Statistical Modeling, Causal Inference, and Social Science

Gay gene tabloid hype update

Posted by Andrew on 10 October 2015, 8:25 pm

Tuck Ngun, one of the researchers involved in the “Twin study reveals five DNA markers that are associated with sexual orientation” project, posted a disagreement with some criticisms relayed by science reporter Ed Yong. I’d thought Yong’s points were pretty good and I was interested in seeing what Ngun had to say. Ngun wrote:

> I wanted to clarify and correct some of the claims made about my work in an article in the Atlantic by Ed Yong. . . . I have reached out to the Atlantic via Twitter about this but have heard nothing back as of this posting.

I hope nobody is ever reaching out to me via Twitter. You can reach out all you want but I won’t hear you!

Ngun continues:

> Ed’s claim that inappropriate statistics were used is not credible because he clearly misunderstood the analytical procedure. . . . All models (from the very first to the final one) were built using JUST the training data. Only after we had created the model did we test their performance on the test data (the algorithm didn’t ‘see’ these during model creation). If performance was unsatisfactory, we remade the model by selecting a different set of predictors/features/data based on information from the TRAINING set and then reevaluating on the test set. This approach is used widely in statistical/predictive modeling field. . . . If this approach is wrong, someone needs to tell Amazon, Netflix, Google, and just about everyone doing statistical modeling and machine learning.

Noooooooo! The problem is here:

> If performance was unsatisfactory, we remade the model by selecting a different set of predictors/features/data based on information from the TRAINING set and then reevaluating on the test set.

Wrong! Once you go back like that, you've violated the "test set" principle.

Now let me say right here that I think the whole training/test-set idea has serious limitations, especially when you're working with n=47. But if you want to play the training/test-set game and the p-value game, you should do it right. Otherwise your p-values don't mean what you think they do.

Ngun continues:

> The second issue I want to discuss is his claim that we needed multiple testing correction. Again he is misunderstanding the approach and rationale. We did not need to correct for multiple testing because we did one hypothesis test. We are not testing whether each of the 6000 marks/loci are significantly associated with sexual orientation. If we had done that, multiple testing correction would have certainly been warranted. But we didn't. The single test we did was to ask whether the final model we had built was performing better than random guessing. It seemed to be because its p-value was below the nearly universal statistical threshold of 0.05.

Ngun is, I believe, making the now-classic garden-of-forking-paths error. Sure, you only did one test on your data. But had the data been different, you would've done a different test (because your remaking of the model, as described in the above quote, would've been different). Hence your p-value is not as stated. See page 2 of this paper.

The big issue

All this is garden-variety statistical misunderstanding, which in many was is excusable. I'm a statistician and I don't understand biology very well, and so it's reasonable enough that a biologist can make some statistical errors.

And at this point I'd usually say the problem is with the scientific publication process, that errors get past peer review, we need post-publication review, etc.

But . . . in this case there is no paper! No publication, not even a preprint.

Talk is cheap. I want to see what Ngun and his colleagues actually did. (I'd say "I want to see the raw data" but I'm no expert in genetics so I don't know that I'd know what to do with the raw data!)

As I wrote in my earlier post: Why should we believe these headlines? Because someone from a respected university gave a conference talk on it? That's not enough: conference talks are full of speculative research efforts. Because it was featured in a news article in Nature? No.

Ngun is providing no evidence at all. I think a healthy skepticism is the appropriate attitude to take when someone makes bold claims based on an n=47 study. Ngun may well feel that he did the statistics right, but how can we possibly judge? Lots of people think they did the statistics right when they didn't. I'm guessing Daryl Bem thought he didn't have multiple comparisons problems either. I agree with Thomas Lumley that it's pretty ridiculous to be having this discussion when there's no actual document by Ngun and his collaborators saying what they did.

To me, the key part of Ngun's note is the following:

> I [Ngun] would have appreciated the chance to explain the analytical procedure in much more detail than was possible during my 10-minute talk but he didn't give me the option.

I'm sorry but that's just ridiculous. Ngun can give all of us the option by just writing up what he did and posting the preprint.

I have some sympathy for Ngun, as this is a tough position for him to be in. It seems kinda weird to me for there to be this high-profile talk without any preprint. Maybe that's how they do things in biology. But it seems a bit much to complain that someone isn't giving you the option to explain your procedure in detail, when you and your colleagues are free to write it up at any time.

P.S. I did not agree with the content of Ngun's note, but I found its tone to be pleasant. I agree with him that vigorous criticism is fine and should not be taken personally. Next step is to dial down the defensiveness and realize that (a) statistics can be tricky, and (b) if you don't make a preprint available, you can't really blame people for doing their best to guess at what you've been doing.

P.P.S. I see from Lumley that the publicity ball got rolling via a press release from the American Society of Human Genetics which includes an interview and a publicity photo of Ngun and a link to an abstract. Based on the abstract and what Ngun wrote on his webpage, it looks like overfitting to me. But, again, we don't really have enough information to judge. In general it seems like you're asking for trouble when you start publicizing technical claims without supplying the accompanying evidence. Everything seems to depend on trust—or perhaps the fear of getting scooped by another news outlet. If you're Nature and the American Society of Human Genetics emits a press release, and you know it's gonna get covered by the Daily Mail etc., there's some pressure to run the story.

But there are incentives in the other direction, too. If you keep with the hype hype hype, and the word gets out that Nature is a less reliable source for science news than BuzzFeed or the Atlantic, that's not so good for your brand.

Filed under Zombies
| Permalink

59 Comments

1. *Mike* says:
   October 10, 2015 at 8:47 pm

   "Now let me say right here that I think the whole training/test-set idea has serious limitations, especially when you're working with n=4."

   care to unpack this?

   - *Andrew* says:
     October 10, 2015 at 9:01 pm

     Mike:

     Sorry—that was a typo, it was supposed to be n=47. The point is that if you hide a big chunk of your data when you're fitting your model, you're throwing away valuable information. According to Ngun's description, though, it seems that they really did use all the data to fit their model, so this is less a concern. But then they lose the integrity of the training/test thing.

     If you're Google or Netflix and you have a million data points, then it's another story. Dividing into multiple stages of training and test sets makes a lot of sense. But when it's n=47 and you're just feeling around, trying to figure out what to do, I don't think the training/test thing makes sense at all.

- *Rahul* says:
  October 11, 2015 at 8:24 am

  Doesn’t n-fold cross validation address this issue to some extent? Also bagging & boosting?

  - *Andrew* says:
    October 11, 2015 at 8:57 am

    Rahul:

    Yes, these methods are all fine, but they eliminate the “wall” between training and test data, hence to get standard errors, p-values, etc., you have to analyze the entire statistical procedure, not simply naively look at the winning model alone, as Ngun seems to have done.

    - *Mike (another one)* says:
      October 11, 2015 at 9:17 am

      Andrew,

      Frank Harrell suggests using the ‘Efron-Gong optimism bootstrap’, in which you use all data to estimate your model, and then do some bootstrapping procedure to estimate how ‘opimistic’ the in-sample fit (e.g., the R^2) is. What is your opinion on that method?

      - *Andrew* says:
        October 11, 2015 at 9:21 am

        Mike:

        As always with the bootstrap, the key question is, what’s your estimator? Except in the easiest and cleanest problems where it doesn’t matter what you do, I’d want to use an estimator that does regularization and partial pooling, so I’d fit a Bayesian model as this is the easiest way for me to do all that. Then if I’m interested in evaluating the predictive error of my model I’d use cross-validation as described in this recent paper with Aki and Jonah. I suppose that the method you describe would do something similar (although requiring more computational effort) if you used an estimator that was similar to what comes out of my Bayesian inference.

      - *Andrew* says:
        October 11, 2015 at 9:22 am

        P.S. It’s fun how comment threads can go in unexpected and interesting directions.

- *Ruben* says:
  October 11, 2015 at 8:33 am

  “Someone should tell Amazon, Netflix, Google”?
  Authors by Microsoft, IBM, Google and others suggest this differential privacy-based approach to avoid overfitting to the test set:
  http://www.sciencemag.org/content/349/6248/636

Or here for a blog post illustrating overfitting to the test set on Kaggle: http://blog.mrtz.org/2015/03/09/competition.html

I think that the training/test set approach should be more widely used, even when data is not humongous (but 47 is indeed stretching it), I'm not sure if you disagree?

The reason is pedagogical: You (AG) are able to adjudicate between various data-dependent decisions, do model selection as part of the analysis, adjust "multiple comparisons" via hyperpriors. All that jazz. I hope to learn it too. But for many people who were trained badly (includes myself) and are too old for way more training (hopefully not including myself), this may be simpler to learn:
Explore all you want, try different models on for size, make gut decisions, do whatever you think is a good way to check you've done it right. And then test it on the test set.

That is so much simpler than exhaustive cross-validation etc. Just lock up one half and give the password to a friend.
Of course a simple train/test split will usually fail, if the methods used on the training set overfit, but seeing this with their own eyes may teach people better statistics in the long run.

- *Andrew* says:
  October 11, 2015 at 9:01 am

  Ruben:

  In my own experience, I have found cross-validation to be an effective convincer (to myself and others) that a particular method really works; see for example figure 3 of this paper from 1996. When I'm really trying to fit a model I will use all the data. But, sure, if you have millions of data points, it can make a lot of sense to divide it into parts and save some of the parts for later. Sometimes we have so many data that we can't analyze them all at once anyway, so why not do some training/test validation? Also, I agree that training/test validation has the potential for reducing the confusion of people such as Tuck Ngun, John Gottman, and Malcolm Gladwell who would otherwise be unduly optimistic about predictive performances of fitted models.

  - *cheese_d* says:
    October 11, 2015 at 4:02 pm

    I'm really not sure why Gladwell was referenced here, but I thought it was quite funny.

    - *Andrew* says:
      October 11, 2015 at 4:14 pm

      Cheese:

      The connection is that Gladwell uncritically promoted Gottman's overfit divorce model in one of his books.

- *ojm* says:
  October 11, 2015 at 4:32 pm

  I think the training/test set distinction can be nice when you're fitting a 'model you like' but it is still likely to underfit on the full dataset. In particular one that is relatively 'simple' but mechanistic/interpretable.

If you try to fit to all the data you are in a sense spreading all the misfit over all the data and hence obscuring what could be interesting discrepancies. By fitting/testing on smaller datasets you can make a sharper comparison and more clearly bring out where the model needs improvement.

- *Andrew* says:
  October 11, 2015 at 9:59 pm

  Ojm:

  If you want to make these sorts of comparisons, I think the way to go is to fit a hierarchical model. There's only so much you can learn from predictive error on random subsets. Cross-validation can be a reasonable way of estimating out-of-sample prediction error, but I think you're putting too much of a burden on cross-validation if you try to use it as a tool for model building.

  - *ojm* says:
    October 12, 2015 at 3:28 am

    I agree in general I think. I suppose I'm thinking of something slightly different. Namely when the subsets are chosen according to information not in the model but which you suspect might be important.

    So really you're looking at p(y|x,s), where s is your subset indicator, but at this point you don't want try to explicitly model that part, rather just see how it might be important or not. Really it's a 'null hypothesis with a model-based null of interest and a graphical check' sort of situation.

    - *Andrew* says:
      October 12, 2015 at 8:43 am

      Ojm,

      Sure, but in that case I'd just fit a hierarchical model. Or, to put it another way, I'd consider the approach you just described as a sort of approximation to a hierarchical model. An approximation that could be useful if it is a lot less work than constructing, fitting, and checking the full model that one might want to use.

      - *ojm* says:
        October 12, 2015 at 10:00 am

        Again I essentially agree. Just to be annoying I'd say that a hierarchical model is of course itself an approximation in that it requires conditional independence assumptions to be satisfied when in principle 'everything might depend on everything' in a more complex manner.

        So at some point you need to draw these lines in the hierarchy according to 'models you like but likely don't capture everything'. What I'm describing is more along the lines of using carefully chosen 'training' and 'testing' subsets for model checking, in particular whether you’ve drawn your boundaries in an acceptable way. I think of this as a type of posterior predictive test but more likely to highlight the misfit than the checks done on the same data you fitted to.

The general intuition that being ‘sharp’ can be good for ‘testing outside the model’ and trying to find misfit while being more ‘continuous’ can be better for estimation ‘within the model’. Isn’t this how you would tend to do model checking? Or would you keep embedding in more and more expanded models?

- *Andrew* says:
  October 12, 2015 at 10:51 am

  Ojm:

  Yes, I agree. I just want to emphasize that this is much different from the usual practice of cross-validation, and certainly different from whatever Ngun and his collaborators did (or indeed anything they could possibly have done given their sample of only 47 pairs).

- *ojm* says:
  October 12, 2015 at 2:01 pm

  sure

2. *amoeba* says:
   October 11, 2015 at 7:58 am

   Very, very few people in biology use preprints. This is unfortunate, but that’s how it (currently) is. I guess 9 times of 10 when somebody gives a conference talk in biology, it is about an unpublished work and there is no preprint. There might be no written up manuscript available until months/years later when it is finally published.

   - *Andrew* says:
     October 11, 2015 at 9:07 am

     Amoeba:

     Wow—any idea why this is? Are the biologists just paranoid about Robert Gallo types in other labs reading the preprint, replicating the experiment, and then beating them to publication? Perhaps the no-preprint-rule makes some twisted kind of sense in a highly competitive lab-science world. But I hope it will change. Certainly I can’t see that there’s anything stopping Ngun and his collaborators from releasing a paper describing exactly what they did.

     - *Ian* says:
       October 11, 2015 at 10:36 am

       I think part of the no-preprint attitude in biology is that biology is just so damn fuzzy. Until the data are final, it’s always possible that the preliminary version is leading you completely astray. That leads to some paranoia, not just of being scooped, but also of making a major public mistake.

(There is also paranoia about being scooped, too. It's not justified in 99% of topics, but there are definitely times when it has happened.)

So the obvious question is, if your preliminary data are so fuzzy and untrustworthy, why are you presenting them at conferences? And again, I think this may be a culture difference, because biologists don't tend to consider conference presentations as particularly trustworthy. Until the data are published, the attitude toward a conference presentation — and especially a 10-minute grad student talk or poster — is that it's mildly interesting but not worth paying much attention to until it's published. This is different from medical conferences, where presentations are often much more complete, and I think very different from, say, physics or computer conferences, where near-complete work is more likely to be shown.

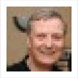
So what may have happened here is that these data got presented with the attitude that this is kind of interesting, let's see where it goes, and then journalists jumped on it and pushed it as a finished product. It's as if a journalist was sitting in at a brainstorming session between a couple people and then published the wildest what-if ideas as finished proofs.

We could ask why biology conferences are OK with these half-finished, incomplete presentations, many of which will never go anywhere or at best be wildly changed before finalization. First, there are also final versions presented at conferences; these tend to be flagged (the author will specifically note that the paper is submitted or published). But the driving force for much of the crap is the widespread attitude that the only way a student or postdoc can attend a conference is if they have a presentation to give. No presentation? We won't pay for your travel and fees. That's a terrible attitude, because it means the people who are likely to benefit most from a conference (the very early-stage students) either can't attend, or have to present embryonic, incomplete studies.

… I could rant about this more, but you see the problem. Structural issues in biology have led to the acceptance of presentations that everyone knows are weak and incomplete, that you're not really supposed to take seriously. And then you end up with this occasional amplification of these things when media get hold of them. It's not the presenters' fault, it's the result of a series of dominoes falling, where no one really wanted to be int this situation but no one wants to be the first to change.

(As a PI, my students and post-docs never needed to present or give posters if they wanted to go to a conference, but I was in a tiny minority.)

- *Jim Woodgett* says:
  October 11, 2015 at 11:34 am

  There is now a concerted effort for biologists to follow the preprint model of physicists and mathematicians with the launch of BioArxiv, a Cold Spring Harbour initiative. It is gaining traction and is modeled on Arxiv but many biologists are reticent to submit preprints because some journals (stupidly) consider this as competitive publication. As behaviour said change, perhaps this stance by some publishers will be reversed.

  There is, unfortunately, a trend in many conferences now to speak only about work that's published (or accepted for publication). Conferences should be places for discussion of on-going work. However, in this case the presenter knew his topic was contentious and likely to attract attention and should have been ready to provide methodology and perspective for the claims. While one could blame Nature, etc. for putting the talk into the headlines, don't blame the messenger.

BTW Andrew, thanks for this very insightful analysis. This is exactly how science should be conducted through discussion and exchange from various perspectives.

- *Michael Hoffman* says:
  October 11, 2015 at 12:20 pm

  It's worth noting that many of the most prominent biological journals and presses are OK with posting pre-prints on bioRxiv. For example, Nature, Science, Cell Press, Proceedings of the National Academy of Science, Springer, Cambridge University Press, Cold Spring Harbor Press, Rockefeller University Press. That's a non-exhaustive list and a more complete and up-to-date list is available at:

  https://en.wikipedia.org/wiki/List_of_academic_journals_by_preprint_policy

  Some of these policies have changed in just the last couple of years and not all authors are aware of it. I see bioRxiv has already gained a lot of traction fast though, especially in some areas like genetics and genomics.

  Competing interests: I am a bioRxiv "affiliate". (This means I recommend bioRxiv submissions for posting and am occasionally asked for advice on their policies.)

- *Steen* says:
  October 11, 2015 at 12:38 pm

  For whatever reason CSH press decided to name their preprint server 'bioRxiv' with no 'a'. (Perhaps it should be pronounced 'björ-kive'). The capital 'R' does indicate an emphasis on analyses done with R, unfortunately. Cf. the capital 'X' in 'arXiv, which archives TeX source.

  The Cell Press journals (owned by Elsevier and including 'Cell', which publishes very long papers and equals Science and Nature in prestige) still consider preprint deposit previous publication. Other journals (e.g. Genome Research, Plant Cell) only changed their policy very recently. More here: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001563#s3

  Some have suggested that the biomedical research community has been slow because it is so large and fragmented. Other biologists feel that peer review frequently catches errors that they would prefer not be made public.

  The recent enthusiasm for biology preprints I have seen has been driven by perceived increases in review time and numbers of figure panels and tables required. See e.g. http://doi.org/10.1101/022368. The newish high profile journal eLife was founded partly as a response to this. I think we'll see some arXiv/bioRxiv overlay journals in the near future, which might drive adoption, at least for non-flashy papers.

The population geneticists have been leading the efforts to share preprints, possibly because they have seen benefits in applied math fields. A few started a blog ‘Haldane’s sieve’ to encourage discussion even before bioRxiv was started.

Andrew, thanks for commenting on this so quickly. It would be great if you’d reduce the blog lag, just a little bit. It seems like the Stan team et al. have been posting here more often, which has been great.

- *Andrew* says:
  October 11, 2015 at 12:46 pm

  Steen:

  I could reduce blog lag by just publishing, all at once, the 100 posts I have currently scheduled. But I’m afraid then people wouldn’t read them all!

  The Stan team have my permission and encouragement to intersperse posts as often as they like. No lag.

  I used to post more often on the sister blog, but recently I’ve been finding it exhausting because they keep telling me that my posts are too bloggy, and I just don’t have the energy to be rewriting things to satisfy their hypothetical reader.

  - *Rahul* says:
    October 11, 2015 at 3:29 pm

    What if you went 2-posts a day till you got the lag under control?

- *Michael Hoffman* says:
  October 11, 2015 at 2:44 pm

  Here is what Cell says:

  “If you have questions about whether posting a manuscript or data that you plan to submit to this journal on an openly available preprint server or poster repository would affect consideration, we encourage you to contact an editor so that we may provide more specific guidance. In many cases, posting will be possible.”

  http://www.cell.com/cell/authors

  - *Steen* says:
    October 11, 2015 at 2:55 pm

    Thanks for clarifying.

- *Steen* says:
  October 11, 2015 at 2:58 pm

whoops—typo: bioRxiv does NOT have ‘an emphasis on analyses done with R’.

    - *Martha* says:
      October 12, 2015 at 12:19 am

      I looked it up (http://biorxiv.org/about-biorxiv) As I had guessed, “bioRxiv” is pronounced “bioarchive” — the R as it is pronounced when reciting the alphabet.

  - *Ian* says:
    October 11, 2015 at 5:06 pm

    To be honest, I think most biologists are not submitting to BioArxiv because they’ve never heard of it. I really think that proponents of preprints and open publishing (I know they’re not the same thing, but there’s a lot of overlap) hugely overestimate their influence, and hugely underestimate the vast numbers of professional biologists who don’t know anything about the subject. It’s not that they’re resistant to the notion or worried about journals not accepting their articles, it’s just that the debate may as well be published on clay tablets in cuneiform for all they know.

    - *Rahul* says:
      October 12, 2015 at 1:30 am

      +1 It’s getting that critical mass of adopters that’s the challenge. Once people realize others they care about will read it there, the posting will start taking off.

- *Rahul* says:
  October 11, 2015 at 11:00 am

  There’s very few preprints in Chemistry too, at least in the areas I’ve been familiar with.

- *Konrad* says:
  October 12, 2015 at 4:26 am

  Yes — and there’s some prominent precedence (see the tRNA structure controversy is one example). The attitude is of course very nonconducive to scientific discussions. It’s changing slowly, however. There’s finally a dedicated preprint server for biology papers (bioRxiv) and people are starting to produce their research in the open.

- *IanSudbery* says:
  October 12, 2015 at 8:53 am

  It can take years to get a biology paper to a place where it has a sufficiently complete story to write a paper. Most PhD students will only ever have one paper (if that, in the UK its fairly normal for it to take several years after your PhD is finished to get it published). Those people that do use preprints tend to publish completed papers that they are ready to submit for peer-review, so as to 1) not let them be delayed by peer-review 2) get round any open access problems.

But this means not talking about your results for 3 or 4 years at once. Hence conferences are seen as places where you can talk about your results before you have a complete enough story to write a paper, preprint or peer-review alike. Indeed, some conferences will only let you speak if you promise to only talk about unpublished work.

- *Martha* says:
  October 12, 2015 at 11:58 am

  "It can take years to get a biology paper to a place where it has a sufficiently complete story to write a paper. Most PhD students will only ever have one paper (if that …"

  This is very different from what is the case in biology Ph.D. committees I've been on. Typically, at least one chapter of the dissertation is based on a paper that has already appeared.

3. *Xi'an* says:
   October 11, 2015 at 8:56 am

   p-values NEVER mean what you think they do..!

   - *Andrew* says:
     October 11, 2015 at 9:08 am

     X:

     See the section "A p-value that worked" on page 71 of this paper.

   - *Martha* says:
     October 12, 2015 at 12:30 am

     "What never?" "No never." "What never?" "Well, hardly ever!" (from HMS Pinafore)

     Andrew: I think you mean p. 70

     - *Andrew* says:
       October 12, 2015 at 8:42 am

       Yes, p.70.

4. *nr* says:
   October 12, 2015 at 1:45 am

   Is he saying he made multiple models till he found one that proved his point?

   He could've just said that at his presentation, I still think there would be some value in showing the different models for discussion. But that is unlikely to have ended up with getting a mention on the Nature website.

   - *Anoneuoid* says:
     October 13, 2015 at 11:40 am

     Yes, this is what it sounded like to me and others as well. His site has been updated to read:

“pjie2 asked:
Your response seems to be self-contradictory. First, you say “If performance was unsatisfactory, we remade the model…”, but then “The single test we did was to ask whether the final model we had built was performing better than random guessing.” How often did you ‘remake’ the model, i.e. how many versions of the model did you try? This is the bit that needs a multiple testing correction. Getting p=0.05 on the best out of 20 different versions of the model is not a significant result.

Hi pjie2,

My collaborators and I are going to issue a joint statement to clear this all up. Until then I have been advised to not talk about this issue on social media. Look out for the statement!”
http://vizbang.tumblr.com/

It is far too common in biomed to read a paper that seemingly makes no sense at all like this (maybe the actual methods made sense but the description was just poor, who knows?). Eventually the time comes to conclude either you are crazy, or something is seriously wrong with the education and publishing system in this area, worldwide.

- *Steen* says:
  October 21, 2015 at 5:53 pm

  And again! http://www.nature.com/news/abstract-thoughts-1.18596

  There is an indirect reference to this post: “at least one [critic] suggested that the authors could have provided preprints of their study when presenting it. These arguments seem to misunderstand the traditional, and still useful and relevant, role of such gatherings.”

  Ha!

  - *Andrew* says:
    October 21, 2015 at 6:43 pm

    Steen:

    Indeed this is ridiculous. If you can have an informal talk, why not an informal preprint? Or why not hold off on the hype (which of course was supplied by Nature itself!)?

○ *Steen* says:
October 15, 2015 at 8:46 pm

This post now featured in the tabloids: http://www.nature.com/news/preprints-called-on-to-support-controversial-talks-1.18568
Congrats! The buzzfeed-style post titles are paying off.

- *Andrew* says:
  October 15, 2015 at 8:57 pm

  3 Secret Ways to Get Traffic—And You’ll Never Guess #2!

5. *Dean Hamer* says:
   October 16, 2015 at 4:21 am

The detailed discussion of the statistical methodology is technically interesting but does not address the real question, which is whether or not there is in fact an association between the five identified epigenes and sexual orientation. The only way to test that is by analyzing an independent series of unrelated gay and straight individuals. The statistical debate also fails to consider the sizes of the detected differences, which is key in understanding whether or not they are biologically meaningful.

When I was doing research in this field, preprints were shared only with trusted colleagues; given the intense competitiveness of molecular genetics that's still a reasonable policy. But it was also strictly observed that preliminary results, whether privately shared or presented for discussion at a conference, should not be publicized or even spoken about with the press. "I'll be glad to talk with you about our results when the paper is accepted" was and still is the correct response to a reporter asking about unpublished work.

The real problem is that the ASHG issued a press release without consulting the PI of the project. That's not cool.

- *Anoneuoid* says:
  October 20, 2015 at 7:59 am

  >"The detailed discussion of the statistical methodology is technically interesting but does not address the real question, which is whether or not there is in fact an association between the five identified epigenes and sexual orientation."

  Taking the description of the methods at face value, there seems to be no new information added here.

  Say I go to https://en.wikipedia.org/wiki/List_of_human_genes, and pick the first five I see (ALB, BCL2, CCR5, CD4, CD8). Then I say the expression of those genes may be associated with sexual orientation. I could even come up with a story ("wow interesting that they seem to be related to the immune system, and CCR5 is said to be a receptor for HIV"). Do we really need to discuss whether this association exists?

6. *rijkswaanvijand* says:
   October 17, 2015 at 5:59 am

   "the real question, which is whether or not there is in fact an association between the five identified epigenes and sexual orientation."
   Not if you want to conclude the existence of a 'gay gene'; correlation does not equal causation.

   Besides, isn't gay mostly a social construct?

   Isn't it more probable that certain genetic traits correlate to rather general character/personality traits involved in social interaction, traits which guide sexual identification -and thus sexual orientation- throughout this interaction with the social environment a society represents?

   Why are 'serious scientists' even looking for gay genes?

7. *Dean Hamer* says:
   October 20, 2015 at 4:21 am

   Right, "gay is a social construct", just like heterosexuality. And sure, sexual orientation is a correlate of some personality trait, just like in every other species that reproduces by sex. LOL…I thought that at least on this site there would be a rational discussion, but I guess when it comes to sex, that's rarely the case.

- *James* says:
  October 20, 2015 at 8:17 am

  I think sometimes there is a tendency for people to conflate the ‘socially constructed’ aspects of an area of research such as all of the naïve beliefs surrounding homosexuality with the actual phenomenon itself. Understanding sexual behavior at a genetic level is a worthy pursuit as is understanding the social aspects that influence the expression of those genes.

  That said, from a political and policy perspective, the biological aspect is properly seen as unimportant because basic human rights and freedoms make it irrelevant. Adults should be able to live, sleep with, and marry any other consenting adults of their choosing irrespective of gender and sexual orientation. That should be a foundational freedom of any civilized society.

  - *Dean Hamer* says:
    October 20, 2015 at 1:04 pm

    I totally agree. Gay rights are fundamental, inalienable rights to which we are all entitled in recognition of our dignity as members of the human family.

    But while science must never play a role in determining these human rights, understanding the biological roots of sexuality does have tangible impact on peoples understanding of the LGBT community; in fact, survey research shows that beliefs about the origins of sexual orientation are the single most important factor in predicting a person’s attitude toward LGBT acceptance. So like it or not, this research does play a role in shaping attitudes, laws, and, ultimately, the ability of people to live free and open lives. To ignore this fact is a disservice to both science and human rights. See: http://www.the-scientist.com/?articles.view/articleNo/40299/title/Going-Beyond-the-Lab/

    - *Jeff Walker* says:
      October 20, 2015 at 1:32 pm

      “in fact, survey research shows that beliefs about the origins of sexual orientation are the single most important factor in predicting a person’s attitude toward LGBT acceptance. So like it or not, this research does play a role in shaping attitudes, laws, and, ultimately, the ability of people to live free and open lives.”

      This of course assumes that our beliefs about the origin of sexual preferences *causes* and not just predicts our attitudes to LGBT. I’m guessing that good evidence for this is rare to non-existent. Science is full of many uncomfortable facts. I would therefore focus on changing people’s attitudes regardless of (or independent of) the science.

      - *James* says:
        October 20, 2015 at 2:05 pm

        Convincing bigots to drop their bigotry is no easy task.

8. *Dean Hamer* says:
   October 21, 2015 at 4:51 am

   Jeff – I agree it’s very difficult to sort out cause from effect when looking at purely correlational data. But if you delve into the literature a bit, you'll see that social scientists have been pursing this question through several approaches, including path analysis (which i don't fully understand, but it seems to analyze what is at the "root" of a series of correlations) and experimental approaches (compare two groups of students +/- information about science of sexual orientation.) I expect there's quite a lot of back and forth between the variable – certainly people with gay friends must be more open to "born that way"! – but at least to some extent the knowledge itself is useful. Anyways, better science than old wives tales!

James – Difficult but not impossible. Check out the pastor in this film I made called OUT IN THE SILENCE a few years ago. He progressed from "if you've ever tried to plumb a house with all male fittings" … to being a friend. Here's the website: http://legacy.wpsu.org/outinthesilence/

- *Andrew* says:
  October 21, 2015 at 8:35 am

  Dean:

  It's worth studying biological correlates of sexual preference. But when people start fishing for correlations in small samples, and then playing around with tools like path analysis, it can end up being nothing more than a search for patterns in noise. It's generally a good idea to do science, but some things that are done by scientists are perhaps unlikely to lead to any new understanding at all. And, unfortunately, this can be true even of studies that are hyped in Nature.

9. *rijkswaanvijand* says:
   October 22, 2015 at 11:54 am

   @James "Understanding sexual behavior at a genetic level is a worthy pursuit as is understanding the social aspects that influence the expression of those genes."
   But in my estimation people seem to conflate the expression of sexual behaviour with the matter of with whom ones sexual behaviour is expressed; the former might easily be correlated to certain gene-expression levels, but the latter seems -to me- an intricate function of behavioural interaction.

   And no, I don't believe a person can be born gay (which is in no way intended as an assault on people with this sexual orientation); it's simply that babies have no sexuality or sexual identity but only a determined sex. It's the tendency to conform or to diverge from socially prescribed gender-roles which determines ones sexuality.
   I do think that without such gender-roles most humans would develop to be bisexual.

   @Dean Hamer "LOL…I thought that at least on this site there would be a rational discussion"
   If you want a fair discussion, maybe you should refrain from the LOLZ and from calling your opponent irrational.

   - *James* says:
     October 22, 2015 at 12:06 pm

     @rijkswaanvijand Your statement rests on a false premise. That if a gene has not been expressed at birth that it is not causally related to development in later years. I do not think any serious scientist could possibly agree with this assumption. A proper analysis of the genetic and the environmental is how, when, and by what process they interact. Broad and sweeping claims about sexuality is not an adequate substitute for careful scientific analysis.

10. *James* says:
October 22, 2015 at 12:07 pm

typo: Broad and sweeping claims about sexuality are not an adequate substitute for careful scientific analysis.