Science AMA Series: I'm John Novembre and I study the genetic diversity of human populations from an evolutionary perspective by developing and applying computational methods.

Dr$_{J}ohn_{N}ovembre^{1}andr/ScienceAMAs^{1}$

[1]Affiliation not available

April 17, 2023

the WINNOWER

REDDIT

Science AMA Series: I'm John Novembre and I study the genetic diversity of human populations from an evolutionary perspective by developing and applying computational methods.

DR_JOHN_NOVEMBRE R/SCIENCE

ABSTRACT

[removed]

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

DATE RECEIVED:
November 20, 2015

DOI:
10.15200/winn.144793.34806

ARCHIVED:
November 19, 2015

CITATION:
Dr_John_Novembre , r/Science , Science AMA Series: I'm John Novembre and I study the genetic diversity of human populations from an evolutionary perspective by developing and applying computational methods., *The Winnower* 2:e144793.34806 , 2015 , DOI: 10.15200/winn.144793.34806

Dr. Novembre,

Thank you very much for offering this AMA. This is an area in which I have a tremendous interest. I've a simple question for you, which is unfortunately prone to enormous social conflict.

Is the concept of "race" a biological construct or merely a social construct?

If race is biological, how many different human races are there, and how is this determined?

Thank you very much.

Zahn1138

Dear Zahn,

Thanks for the question… It seems severals others touch on this same basic issue. I will do my best here and I will ask for some leniency from everyone up front so that we handle this issue collectively with care. This was a big one and hence the slow first response.

First let me highlight two basic results about human genetic diversity:

1. On average a pair of human DNA sequences varies at approximately 1 out of every 1000 basepairs.
2. When we focus on the basepairs that are variable – if make the assumption that globally all humans are randomly mating and we compute a basic prediction for the number of heterozygotes we will see in a global sample, our answer is only off on average by approximately 10%. (e.g, we can use the famous 2pq of Hardy-Weinberg and compare to the observed heterozygotes we find in real data like that from the 1000 Genomes project).

These are two facts that exemplify how similar humans are to one another across the globe. The best explanation for these facts (and many other related genetic data) is that humans emerged and spread out within Africa and across the globe in a relatively short time period. There simply hasn't been

enough time since we spread across the globe for extensive differences to have accumulated across the genome.

Now, let's get a notch more complicated. If we search hard amongst all the sites in the genome, we will find some sites where the random mating assumption (part 2 above) results in a prediction error much greater than 10%. Alleles at these sites vary across the geography of the globe much more than would be expected from the average. When one asks what kinds of traits these variants affect, you will find many of them are disproportionately found in genes affecting skin pigmentation, eye color, and hair texture. This is stunning and something I've seen in my own data analyses. What it means in the context of thinking about race, is that, when we see in each other skin pigmentation differences or eye color or hair color differences, we are looking at parts of our biology that are outliers from a general genomic average.

Besides variants like those affecting skin pigmentation, one also sees dietary trait variants (like lactose tolerance) and immunologically-related ones (like the sickle-cell variant) can have elevated levels of differentiation. These phenotypes all have something in common – which is that they have likely been under selection in response to selective pressures that vary across the globe (e.g. because humans have been exposed to different sunlight regimes, different diets, and different pathogens depending on where they have lived).

So, race to me, as I see it used in the world today and in US census categories, is something much more driven by historical legacy than biological understanding – it stems from a legacy based on judging a small number of external characteristics that hide the great amount of genetic similarity that exists under the surface.

Hey John! Thanks for taking the time! We all know about the lactose intolerance story... what's your favorite signature of evolution in the human genome?

p1percub

The story of salivary amylase copy number variations in humans is a fun one. P.J. Perry and his colleagues found the number of copies of the amylase gene varies among humans and the average number is slightly higher in populations that have had longer histories of consuming diets heavy in starch. My colleagues Graham Coop and Jonathan Pritchard and I wrote a review of the work many years ago which we enjoyed titling: "Adaptive drool in the gene pool"

Hello John, thank you for coming here to answer questions.

Are there any geographic results that are truly puzzling? For example, some tribes of native Americans with unexpected ties to other geographical areas that may seem unlikely, or any other anomalies? In general, what are the big unsolved mysteries or unlikely results in this field or in your work?

semitones

Thanks - this is another big one.

Some of my research recently has stemmed from the surprising observation by several colleagues that the closest living relatives to Otzi the Iceman (a Copper Age man whose remains were found in the Tyrolean Alps) are in fact Sardinians. A model has emerged to explain this but it's still being refined and we'd like to understand the history in more detail.

There are many other interesting facts that are similar but here are some of the bigger issues for the field:

- Understanding archaic introgression - Who were the archaic humans that contributed to genetic diversity? How many groups / where were they living / what were their population histories? How in more detail did modern humans come into contact with them?
- Reconciling the smooth geographic patterns seen in modern DNA data for instance, in levels of heterozygosity and in PCA plots, with the dynamic history of population movements being suggested by a bevy of recent ancient DNA papers. Are the smooth patterns the result of many population movements continually overlaid on each other?
- Regarding specific area histories - what have been the major important demographic events and movements? We have a long way to go. Even in Europe - which is arguably the best studied region - the field is still only beginning to piece together the history, with major adjustments to the basic models emerging in just the past few years.

Dr. Novembre

I wanted to ask how you see genomic field overcoming the analytical limitations due to difficulty in sequence assembly. Has the field progressed significantly since this Alkan et al 2011 paper demonstrating a large number of exons were missed in human *de novo* assembly? Do you think that improvements in the longer read sequencing technologies (Pac Bio, Oxford nanopore, etc) are the answer? Do you think it is more related to necessary advancements in bioinformatics? Or something entirely different?

In particular, are you aware of any techniques to address the issues caused by heterozygosity? As a student working in a non-model organism, it seems the answer to that question would come more easily in human systems.

Thank you!

kto_jest

Sequence assembly is still a hard problem, but progress is being made. The solution is partly technological - as you mention, new technologies with longer reads are emerging and they are helping. I am most excited about some of the computational and bioinformatic developments to make progress on this problem. One solution is not to attempt full de novo assembly, nor map reads to a single human reference, but to map reads to a human sequence variation graph. The idea is that a single graph structure can represent known human sequence variants (both single nucleotide variants and complex structural polymorphisms) and aligning reads to such graphs will result in much less reference bias than using a single linear reference. This will allow us to move forward for the time being until high quality de novo assemblies become affordable to do on every sample.

What are the main computational methods that you use to help with your research? As a computer science major, I've experimented with genetic algorithms and genetic programming. Is it anything like that, or is it more of data mining and analysis of data?

Marthinwurer

Great question! Ironically, we rarely use genetic algorithms or genetic programming in our work - and I remember being surprised when I learned that getting into the field.

Broadly there are 3 main important quantitative/computational skillsets that get used in our work: probabilistic modeling (e.g. Markov models of how populations evolve through time); statistical inference, especially using tools of computational statistics (e.g. Monte Carlo methods, hierarchical

modeling, graphical models); and computational algorithms/structures (e.g. using dynamic programming algorithms, numerical optimization tricks, clever data structures for compressing data). Data visualization is another area I would add (e.g. using PCA or clustering methods or simple geographic maps).

So - it's really a mix of computer science, statistics, probability - though we can't forget about the biology!

Thank you for doing this AMA.

When referring to the fourth, unnamed, branch of humans what are they called in your field? The graphic on the referenced page indicates the fourth branch mixed with Denosivans, and Denosivans with modern Oceana and Asia. Have hints of this fourth branch turned up in living modern humans?

catharticwhoosh

I haven't looked at this deeply myself but from those I know who have, the consensus seems to be there is increasing evidence of this "fourth branch". I would caution it's still at the edge of what is known and a topic of considerable uncertainty.

Is visualising these obviously complicated genetic variations difficult? What kind of tools do you use to tackle this problem? I imagine you're dealing with a lot of variables here.

Edit: Spelling.

biledemon85

Yes - it is a large number of variables indeed!

For visualization we use several tools: 1) Ordination methods, like principal components analysis that reduce the millions or 100s of thousands of dimensions we study (i.e. each variant is a dimension in our data) - down to a small number we can plot and explore visually.

2) Model-based clustering methods, like the program structure or admixture, which model individuals as proportionally descended from a mixture of different source populations. The resulting mixture proportions are a useful visual summary of structure in the data.

3) Tree-based methods (e.g. treemix or mixmapper software): That try to infer a tree that describes the history of how populations are related, sometimes with additional edges to represent migration/admixture events between population lineages.

4) We are increasingly using a new geographic method we are developing that fits an "effective migration surface" over space. It can show where long-term effective barriers and corridors to gene flow are.

5) Compressing the data down to the "frequency spectrum" of variants - i.e. a histogram showing the number of sites found whose frequencies fall within a binned range of allele frequencies; with bins spanning from 0 to 1.

6) Basic geographic maps of variants. For example, see this site we've made: http://popgen.uchicago.edu/ggv/

Thank you for doing this ama! My son has a very rare de novo (not carried by either parent) genetic

mutation that is believed to cause his mental disability. We have been told that one of us may still carry the gene despite it being a de novo mutation if one of us has some form of mosaicism.

Based on your experience do you believe genetic mosaicism is more common than we initially have thought?

tavenger5

Thanks for your message. I wish the best for you and your son, but unfortunately I haven't done research related to mosaicism, so I can't speak with any authority on this question. What I can say is that as more fine-grained data is emerging, genetics as a field is finding some of the basic processes are more complicated than initially thought and there are often rare cases that stray from the general patterns. Regarding de novo mutations, two recent surprises are a number of studies suggesting human mutation rates are lower than thought in the past, and that there is a paternal age effect. I suspect there is still much more to learn...

Hi John, thank you for doing this AMA!

Human evolutionary genetics is a fascinating topic and has the power to drive meaningful insights into human history and disease. But as I am sure you are all too aware, there is a nasty subset of our culture which is obsessed with the prospect of using human evolutionary genetics research to advance racist agendas. They show PCA-derived synthetic maps from research such as this and use it to justify their views on race. I was hoping you can comment on what 'race' means to you, and what it means that it is possible to allot people to different groups using genetic markers. I've always been fond of the quote from Johnathan Marks:

correlations between geographical areas and genetics obviously exist in human populations, but what is unclear is what this has to do with 'race' as that term has been used through much of the twentieth century - the mere fact that we can find groups to be different and can reliably allot people to them is trivial. Again, the point of the theory of race was to discover large clusters of people that are principally homogeneous within and heterogeneous between, contrasting groups. Lewontin's analysis shows that such groups do not exist in the human species, and Edwards' critique does not contradict that interpretation.

but I would love to hear your thoughts as well. Thanks!

SirT6

That's a great quote!

I responded to a related question above - so hope you don't mind reading there for my thoughts on this issue.