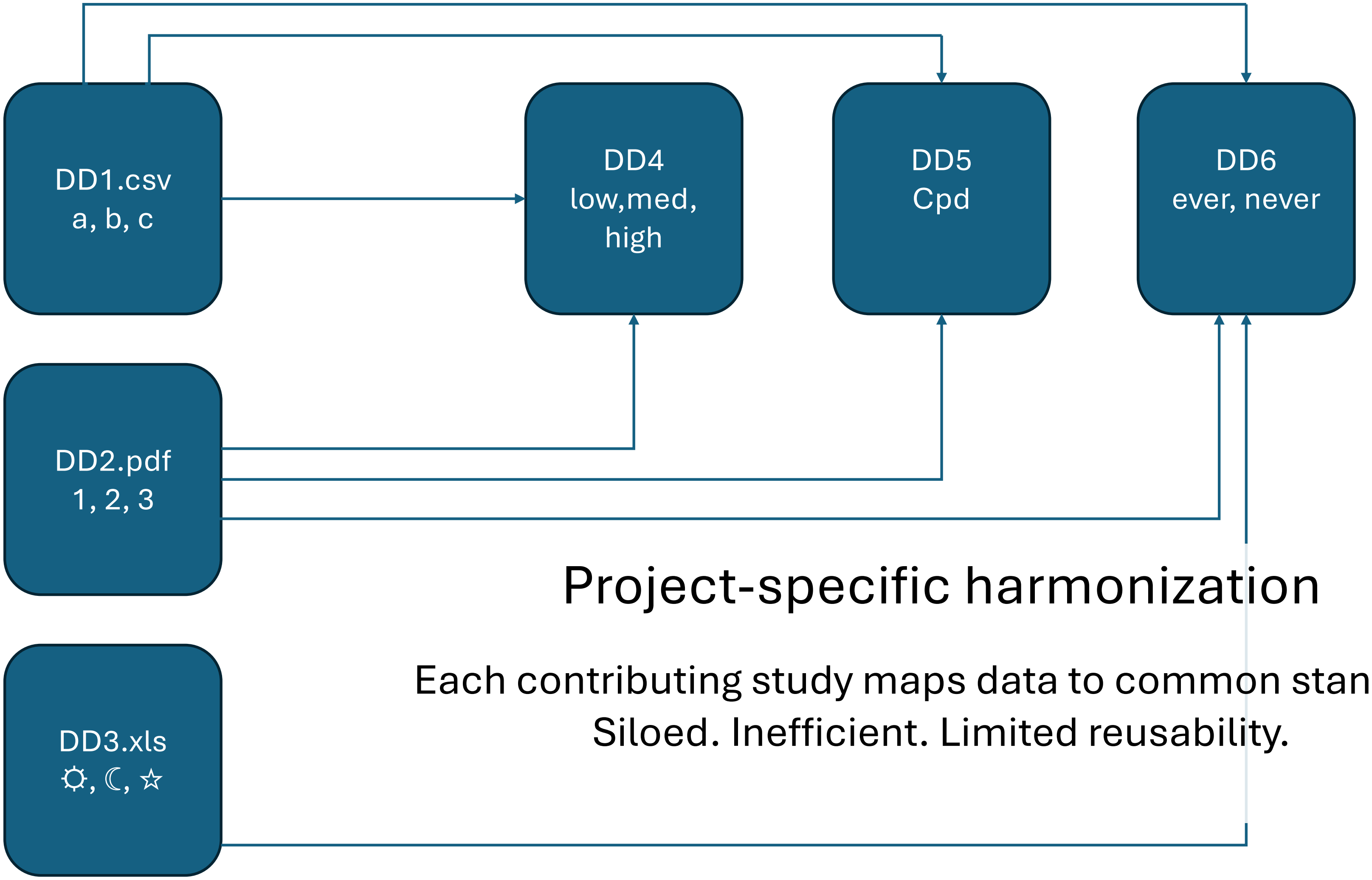
DD1.csv
a, b, c

DD2.pdf
1, 2, 3

DD3.xls
☼, ☾, ☆

Data dictionaries among studies have similar content but variable structure

Hard to search and sort
Barrier to harmonization within and between studies



DD1.csv
a, b, c

DD2.pdf
1, 2, 3

DD3.xls
☼, ☾, ☆

DD4
low,med,
high

DD5
Cpd

DD6
ever, never

Project-specific harmonization

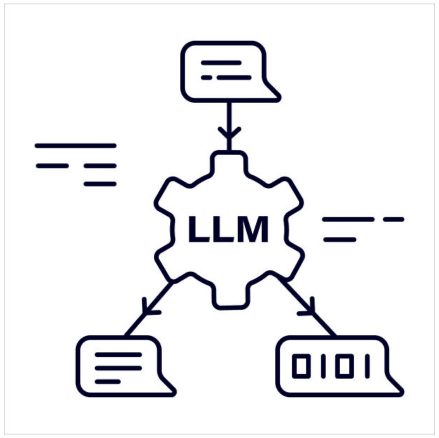
Each contributing study maps data to common standard.
Siloed. Inefficient. Limited reusability.

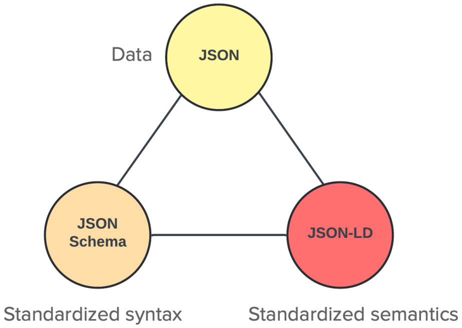
DD1.csv
a, b, c

DD2.pdf
1, 2, 3

DD3.xls
☼, ☾, ☆

Automated Harmonization





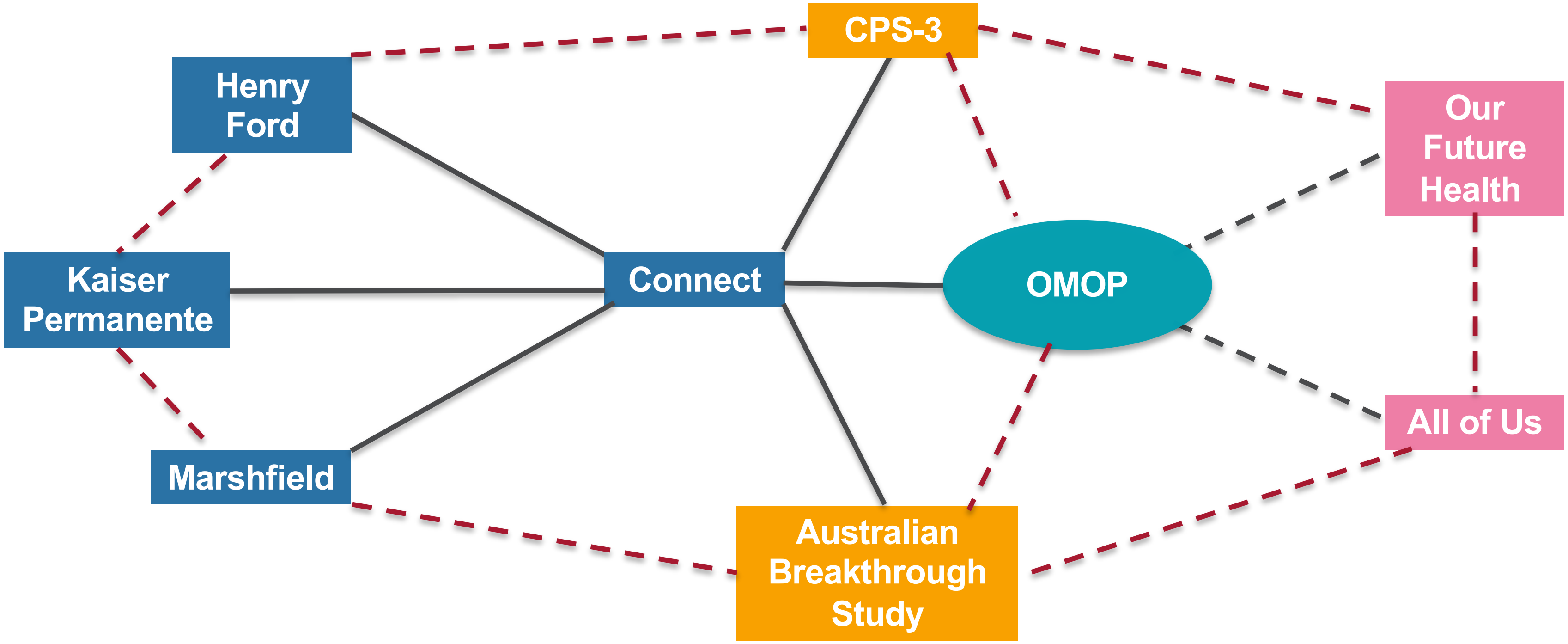
Vocabulary Mapping

Data Transform

QA/QC

Jeya Balasubramanian

Mapping to a common data model facilitates interoperability, accelerates collaboration



Nicole Gerlanc

Why study genes and environment?

- Leverage assumed effect modifiers to increase power
- Provide insights into biological mechanism
- Improve risk prediction and prognostic models

Kraft and Hunter (2010); Garcia-Closas et al. (2010)

Paths forward

- Increase sample sizes, facilitate cross-study collaborations
- More and more detailed exposure measurements
- Increase participant diversity

NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics

Discovering the causes of cancer and informing the means for prevention

Thank You!

https://dceg.cancer.gov/