Using fine-grained population based grid data to assess genetic self-selection into residential neighborhoods across the lifespan

Rafael Ahlskog, Sven Oskarsson
Uppsala university

Background

- Gene-environment correlations, or rGE, can cause bias in conventional genetic discovery studies. rGE can come in many different varieties, and not all are problematic.
- Most notably, *self-selection* behavior driven by genetic differences most likely form an integral part of the causal signal of genes on many social science outcomes.
- This is often called *active rGE*, and controlling for this component of rGE is therefore, in most cases, not a good idea.

Background

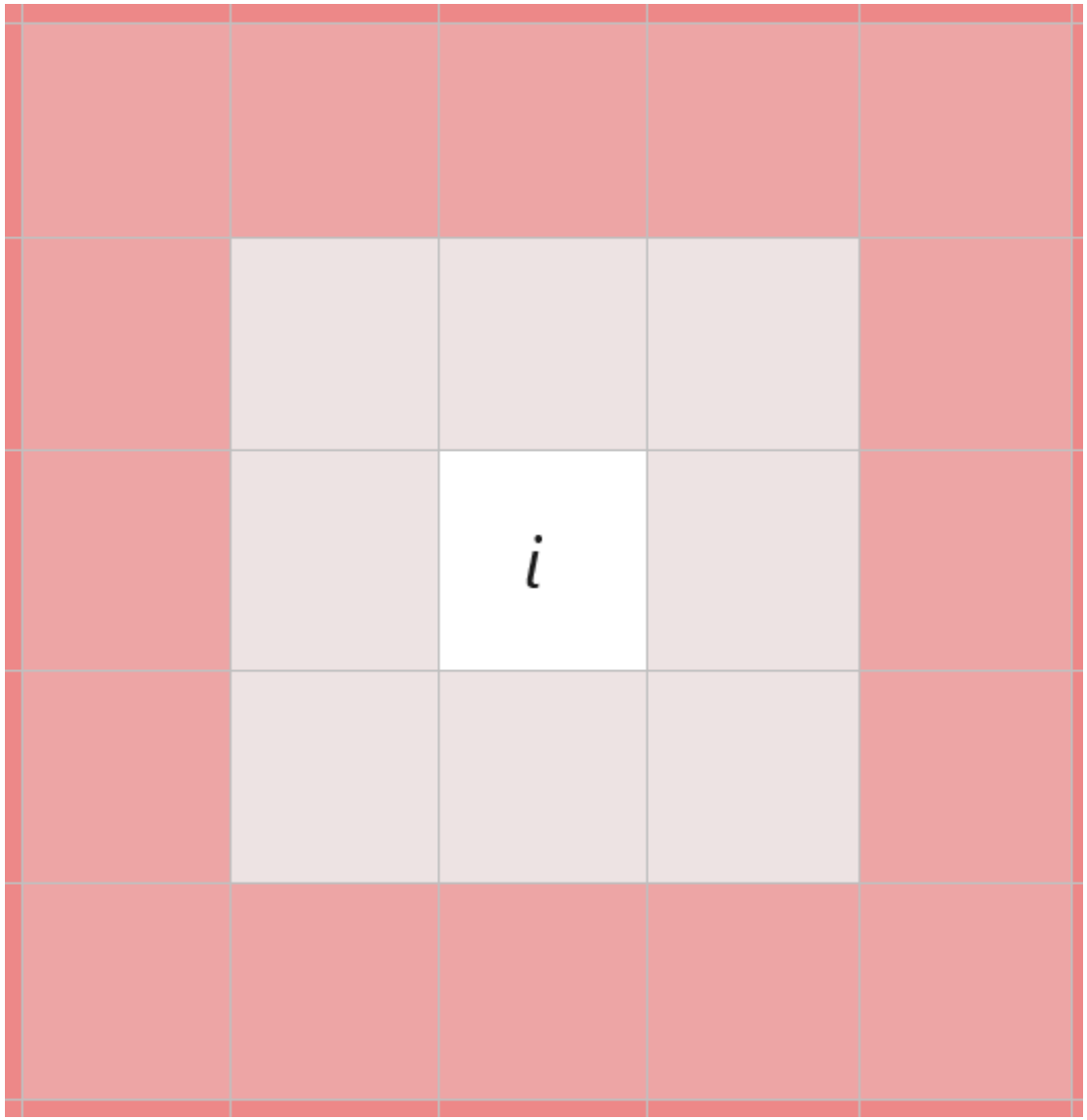
- A salient example of active rGE is selection into residential neighborhood environments, which can be important influences on many life outcomes. ("direct" genetic effects are not at all direct)
- Structure of genetic self-selection into residential environments largely unknown:
  - What neighborhood characteristics do people sort on?
  - What genetic factors are driving selection behavior?
  - When during the life cycle does the sorting happen?
  - To what extent is this active rGE passed on to the next generation in the form of *passive* rGE?
- This is an early attempt to illuminate some of these dynamics, using a unique combination of full-population register data and a large sample of genotyped twins, both of which have fine-grained geographical identifiers over time.

Data – education LCs

- Sample of the complete Swedish population with annual register data on a wide range of indicators, as well as...
- ...geographical data, showing the location of an individual's place of residence, annually, at 250x250 meter grid granularity (or 1000x1000 in very sparsely populated areas).
- For most purposes, this is fine-grained enough to locate distinct neighborhoods in e.g. the same area of the same city.

Data – education LCs

- For each grid in the population data, we aggregate years of education (residualized on sex and birth cohort), for the area of grids surrounding each index grid that contains at least 30 people (or at most five layers of surrounding grids).
- We do this for each year between 1990-2018 (which is when we have good quality annual data on education level).


i

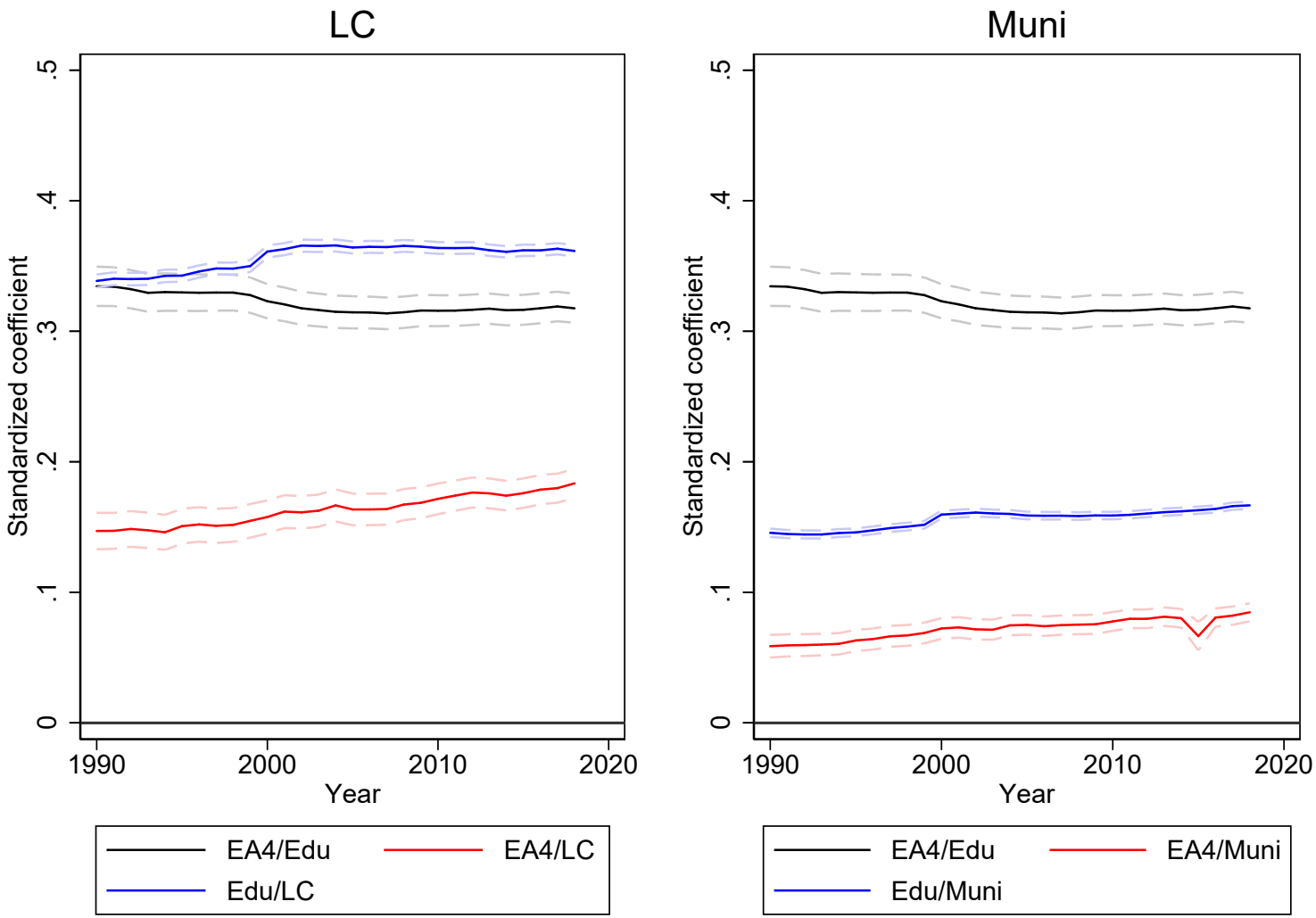
Data – education LCs

- These population-pased grid characteristics are merged to a sample of genotyped DZ twins from the Swedish Twin Register. Their own residualized education is jacknifed out.
- Gives an individualized, fine-grained measure of the average level of education for everyone *else* in the immediate surrounding area.
- Procedure similar to e.g. KNN or the EquiPop algorithm, but complicated by the fact that these are two different samples with different anonymized pseudo-IDs.

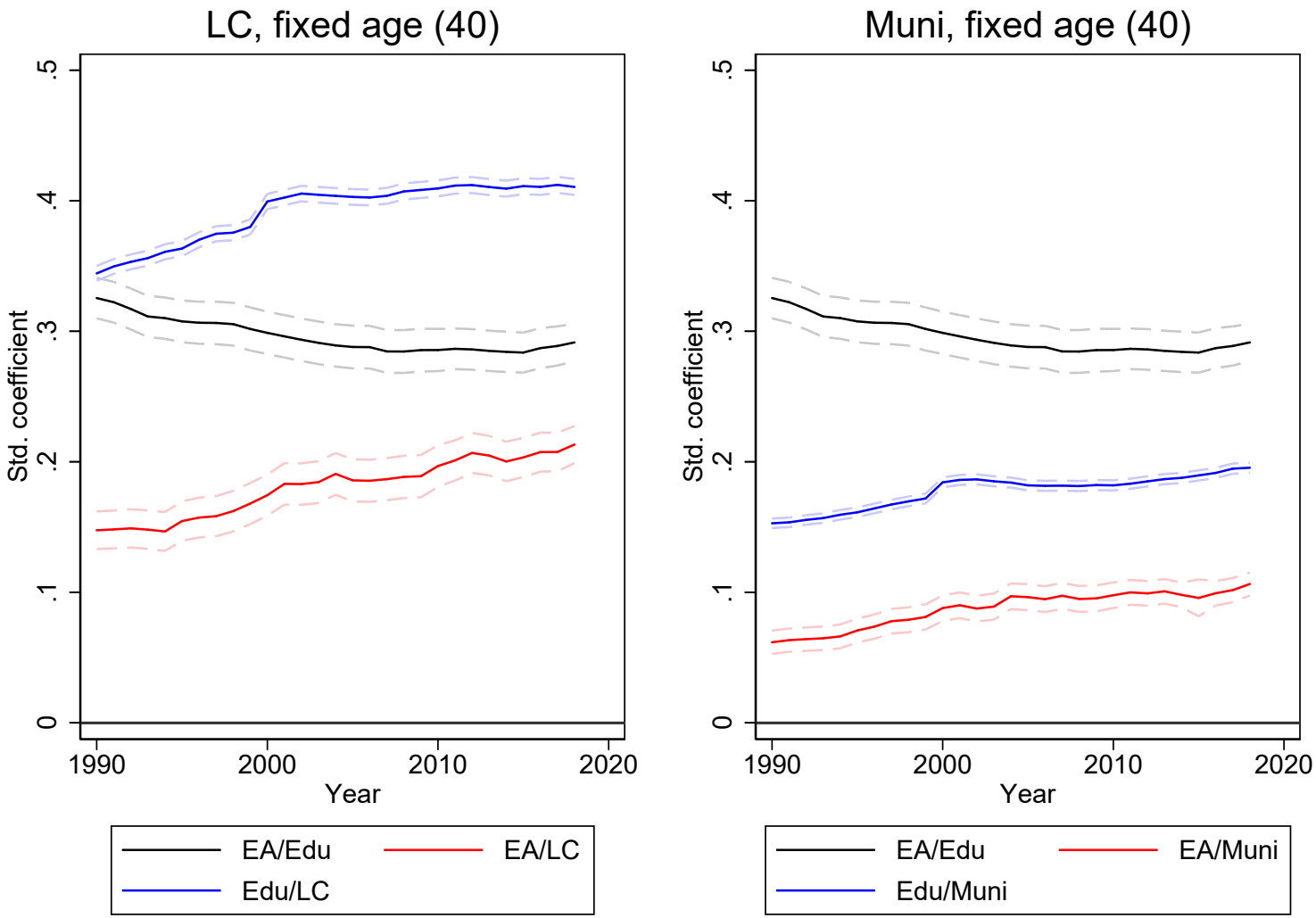
Some stats on education LCs

- Average population size in the individualized local contexts vary by year between 160-250 people. The total number of grids is >200'000.
- As a point of comparison, the average population size per municipality is between 54'000 and 111'000, across 280 municipalities.

Do LCs work?


LC
Muni
Standardized coefficient
0
.1
.2
.3
.4
.5
1990
2000
2010
2020
Year
EA4/Edu
EA4/LC
Edu/LC
EA4/Muni
Edu/Muni

Do LCs work?


LC, fixed age (40)
Std. coefficient
0
.1
.2
.3
.4
.5
1990
2000
2010
2020
Year
EA/Edu
EA/LC
Edu/LC
Muni, fixed age (40)
Std. coefficient
0
.1
.2
.3
.4
.5
1990
2000
2010
2020
Year
EA/Edu
EA/Muni
Edu/Muni

Do LCs work?

- Substantial – and increasing – level of sorting on both phenotype and genotype.
- LCs capture this much better (roughly twice as well) than e.g. municipalities.

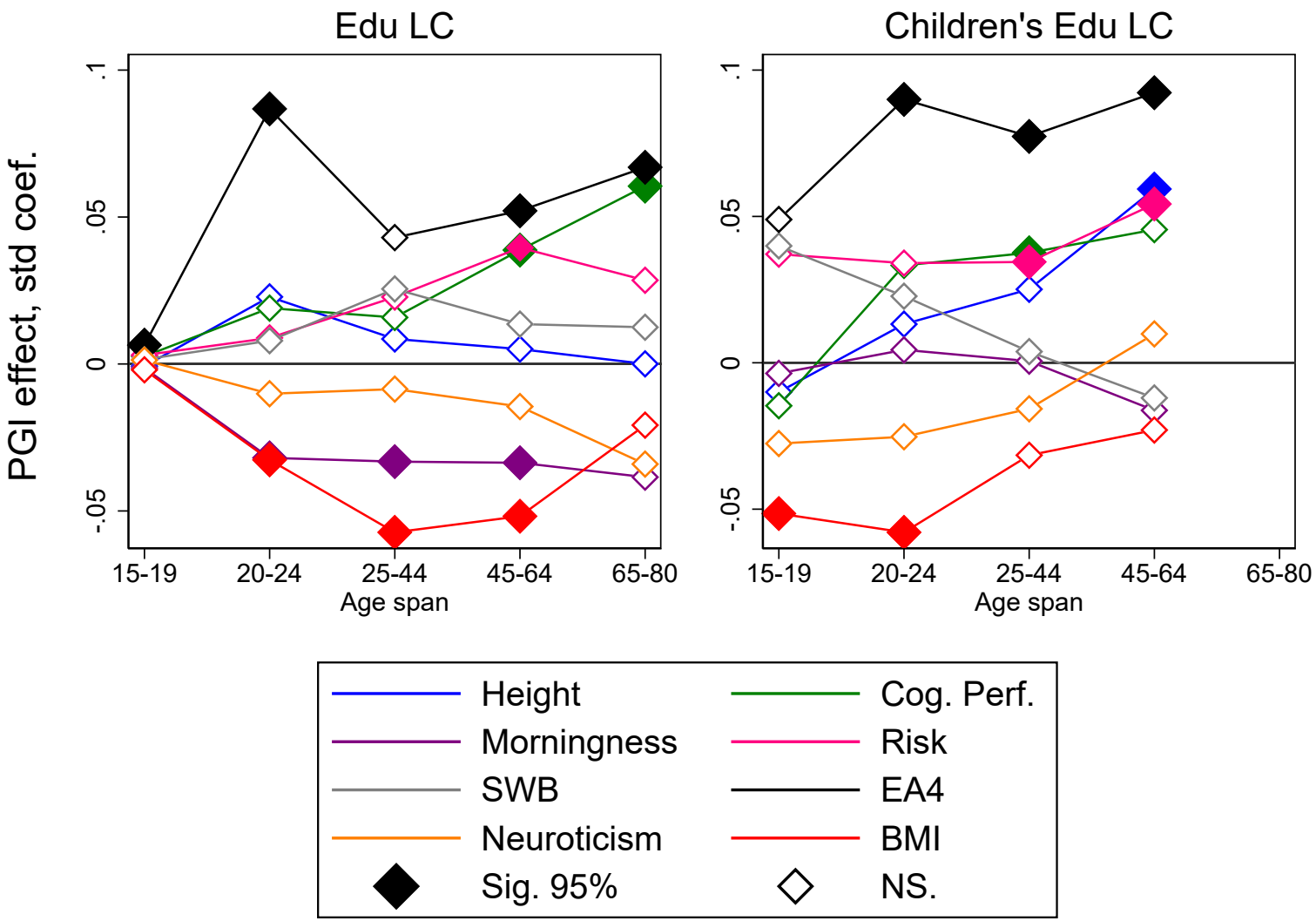
Disentangling active rGE

- We can use within-family models with the DZ twins to look directly at the active component of the correlation between individual genotype and contextual education:

$$(LC_{1j} - LC_{2j}) = \alpha + \beta(PGI_{1j} - PGI_{2j})$$

- Furthermore, we can utilize differences among the *children* of the DZ twins to look at both generational persistence, and the passive rGE component stemming from active sorting in the previous generation:

$$(LC^c_{1j} - LC^c_{2j}) = \alpha + \beta(PGI_{1j} - PGI_{2j})$$


Edu LC
Children's Edu LC
PGI effect, std coef.
.1
.05
0
-.05
15-19
20-24
25-44
45-64
65-80
Age span
Height
Morningness
SWB
Neuroticism
Sig. 95%
Cog. Perf.
Risk
EA4
BMI
NS.

Preliminary takeaways

- Gene-environment correlations between EA alleles and local contextual education partially reflect active rGE, i.e. self-selecting into neighborhoods with different socioeconomic characteristics – and does so at the point in the life cycle when we should expect it.
- Active selection effects (mechanically) translate to passive rGE in the next generation.
- Effects of parental genotype show a tendency to be stronger than effects of own genotype later in life. This could reflect e.g. assortative mating or Matthew-type GxE effects.
- Selection effects into educational LCs are most prominent for the EA4 PGI, but can also be seen for BMI, morningness, cognitive performance and risk preference.

Thank you!

This is preliminary work – any and all suggestions and ideas are highly appreciated!