Burden analysis of centenarian and control genomes

Elizabeth T. Cirulli, Mingfu Zhu, Min He, Dongliang Ge, Kevin V. Shianna, David B. Goldstein

Center for Human Genome Variation, Duke University School of Medicine, Durham, NC 27708

Introduction

•Heritability of longevity estimated at ~25%
•Exceptional longevity runs in families
•Genetic variants in model organisms shown to affect longevity
  •*e.g.*, mutations decreasing the activity of daf-2 double the lifespan of *C. elegans*
•Previous techniques to study genetics of longevity:
  •Linkage studies
    •Interesting regions identified but no causal variants
  •GWAS (genome-wide association study)
    •Only consistent signal is for APOE
      •Associated with Alzheimer's disease and longevity
•Most of the genetic factors influencing longevity remain unknown
•Here, we investigate genetics of longevity by sequencing genomes of centenarians
  •Focus on burden analysis
  •Do centenarians have fewer rare, deleterious variants than controls?

Methods

•Genomes of 178 individuals sequenced to an average of 33.4x coverage
  •used either llumina Genome Analyzer II or the Illumina HiSeq2000
  •13 centenarians
    •11 European ancestry, 2 African
  •165 controls
    •117 European ancestry, 43 African, 5 Hispanic
    •Sequenced as part of other projects
      •Healthy controls, schizophrenics, epilepsy patients, etc.
•Identify all nonsynonymous variants in canonical transcripts
    •Nonsynonymous variants are the largest class of variants that are likely to have a functional effect and be enriched for deleterious variants
    •Use MAF cutoffs to enrich further for deleterious variants
      •MAFs calculated using all 178 samples
    •Use PolyPhen2 [1] predictions to further enrich for deleterious
    •Compare to synonymous variants, which are less likely to have a functional effect and be deleterious
•Compare number of nonsynonymous variants found in centenarians and controls
  •Linear regression to identify significant differences

Results

Effect of sequencing platform

•In burden analysis, important to remove sources of artifact
•Clear differences found between samples sequenced on Genome Analyzer and HiSeq (Figure 1)
•Therefore, burden analysis will focus on samples sequenced on HiSeq2000

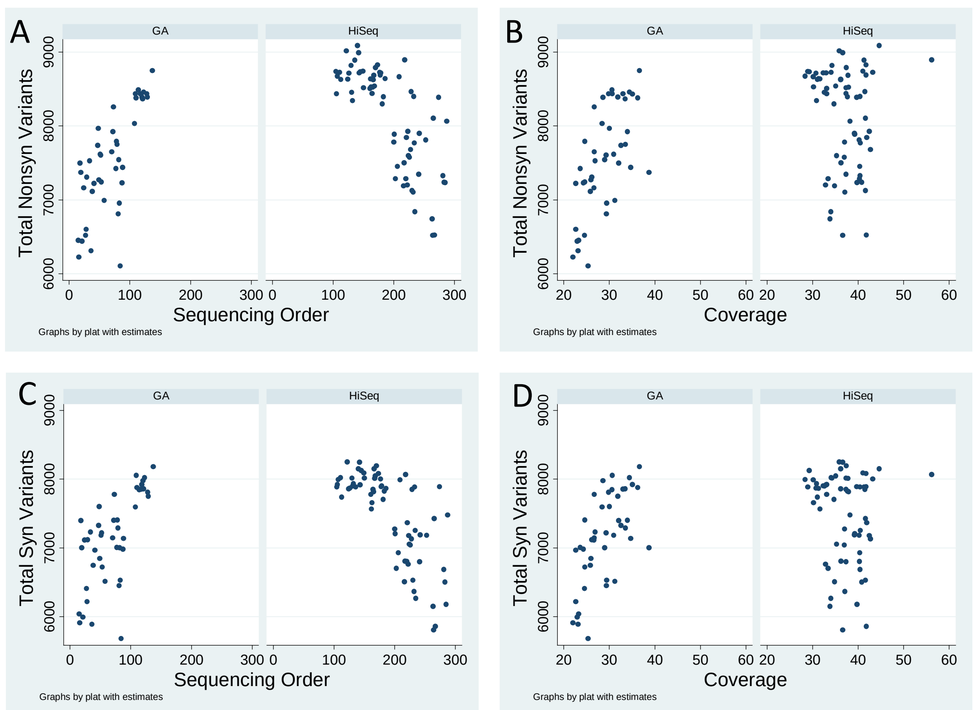

Fig. 1. The total number of nonsynonymous (A,B) and synonymous variants (C,D), are plotted against the sequencing order (A,C) and the average genomic coverage (B,D). The sequencing order is a proxy for the date on which the sample was sequenced; as technology has changed over time, it has affected the number of variants called, with the number of variants increasing with time on the Genome Analyzer (GA, left) and decreasing with time on the HiSeq2000 (HiSeq, right). The number of variants called increases with coverage up to approximately 30x coverage, at which point it levels off. Only European American controls are shown.

Burden analysis

•Compare 11 centenarians to 117 controls
  •All European ancestry, all sequenced on HiSeq2000
•Trend for centenarians to have significantly fewer rare, nonsynonymous variants (Table 1, Figure 2)
  •Association driven by variants predicted by Polyphen2 [1] to be benign or of unknown consequence (Figure 3)
  •No trend for fewer nonsynonymous variants predicted to be damaging in centenarians
•No trend for the number of synonymous variants to be different in centenarians (Figure 2)

	Nonsynonymous	Probably damaging	Possibly damaging	Benign	Unknown	Synonymous
All variants	0.723					0.811
MAF<5%	0.481	0.660	0.859	0.33	0.465	0.889
MAF<1%	0.029	0.329	0.351	0.127	0.001	0.427
Singletons	0.024	0.414	0.349	0.028	0.008	0.435

Table 1. Statistical results of comparison of European HiSeq2000 centenarians to controls. We ran a linear regression model that had the total number of variants as the outcome and the sequencing order, coverage, and centenarians status as variables; shown are the p-values for the centenarian variable.. The rows indicate all variants, those with MAF<5%, <1%, and those that were seen only once (singletons). The columns indicate all nonsynonymous variants, four classes of nonsynonymous variants as predicted by Polyphen2, and synonymous variants. Cells are highlighted if p<0.05.

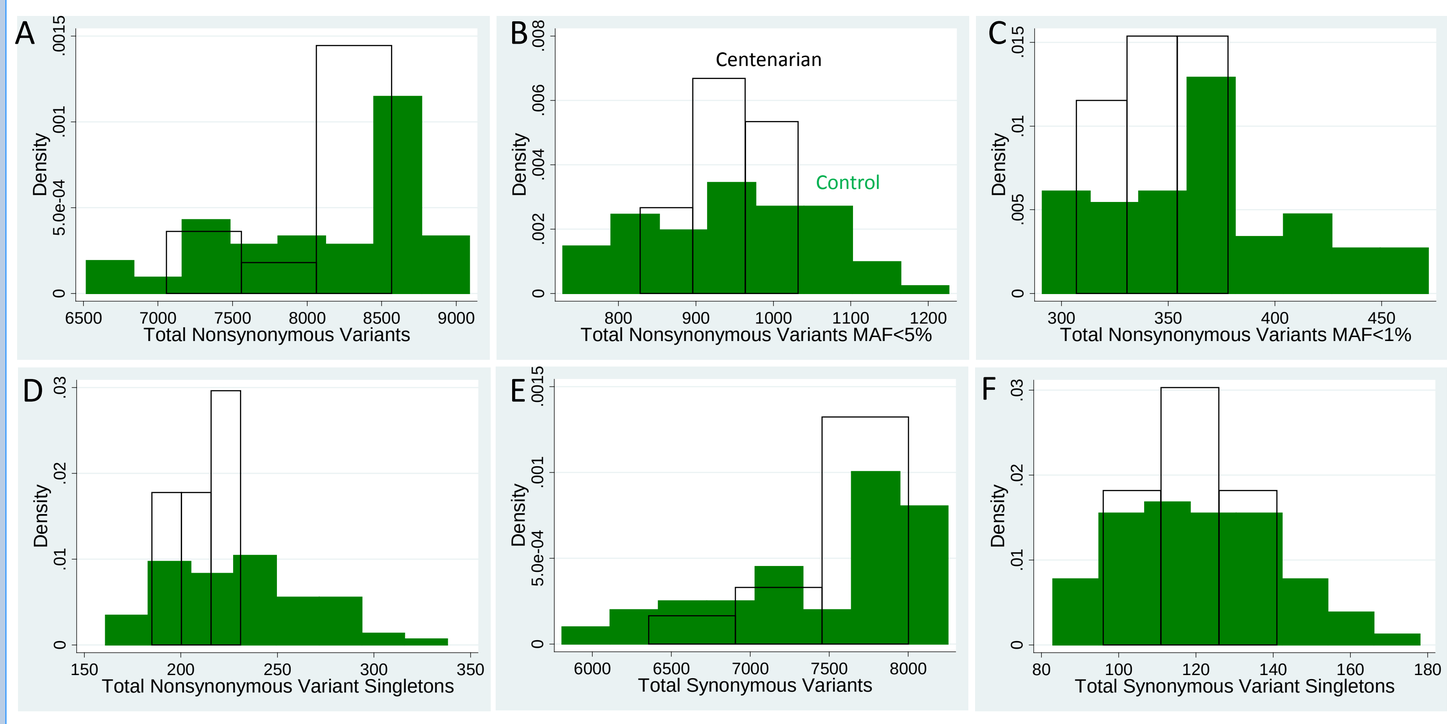

Figure 2. Distribution of variants in centenarians and controls. The centenarians are shown in black with no background, and the controls are shown in green. Shown are (A) the total number of nonsynonymous variants found in each person, (B) the number with MAF<5%, (C) the number with MAF<1%, (D) the number that are singletons, (E) the total number of synonymous variants, and (F) the number of synonymous variants that are singletons

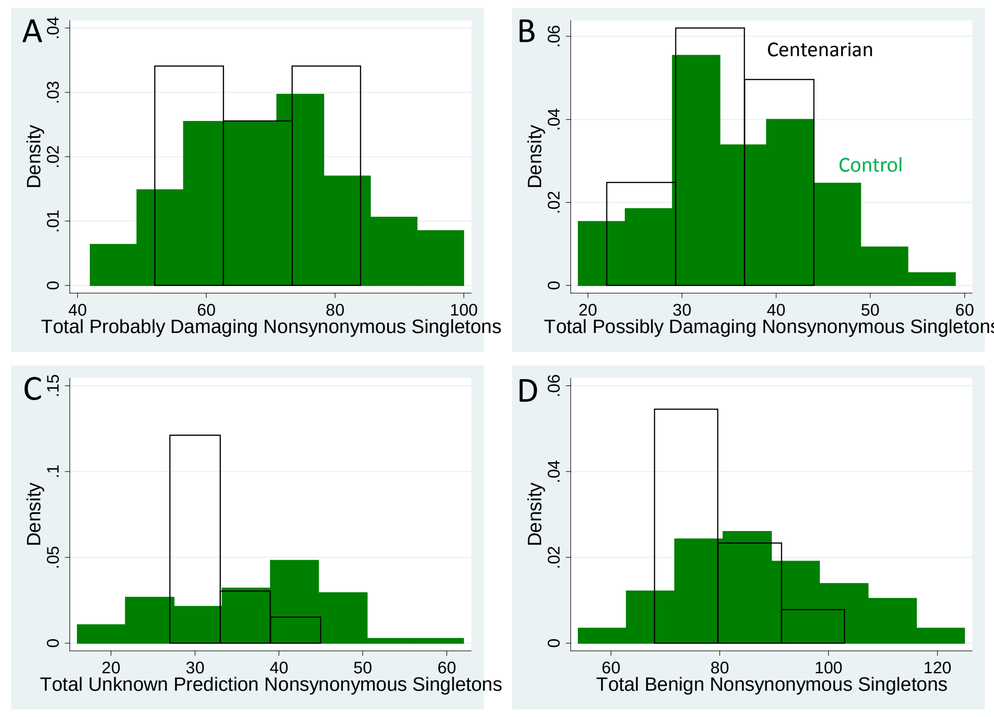

Fig. 3. Distribution of nonsynonymous variants in centenarians and controls. Centenarians are shown in black with no fill and controls are shown in green. Shown are (A) the number of nonsynonymous singletons in each person that are predicted to be probably damaging, (B), possibly damaging, (C) unknown, and (D) benign. The clearest differences are found in the unknown and benign categories.

Conclusions

•Sequencing platform can affect number of variants called even at similar sequencing levels
  •Important to control for artifacts from sequencing methods in any burden analysis
  •Best if cases and controls are sequenced simultaneously using the same methods
•Trend for centenarians to have fewer rare, nonsynonymous variants than controls
  •However, the association is are driven by nonsynonymous variants predicted by PolyPhen2 [1] to be benign
  •Larger sample size needed to prove trend, investigate its source

Future work

•Increase sample size
•Analyze structural variants
  •Preliminary data does not support fewer or smaller structural variants in centenarians
•Expand African analysis
  •Current small sample does not show trend for African centenarians to have fewer rare, nonsynonymous variants
•Study other types of variants
  •Preliminary study of stop gains, stop loss and splice sites showed no trends in these smaller classes
•Look for individual variants contributing to longevity
  •This will require a larger sample size

Literature cited

1. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods 7(4):248-249 (2010).

Acknowledgments

The centenarians were collected under the Murdock study community registry and biorepository Pro00011196. We thank C. Gumbs, K. Cronin, and D. Hughes for DNA extraction and L. Little, H. Kim, J. Maia, J. Li, A. McKenzie, M. McCall, L. Hong, and C. Campbell for sequencing the samples and processing the resulting data.

For further information, contact *etc3@duke.edu*.

Duke University School of Medicine

CHGV CENTER FOR HUMAN GENOME VARIATION