# Ontario Institute for Cancer Research

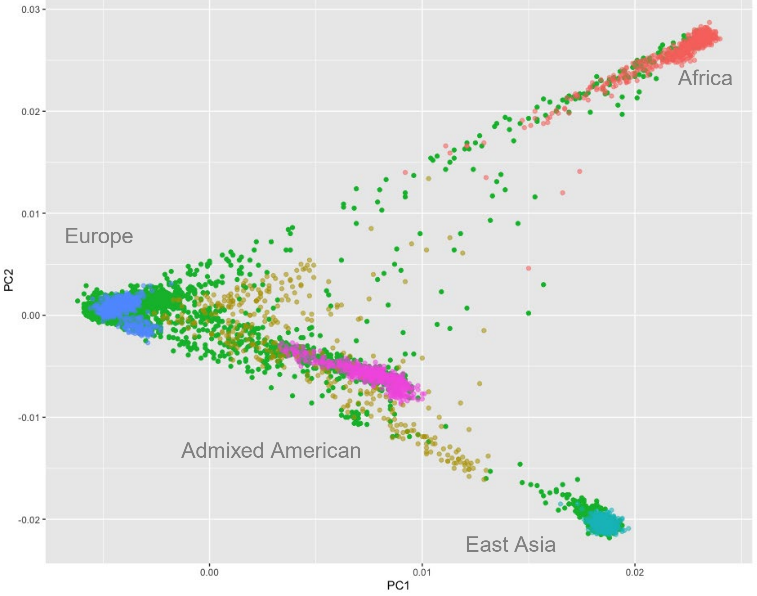
## From Genes to Environment: Exploring Disease Risk Factors through large Population Cohorts

Marie-Julie Favé, PhD

OICR
Ontario Institute
for Cancer Research

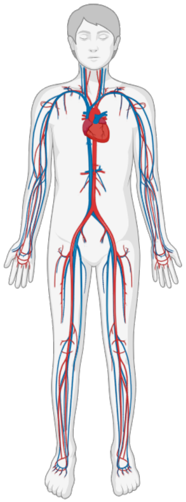
# CanPath genetic diversity as of July 2024 (OHS and CaG)

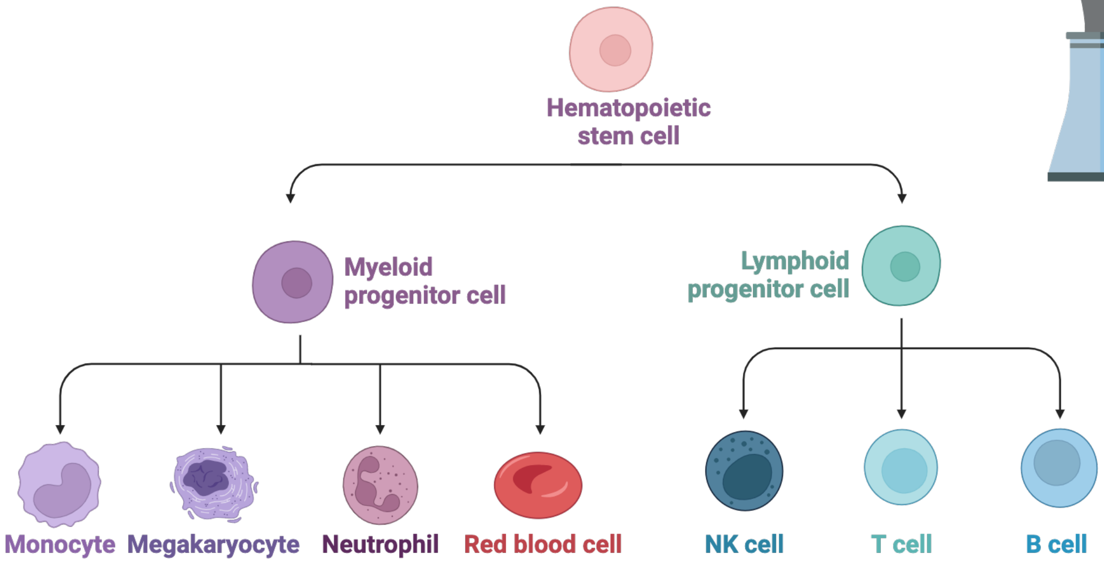

Ontario


Québec

PC plots include individuals from 1000 Genomes project for ancestry inference

# Hematopoiesis: The making of blood and circulating immune cells





Hematopoietic stem cell

Myeloid progenitor cell

Lymphoid progenitor cell

Monocyte

Megakaryocyte

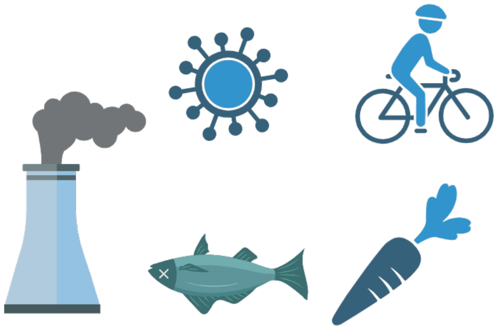
Neutrophil

Red blood cell

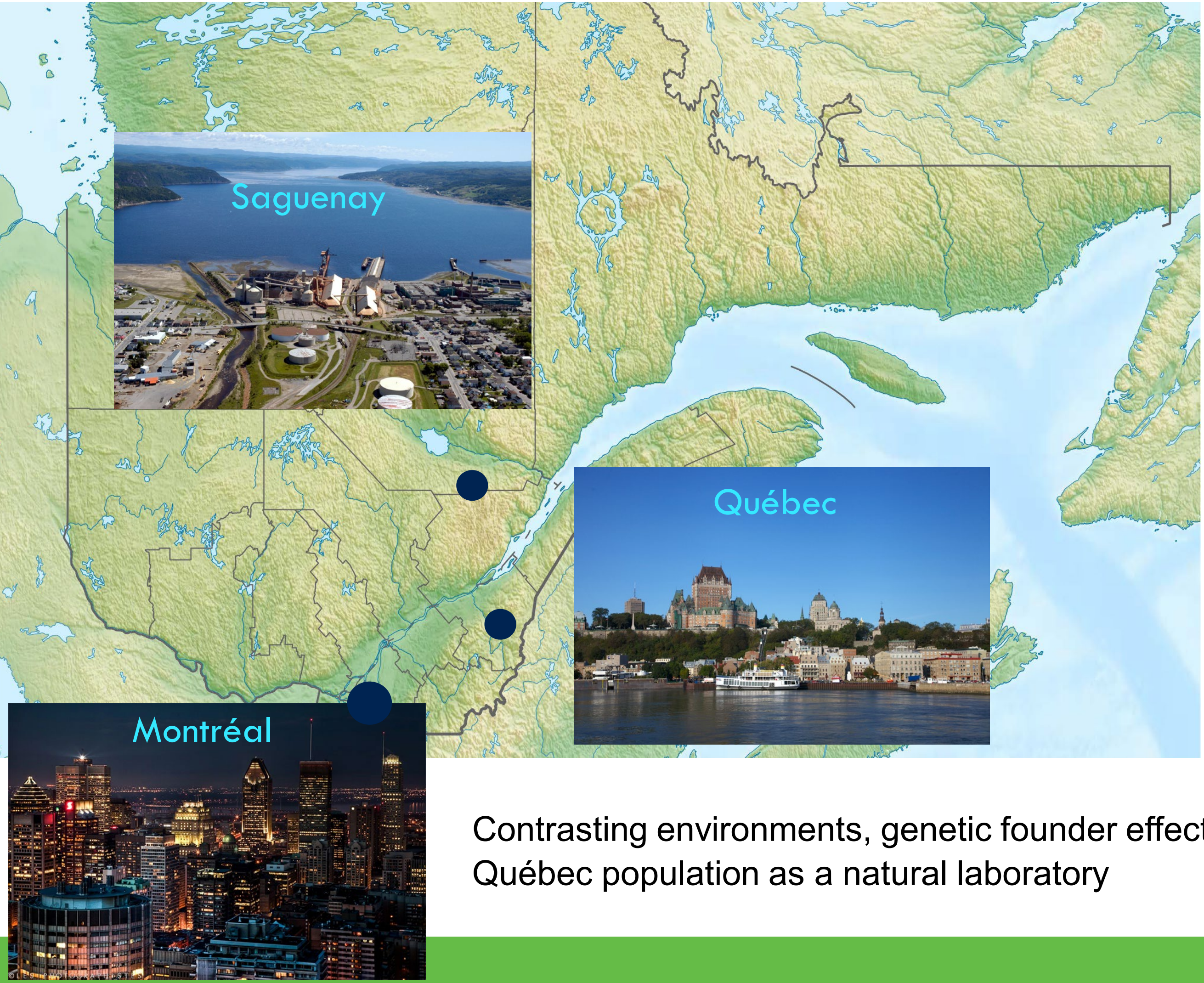
NK cell

T cell

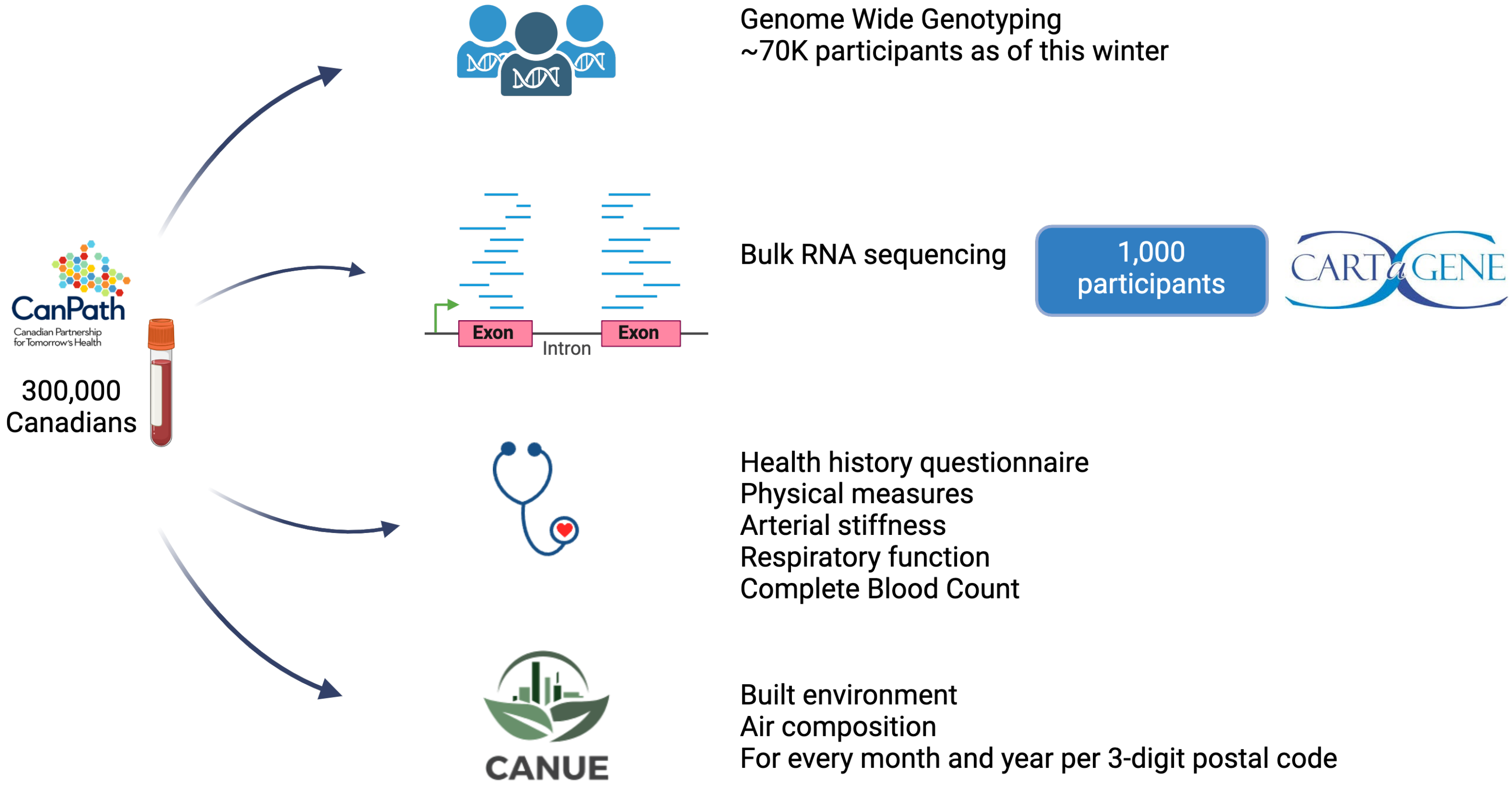
B cell



Blood and circulating immune cells are messengers going around your body



Contrasting environments, genetic founder effects, and disease prevalence
Québec population as a natural laboratory



300,000
Canadians

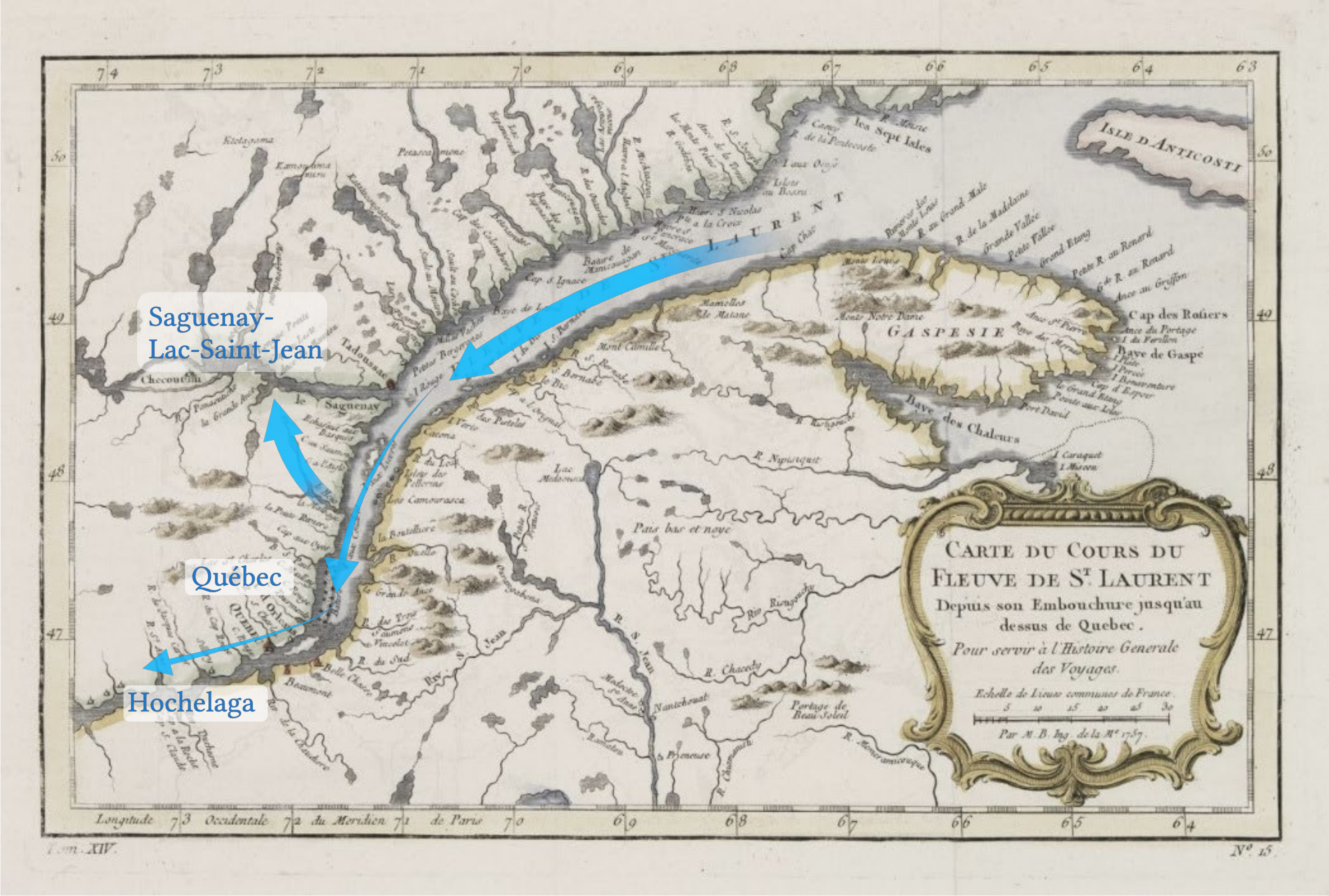
Genome Wide Genotyping
~70K participants as of this winter

Bulk RNA sequencing

1,000 participants

Health history questionnaire
Physical measures
Arterial stiffness
Respiratory function
Complete Blood Count

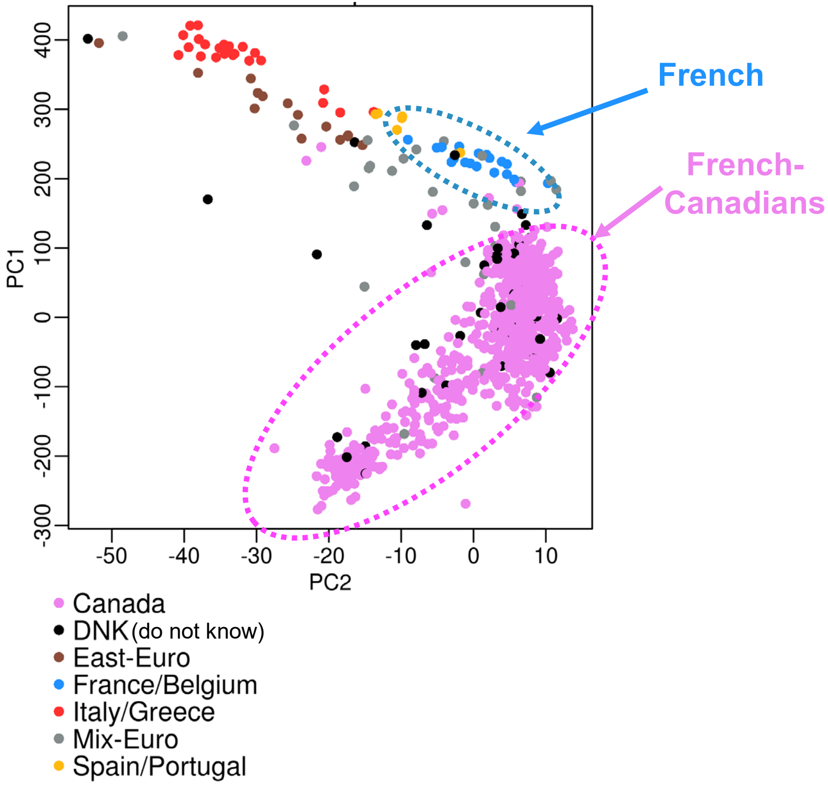
Built environment
Air composition
For every month and year per 3-digit postal code


Saguenay-
Lac-Saint-Jean
Québec
Hochelaga
Carte du Cours du
Fleuve de St. Laurent
Depuis son Embouchure jusqu'au
dessus de Quebec.
Pour servir à l'Histoire Generale
des Voyages.
Echelle de Lieues communes de France.
5 10 15 20 25 30
Par M.B. Ing. de la Mᵉ 1757.
Isle d'Anticosti
les Sept Isles
Gaspesie
Baye des Chaleurs
Cap des Roziers
Baye de Gaspé
le Saguenay
Tadoussac
Checoutimi
Quebec
Longitude 73 Occidentale 72 du Meridien 71 de Paris 70
Tom. XIV.
Nº. 15.

# French-Canadians have diverging allele frequencies from Europeans

CanPath

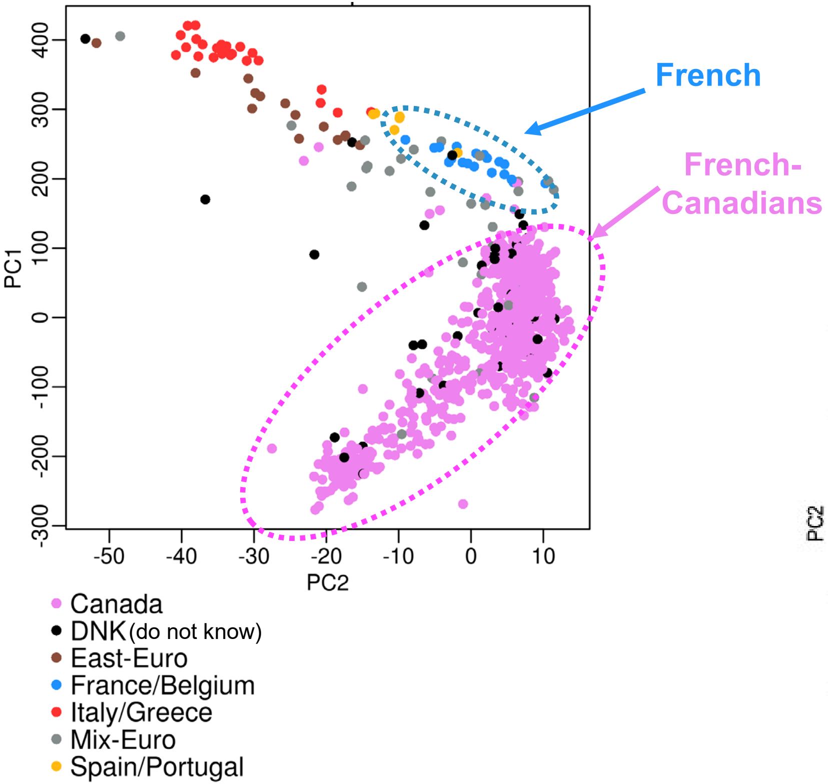
Genotyping
Illumina Omni 2.5
~2.5M SNPs



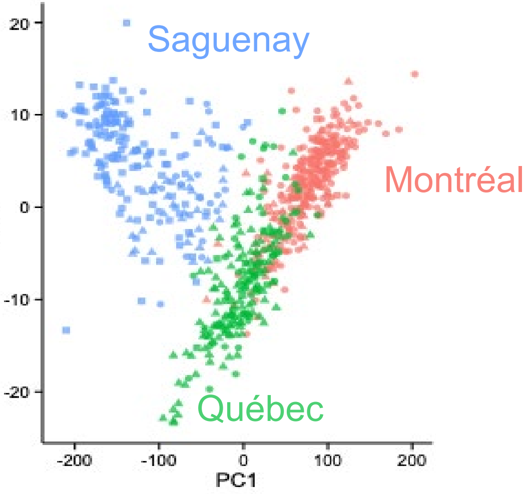
# French-Canadians have diverging allele frequencies from Europeans

CanPath

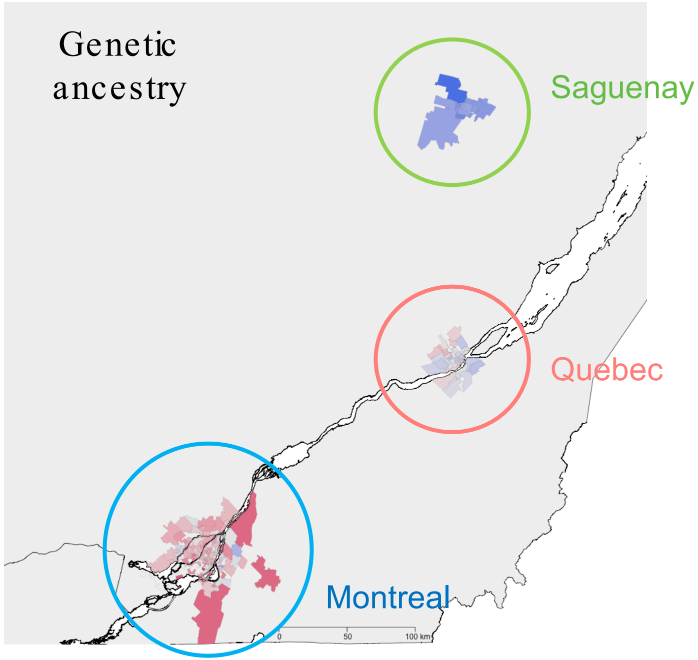
Genotyping
Illumina Omni 2.5
~2.5M SNPs



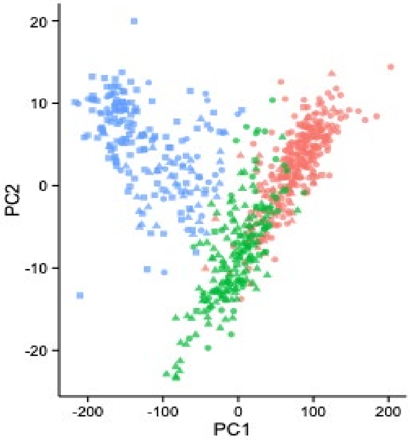
PCA with regional ancestry



# Regional ancestries within French-Canadians



PCA with regional ancestry



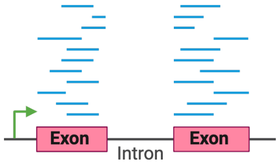
Previous and subsequent studies have shown that French Canadian genes mirror Quebec's geography and demographic history of sequential bottlenecks.

*Anderson-Trocmé (2023)*

*Gauvin et al. (2014)*

*Roy-Gagnon et al. (2011)*

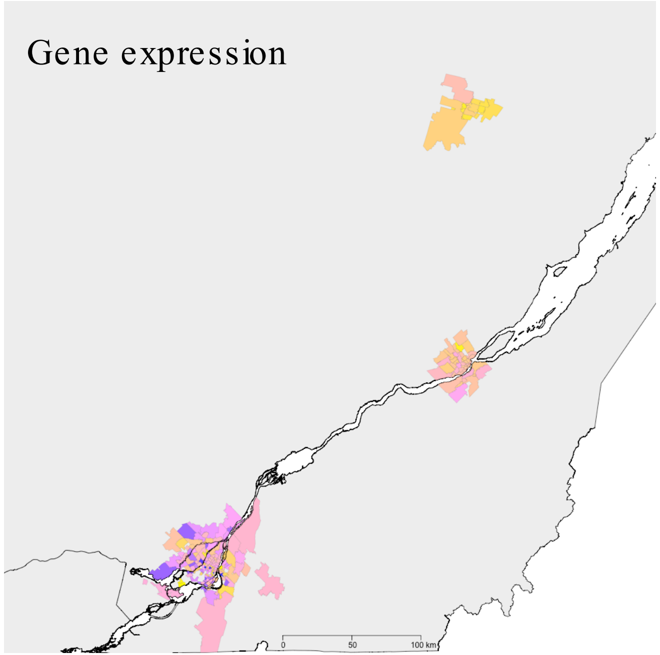
# Gene expression varies across regions



Bulk RNA sequencing

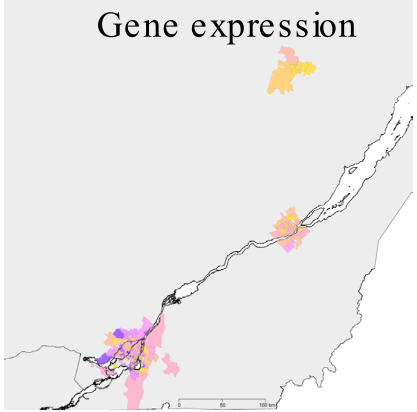
1,000 participants

~17,000 expressed genes

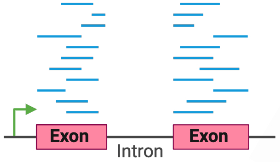


Mean PC1 value from gene expression

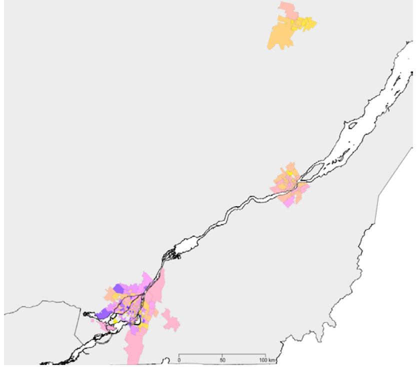
# What drive this gene expression cline?


Gene expression

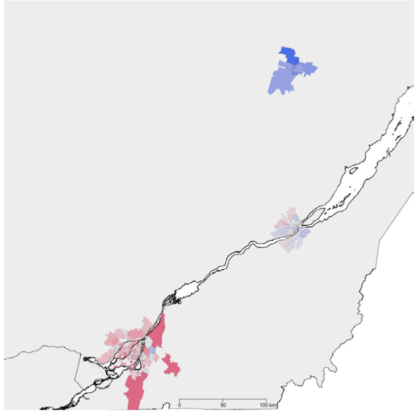
# What drive this gene expression cline?



Gene expression



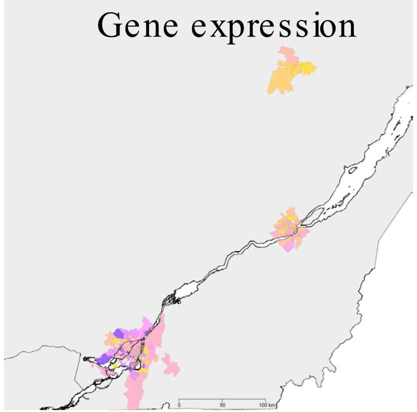
Genetic ancestry

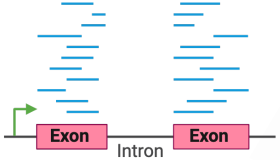


Color = PC1 value for the data type

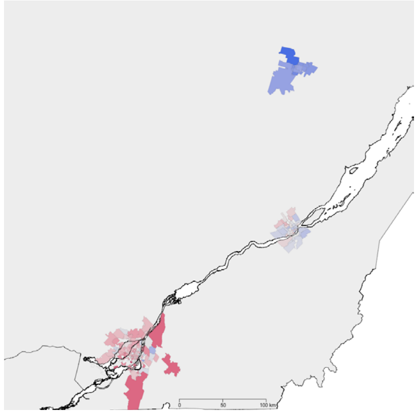
# What drive this gene expression cline?

Gene expression



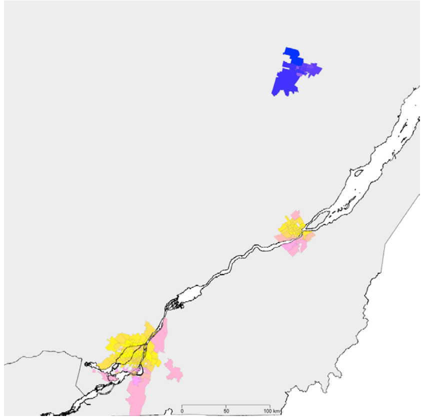


Genetic ancestry



CANUE

Environment



Is gene expression variation modulated by genetic ancestry

or

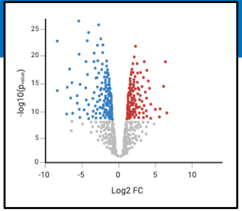
shared environment (city)?

Color = PC1 value for the data type

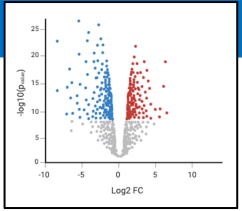
# The **environment** has an overwhelming impact on gene expression

15 Pairwise comparisons of genetic ancestry versus current living city

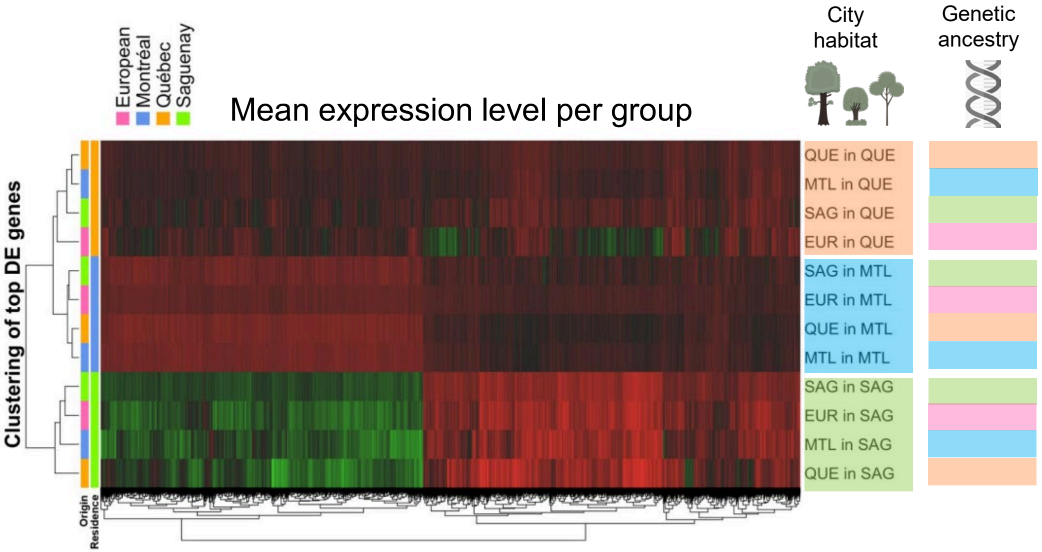
Differential gene expression analyses (negative binomial regression models)



# The **environment** has an overwhelming impact on gene expression



15 Pairwise comparisons of genetic ancestry versus current living city
Differential gene expression analyses (negative binomial regression models)



Mean expression level per group

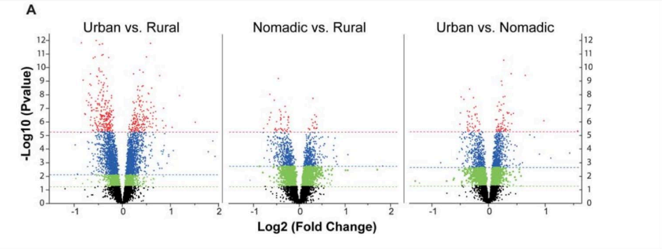
Top differentially expressed genes (n=170)

**Gene expression cluster by environment and not ancestries**

Enrichment of genes involved in gas transport and exchanges, GPCR pathways, and inflammatory pathways

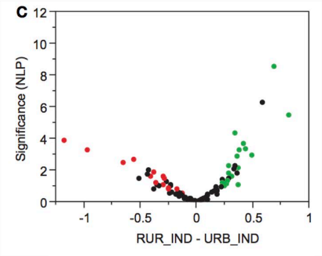
Favé et al. (2018)

## Morocco



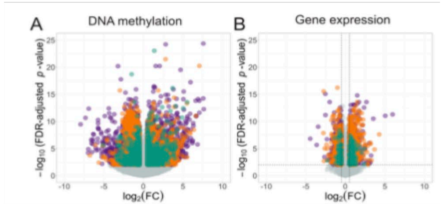
Idaghdour et al. 2008,2009

## Fiji



Nath et al. 2012

## Indonesia

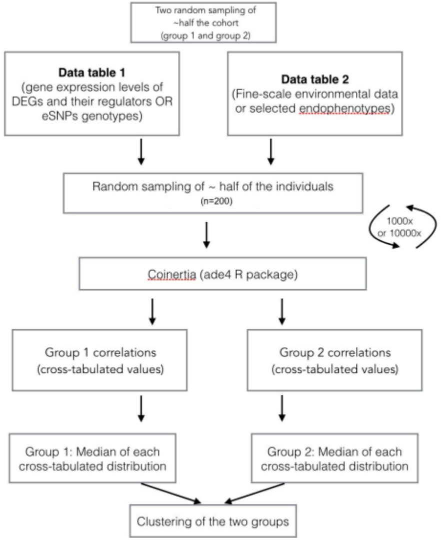


Natri et al. 2020

# Covariation of expression, environmental variables and clinical measures

Co-inertia analysis is a multivariate method for coupling two tables
(with large number of features)

Dr. Fabien Lamaze



Two random sampling of
~half the cohort
(group 1 and group 2)

**Data table 1**
(gene expression levels of DEGs and their regulators OR eSNPs genotypes)

**Data table 2**
(Fine-scale environmental data or selected endophenotypes)

Random sampling of ~ half of the individuals
(n=200)

1000x
or 10000x

Coinertia (ade4 R package)

Group 1 correlations
(cross-tabulated values)

Group 2 correlations
(cross-tabulated values)

Group 1: Median of each cross-tabulated distribution

Group 2: Median of each cross-tabulated distribution

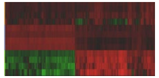
Clustering of the two groups

# Covariation of expression, environmental variables and clinical measures

Dr. Fabien Lamaze

Co-inertia analysis is a multivariate method for coupling two tables (with large number of features)

CoIA



x CANUE

x CANUE

**DGE and environment**

Differentially expressed genes across regions are associated with air composition measures : PM2.5, $NO_2$, $SO_2$ (mean conc. 2 weeks prior to blood draw)

Enriched in coagulation pathways

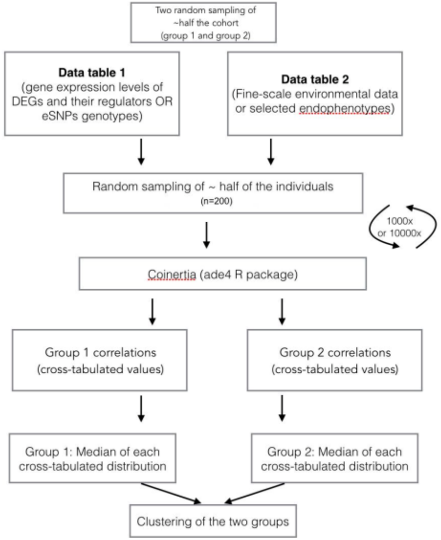
**Clinical features and environment**

Arterial stiffness higher with high $SO_2$ exposure

Risk factor for cardiovascular diseases

Also associated with Asthma and lung diseases

Sensitivity analyses within-region along air quality gradient replicate DEG across regions



Two random sampling of ~half the cohort (group 1 and group 2)

**Data table 1** (gene expression levels of DEGs and their regulators OR eSNPs genotypes)

**Data table 2** (Fine-scale environmental data or selected endophenotypes)

Random sampling of ~ half of the individuals (n=200)

1000x or 10000x

Coinertia (ade4 R package)

Group 1 correlations (cross-tabulated values)

Group 2 correlations (cross-tabulated values)

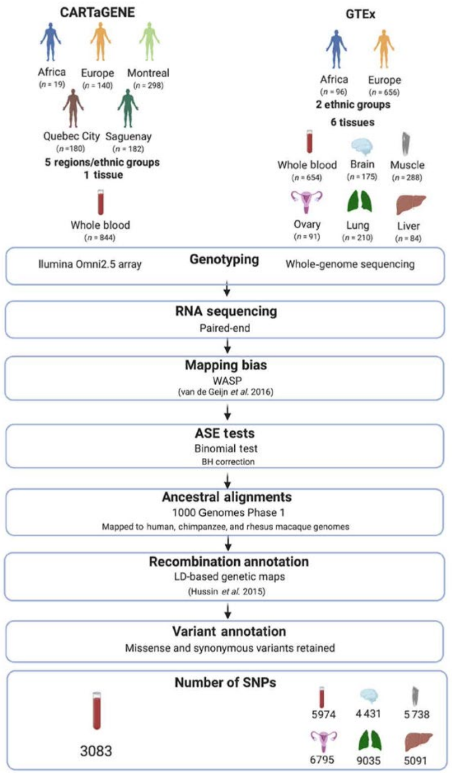
Group 1: Median of each cross-tabulated distribution

Group 2: Median of each cross-tabulated distribution

Clustering of the two groups

Favé et al. (2018)

# Allele-specific expression across populations and tissues



**CARTaGENE**

Africa (n = 19) Europe (n = 140) Montreal (n = 298)
Quebec City (n = 180) Saguenay (n = 182)
**5 regions/ethnic groups**
**1 tissue**
Whole blood (n = 844)

**GTEx**

Africa (n = 96) Europe (n = 656)
**2 ethnic groups**
**6 tissues**
Whole blood (n = 654) Brain (n = 175) Muscle (n = 288)
Ovary (n = 91) Lung (n = 210) Liver (n = 84)

Ilumina Omni2.5 array **Genotyping** Whole-genome sequencing

**RNA sequencing**
Paired-end

**Mapping bias**
WASP
(van de Geijn *et al.* 2016)

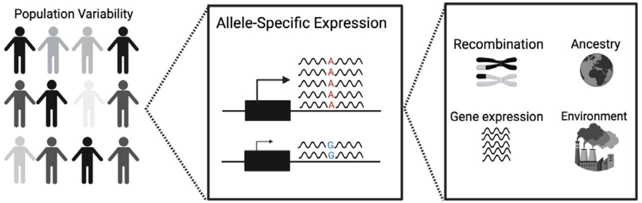
**ASE tests**
Binomial test
BH correction

**Ancestral alignments**
1000 Genomes Phase 1
Mapped to human, chimpanzee, and rhesus macaque genomes

**Recombination annotation**
LD-based genetic maps
(Hussin *et al.* 2015)

**Variant annotation**
Missense and synonymous variants retained

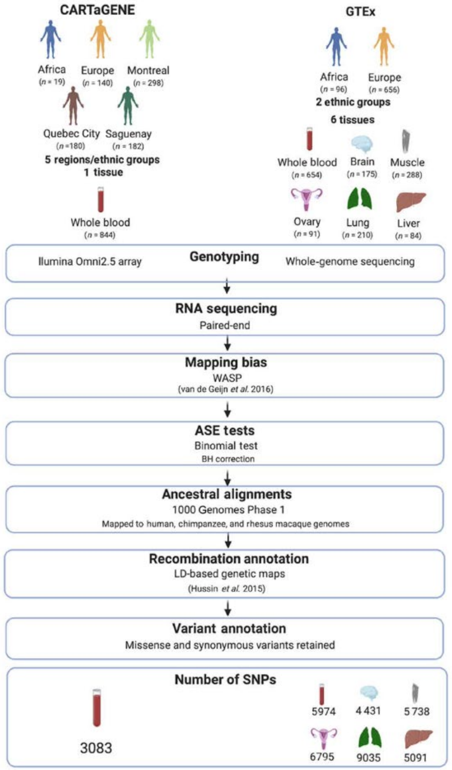
**Number of SNPs**
3083
5974 4 431 5 738
6795 9035 5091



Dr. Michelle Harwood

Harwood et al. (2022)
Harwood et al. (under review)

# Allele-specific expression across populations and tissues



**CARTaGENE**

Africa (n = 19)
Europe (n = 140)
Montreal (n = 298)
Quebec City (n = 180)
Saguenay (n = 182)

5 regions/ethnic groups
1 tissue

Whole blood (n = 844)

**GTEx**

Africa (n = 96)
Europe (n = 656)

2 ethnic groups
6 tissues

Whole blood (n = 654)
Brain (n = 175)
Muscle (n = 288)
Ovary (n = 91)
Lung (n = 210)
Liver (n = 84)

**Genotyping**
Illumina Omni2.5 array
Whole-genome sequencing

**RNA sequencing**
Paired-end

**Mapping bias**
WASP
(van de Geijn *et al.* 2016)

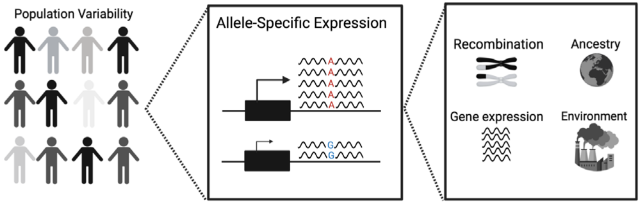
**ASE tests**
Binomial test
BH correction

**Ancestral alignments**
1000 Genomes Phase 1
Mapped to human, chimpanzee, and rhesus macaque genomes

**Recombination annotation**
LD-based genetic maps
(Hussin *et al.* 2015)

**Variant annotation**
Missense and synonymous variants retained

**Number of SNPs**
3083
5974
4 431
5 738
6795
9035
5091



Dr. Michelle Harwood

Recombination affects allele-specific expression (ASE) of deleterious variants in human populations

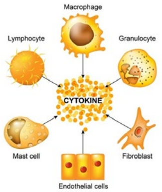
Founder populations show **less efficient selection** on ASE to **buffer** deleterious polymorphisms

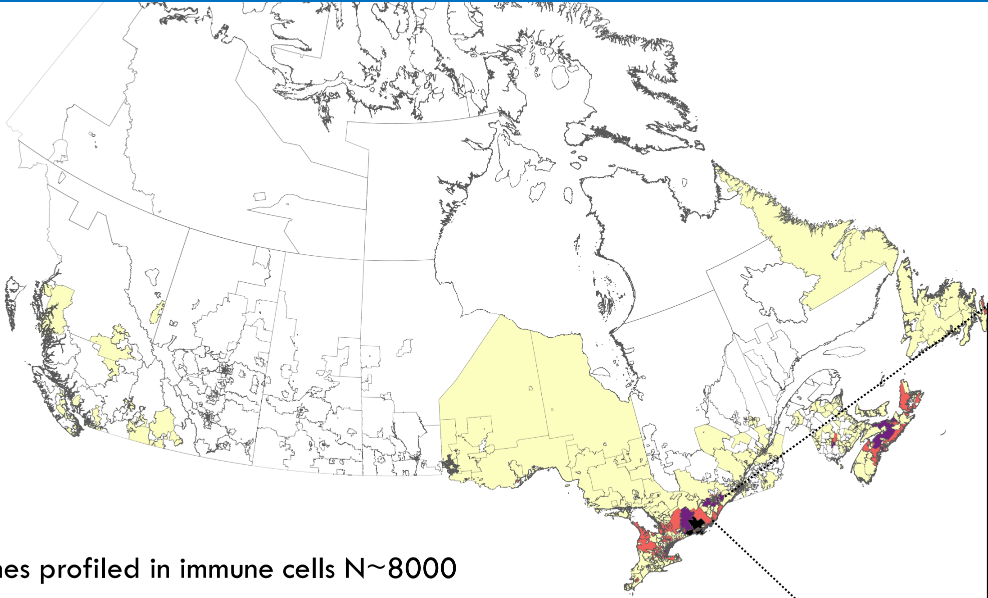
Medication response is associated with ASE

Harwood et al. (2022)
Harwood et al. (under review)

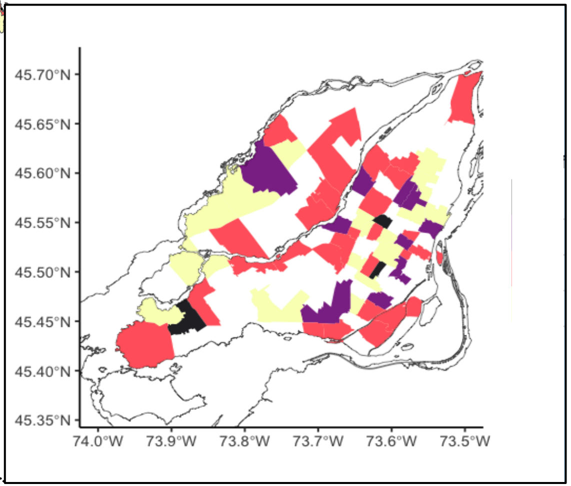
# Plasma cytokine profiling across CanPath neighborhoods

CanPath
Canadian Partnership
for Tomorrow's Health





Neighborhoods (FSA) in Montréal with n>50 participants with cytokine profiling

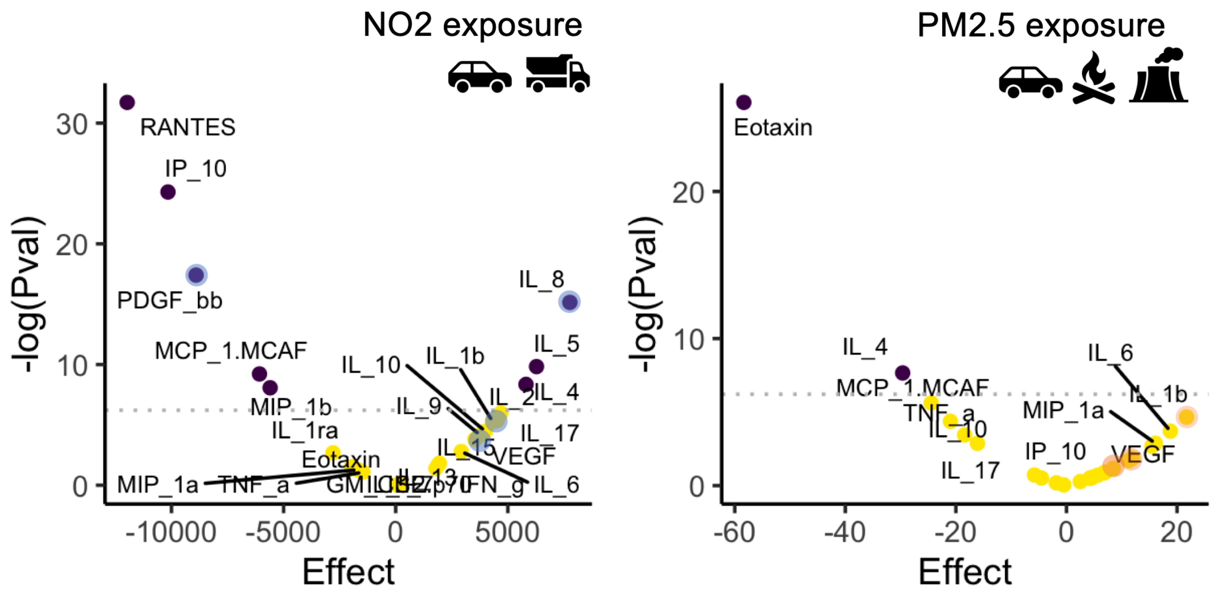


Participants with cytokines profiled in immune cells N~8000

Small scale geographic information (3 digits postal code)

Cytokines have diverse functions in the immune system and beyond. They can promote inflammation to help fight infection or injury, stimulate the production and activation of immune cells

# Inflammatory cytokines replicate some associations with air quality



Underexpressed in high NO2

Overexpressed in high NO2

Underexpressed in high PM2.5

Overexpressed in high PM2.5

Gene encoding the cytokine also found associated with
● NO2 levels or ● PM2.5 levels in Favé et al. 2018

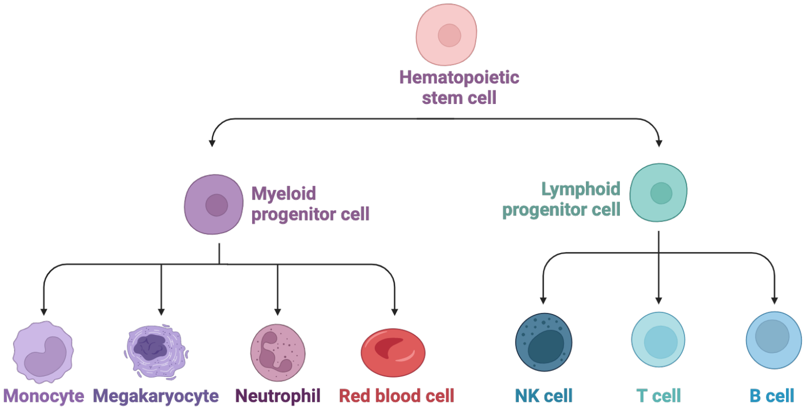
Pilot project, unpublished

# Immune Risk Score is associated with mortality and disease

**Immune Risk Score**

- Hematocrit
- White blood cell concentration
- Platelet concentration
- Mean corpuscular volume
- Mean corpuscular hemoglobin concentration
- Red blood cell distribution width coefficient of variation

*Horne et al. (2009) Am J Med*



0 → 10+

LOW RISK HIGH RISK

Young Low Risk

Aged Low Risk

Young High Risk

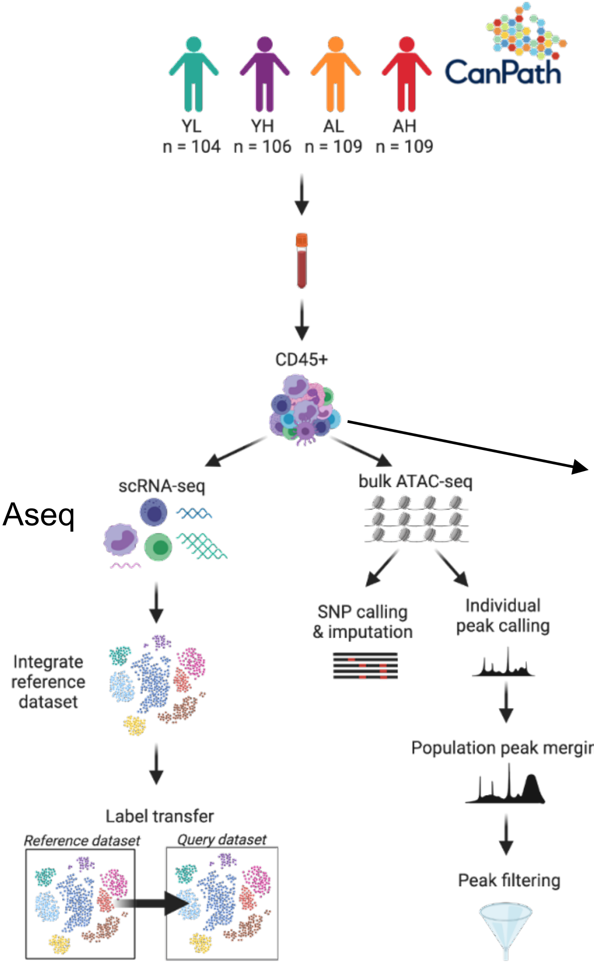
Aged High Risk


Dr. Elyssa Bader


Elias Gbeha


Mawussé Agbessi


Vanessa Bruat



Single-cell RNAseq

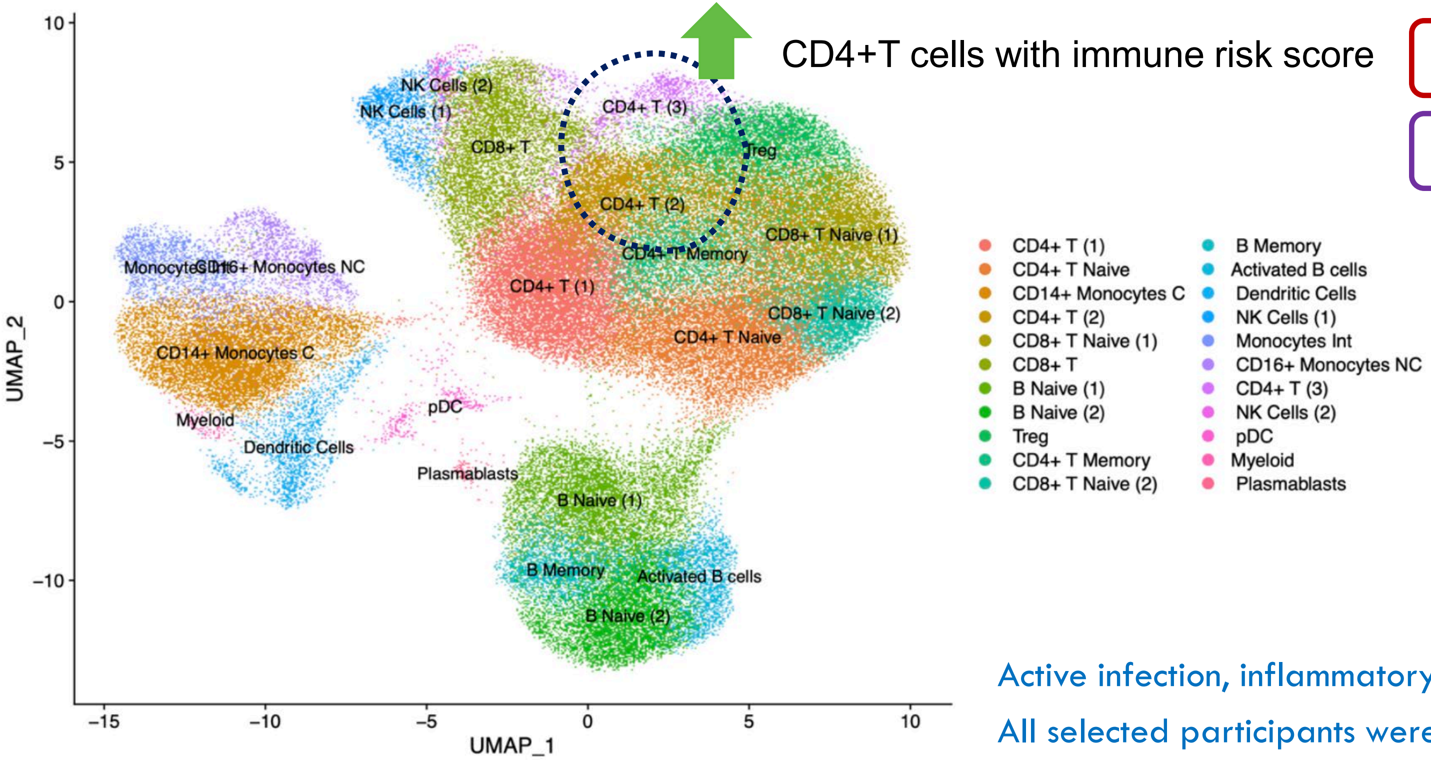
WGS low coverage

imputation

genotypes

Bader, Favé et al. (under review)

# Increase of CD4+Tcells (adaptive immune cells) with immune risk score



CD4+T cells with immune risk score

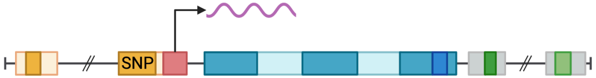
Aged High Risk

Young High Risk

Active infection, inflammatory state, cancer

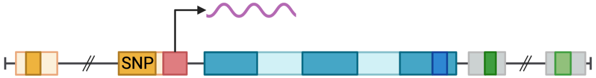
All selected participants were cancer free at baseline and after recent recontact (no incident cases)

# Discovering cell type specific eQTLs across the immune system and aging

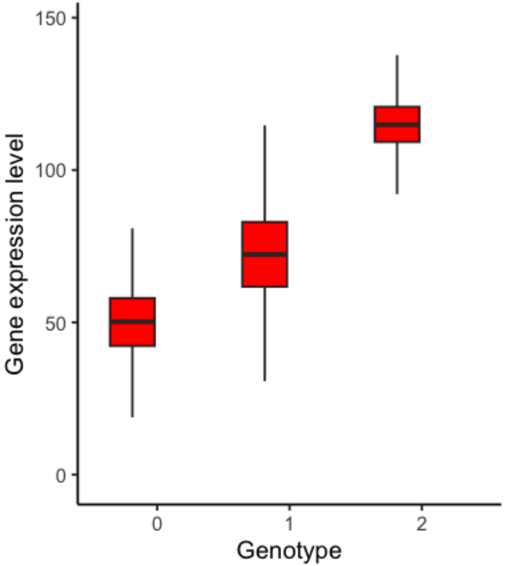


Expression quantitative trait loci

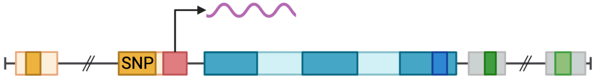
# Discovering cell type specific eQTLs across the immune system and aging



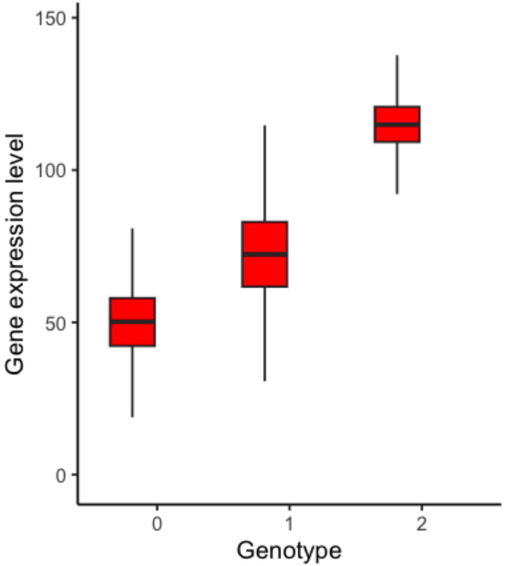
Expression quantitative trait loci



# Discovering cell type specific eQTLs across the immune system and aging

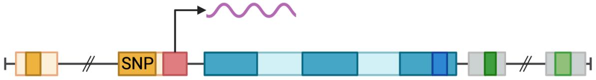


Expression quantitative trait loci

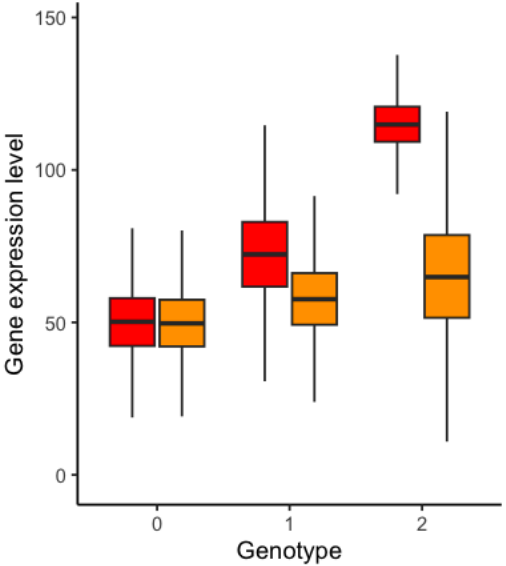


Aged High Risk

# Discovering cell type specific eQTLs across the immune system and aging



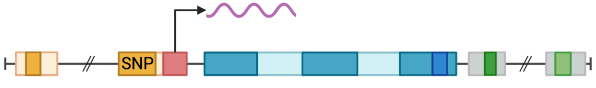
Expression quantitative trait loci



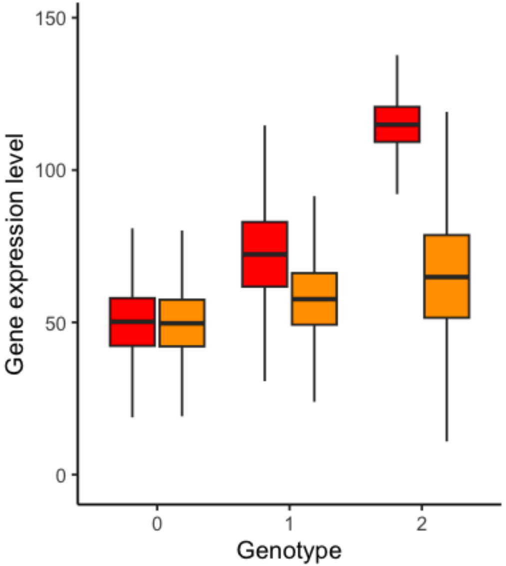
Aged High Risk

Aged Low Risk

# Discovering cell type specific eQTLs across the immune system and aging



Expression quantitative trait loci

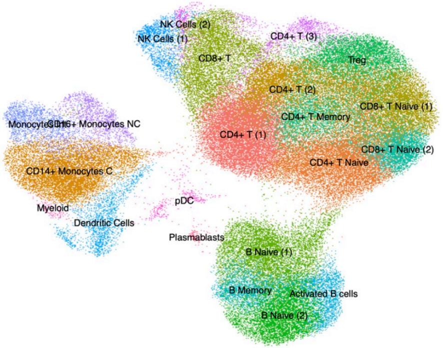


Aged High Risk

Aged Low Risk

Expression ~ SNP + RiskScore + sex + 10 gPCs + 4 SVs + **SNP:RiskScore** + e

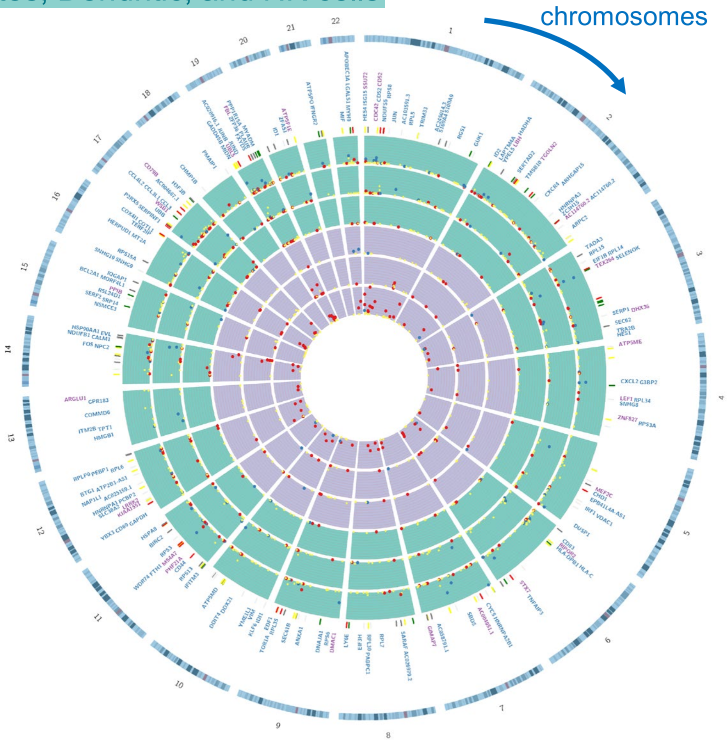
21 cell types, across all genes

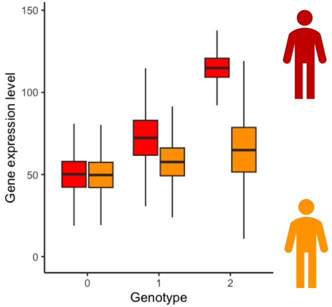


# Regulatory SNPs that are associated with healthy aging are cell type specific

CD4+, CD8+ Tcells and B cells

Monocytes, Dendritic, and NK cells





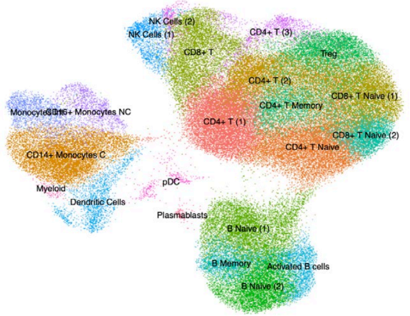
**1105 cis-eQTL** dependant on immune risk

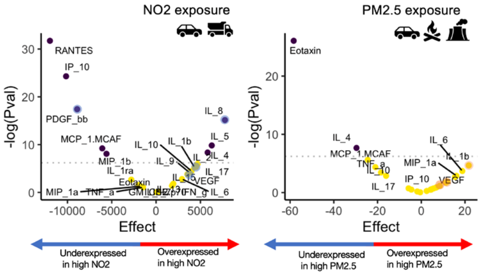
Most are **cell type specific**

SNPs with low MAF have higher effect sizes

Located in **quiescent** chromatin regions (ATACseq data)

Replicated in Cartagene bulk RNAseq

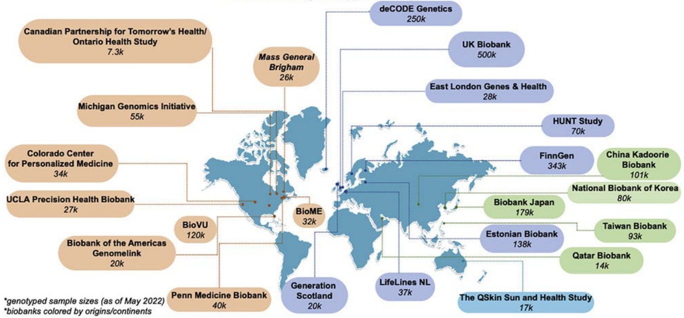




Gene encoding the cytokine also found associated with
● NO2 levels or ● PM2.5 levels in Favé et al. 2018

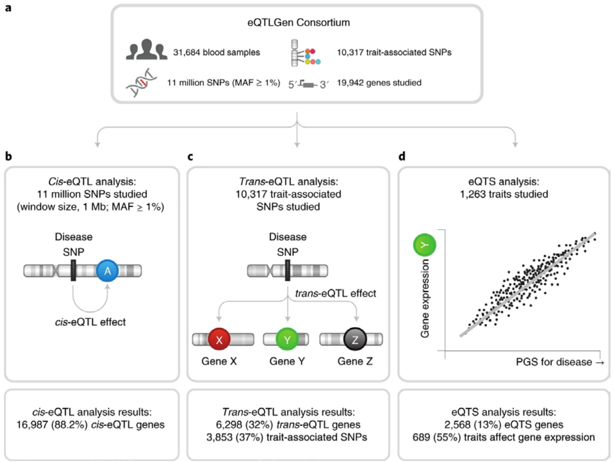
## 23 biobanks with different origins and ancestries have joined GBMI

**> 2.2 million genotyped samples**



**deCODE Genetics** 250k
**Canadian Partnership for Tomorrow's Health/ Ontario Health Study** 7.3k
**UK Biobank** 500k
***Mass General Brigham*** 26k
**East London Genes & Health** 28k
**Michigan Genomics Initiative** 55k
**HUNT Study** 70k
**Colorado Center for Personalized Medicine** 34k
**FinnGen** 343k
**China Kadoorie Biobank** 101k
**National Biobank of Korea** 80k
**UCLA Precision Health Biobank** 27k
**Biobank Japan** 179k
**BioME** 32k
**BioVU** 120k
**Estonian Biobank** 138k
**Taiwan Biobank** 93k
**Biobank of the Americas Genomelink** 20k
**Qatar Biobank** 14k
**LifeLines NL** 37k
**Generation Scotland** 20k
**Penn Medicine Biobank** 40k
**The QSkin Sun and Health Study** 17k

**genotyped sample sizes (as of May 2022)*
**biobanks colored by origins/continents*



a

eQTLGen Consortium

31,684 blood samples
10,317 trait-associated SNPs
11 million SNPs (MAF ≥ 1%)
19,942 genes studied

b

*Cis*-eQTL analysis:
11 million SNPs studied
(window size, 1 Mb; MAF ≥ 1%)

Disease SNP
*cis*-eQTL effect

*cis*-eQTL analysis results:
16,987 (88.2%) *cis*-eQTL genes

c

*Trans*-eQTL analysis:
10,317 trait-associated SNPs studied

Disease SNP
*trans*-eQTL effect
Gene X
Gene Y
Gene Z

*Trans*-eQTL analysis results:
6,298 (32%) *trans*-eQTL genes
3,853 (37%) trait-associated SNPs

d

eQTS analysis:
1,263 traits studied

eQTS analysis results:
2,568 (13%) eQTS genes
689 (55%) traits affect gene expression

**Awadalla Lab at OICR (& formerly UdeM)**

Dr. Philip Awadalla
Mawussé Agbessi
Dr. Armande Ang Houle (Genentech)
Dr. Elyssa Bader (Genentech)
Vanessa Bruat
Nicholas Cheng
Elias Gbeha
Dr. Heather Gibling
Elizabeth Hall (UoT Medicine)
Dr. Michelle Harwood (Genentech)
Dr. Alan Hodgkinson (King's College)
Dr. Youssef Idaghdour (NYU Abu Dhabi)
Dr. Fabien Lamaze (IUCPQ, U. Laval)
Ido Nofech-Mozes
Tom Ouellette
Dr. Kimberly Skead (Regeneron)
Dr. David Soave (Wilfrid Laurier)
Jasmina Uzunović (Genentech)
June Kim
Jasmine Kang
Yiran Shao
Zixuan Lan

OICR Ontario Institute for Cancer Research Collaborate. Translate. Change lives.

MONTREAL HEART INSTITUTE

VECTOR INSTITUTE

CDIC Canadian Data Integration Centre

OICR
Ontario Institute
for Cancer Research

Funding for the Ontario Institute for Cancer Research
is provided by the Government of Ontario

Ontario