[Dr. Greg Feero]
Good afternoon.
This is Greg Feero, chief of the Genomic Healthcare Branch at the National Human Genome Research Institute.
I'd like to welcome all of you to the third webinar in NHGRI's webinar series.
Today we'll be talking about genome-wide association studies, describing the latest on genome-wide association study results and what they can tell us about genomics and health.
Today we'll be hearing from Teri Manolio and after that we'll be taking questions from you.
The questions will be taken over the phone, and to access the system dial star 1 to speak to the operator and you'll be put in the queue for the questions.
It's now our pleasure to welcome Dr. Teri Manolio, director of the Office of Population Genomics here at the NHGRI.
She is currently senior advisor to the director at NHGRI for Population Genomics.
She's been involved deeply in large-scale cohort studies such as the Cardiovascular Health Study and the Framingham Heart Study.
She joined the NHGRI in 2005 and leads efforts in applying genomic technologies to population research, including the Genetic Association Information Network, GAIN, and the Genes and Environment Initiative, GEI.
Dr. Manolio, I will put your slides up shortly.
[Dr. Teri Manolio]
Super, thank you.
And I'm glad that everyone was able finally to join.
Again, our apologies for the delay.
Because we have lost a few minutes I may skip over a couple of slides.
I hope that doesn't disturb anyone.
And there may be someone who is breathing a bit heavily on the phone, if you could just hit your mute that would be grand.
So moving on then, if you are seeing my first slide, to talk about these being interesting times for doing genome-wide association studies and really looking at the genome in general.
You're probably familiar with Robert Kennedy's quote, "May he live in interesting times.
Like it or not, we live in interesting times," which is actually part of a speech he gave in Cape Town in 1966, well worth reading.
There are two other parts to that proverb that I'll just kind of skip over here for the time being.
And actually, if one were to look at the associations that were known through any kind of, really, genetic studies, there were maybe six or seven of them prior to 2005.
And those were of -- there was some question as to how strong those were.
There were many that had been reported but these six or seven were sort of pretty solid.
But just looking at what has been learned in genome-wide association since 2005, you should be seeing a slide that shows the entire genome and then on chromosome 1 at the bottom of it, a compliment factor, age related to age-related macular degeneration.
And that was reported in March, I believe, of 2005.
And then really nothing much more -- oops -- nothing more until late in 2006, when there were three more associations, as shown here.
2007, things started really to

pick up, and as time went by
we really have sort of filled
out the genome dramatically,
to the point where we're almost
near asking people to stop
working on chromosomes 1 and 6
because there isn't anymore
room on the graph
to show them.
But this work has really led
2007 to be called the year
of genome-wide association
studies because much of the
work really kind of
took off in 2007.
This is a paper in "Science"
at the end of that year,
and just shown here are all of
the diseases and traits that
have published genome-wide
association studies done.
We keep sort of a running
catalog of these,
and there are over 75 of them
as of a couple of days ago,
so it's really going
very, very rapidly.
This has been referred to by
Hunter and Kraft from Harvard
as drinking from the fire hose
and trying to talk about the
massive amounts of data that are
coming out of these studies.
They point out there have been
few, if any, similar bursts
of discovery, really, in the
history of medical research,
and I think most would agree
with that in terms of the
number and rapidity with which
findings have been reported.
So, what is a genome-wide
association study?
Well, it's basically a way for
interrogating all of the
10 million variable sites
across the genome.
So we have three billion
spots in our genome,
letters in the spelling of our
DNA, and about 10 million of
those differ between
any two individuals.
This variation is inherited
in groups or blocks,
so you don't have to test
all 10 million points,
you can test maybe a subset of
those and then infer what the
other, you know, nine million
or whatever are.
The blocks are shorter, so you
have to test more points
the less closely the
people are related.
So, when we started
doing family studies,
they have very close
relationships and so
you might only have needed
400 or 500 markers,
but technology now allows us
to study unrelated people,
assuming that there are much
shorter base pair links
in common so you need
many more markers.
This is just a stretch
of DNA on chromosome 7,
and as you can see at the top,
you know, we're all really
pretty similar in 99.9
percent of the genome.
But every now and then there
will be one that sort of
pops up like this C over A here,
where some people have a C
and some have an A up in that
upper left-hand corner.
And then you go on and everybody
is the same for a while
and then there's a C or a T,
et cetera, and you have these
single nucleotide polymorphisms
about one every 300 bases or so.
This is a nice figure from a
review by Christensen and
Murray last year that basically
took an example chromosome,
just sort of this cartoon up at
the top, and then from there
took a, you know, example gene,
essentially in that sort of
second middle bar that
shows various SNPs;
some of them are in exons,
which are the red sections
of that gene.
Some of them are in introns,
which are the white sections
of that gene.
There tend to be a few more
in the introns than there
are in exons, perhaps because

they are better tolerated
in introns
than exons.
And then you see this
triangular shaped diagram
toward the bottom.
These tend to throw people,
but really what these are
is just the relationship
among each of these SNPs,
each to each other.
And this is essentially a
matrix, and we've all been
looking at matrices like
these for years and years
maybe without
realizing it.
When you ask the AAA for a
road map and a set of --
I'm sorry, here's another
example of one on chromosome 9,
a little bit more extended,
and we'll come back to this
one in a second.
So you ask AAA for a map of the
East Coast and they'll tell you
that driving from Boston
to Providence is 59 miles
and from Boston to
New York is 210 miles
and Providence to New York
is 152 miles, et cetera.
That's the same sort of matrix
as we're looking at with
these SNPs.
And if you wanted to color code
these and say that, you know,
the distances that were really
close, less than 100 miles,
were dark red, and those
that were much further,
say, more than 400 miles,
were white, you could do
that and you could sort of
overlay those colors on
this matrix here.
And if you kind of turned it
on its side and made it into
squares, you'd basically have
the same thing that you're
looking at with a
linkage diagram.
And that's all that we're
looking at when you see
this dark red between
two SNPs.
It's just, you know, if you look
at SNP 3 and 4 in that diagram,
that's just very much like
Boston to Providence
essentially.
So, because of this one tag SNP,
or a SNP that sort of stands up
for several that it's
strongly related to,
can really serve as a
proxy for many of them,
and shown here is a stretch of
DNA on two chromosomes from,
say, one individual and then
the same stretch from another
individual's two chromosomes
and then another individual's
two chromosomes.
And as you can see, this
first SNP here in blue,
SNP number 3, can either be a
G or a C, depending on which
chromosome you're looking at,
and SNP 4 in gold right next
to it, actually moves pretty
much in concert with it.
So when SNP 3 is a G,
SNP 4 is an A,
and every time there's a G at
SNP 3 there's an A at SNP 4.
And likewise, when SNP 3 is a C,
there's a G at SNP 4.
SNP 5, on the other hand,
in bright green,
does not always move together
with SNPs 3 and 4,
so sometimes when SNP 5 is a G,
there's an A in SNP 4.
Sometimes when SNP 5 is a G,
there's a G in SNP 4,
and so on.
SNP 2, just take my word for it
if you don't want to check them
all, but it's also exactly
correlated with SNPs 3 and 4
and so is SNP 1, again,
just in this cartoon.
And these four SNPs could be
said to move as a block,
so these are what are often
known as a haplotype block,
haplotype just being a string of
SNPs of sort of the same flavor
along one stretch
of the genome.
SNP 5 has a SNP next to it,
SNP 6, with which it is in
perfect correlation, also called

linkage disequilibrium,
which is kind of an awful
name but so be it,
that's what it's called.
And then SNP 7 in light blue
here and those three form
another block.
And then there's this SNP sort
of in brown on the side that
kind of moves
by itself.
So if we were to take out the
SNPs in between here and just
focus on the places where people
differ between chromosomes,
you could see that for block one
you could measure any of these
four SNPs and still get all of
the information if you had
measured all of them.
So, you might just
pick one of them.
You could pick the one with the
prettiest colors, I've done,
or you could just probably pick
the one that's either cheapest
or most easy
to type.
And you could also pick any
one of block two and then
the singleton on block three and
you measure three SNPs instead
of, you know, probably 1,000 or
10,000 or so to be able to get
all the information that
you would from all those
different SNPs.
And this just shows how
these kind of break up
into haplotypes and very often
there are just a few haplotypes
that are very common as
these top three are,
and sometimes then there are
others that are much rarer.
So coming up with these blocks
and the way that the SNPs
travel together in the genome
was the whole purpose of the
haplotype map, and the HapMap
Project published its first
paper in 2005 that summarized
over a million SNPs, I believe.
And then in 2007 there was a
follow-up paper that reported
over three million SNPs,
and there will be multiple
follow-up papers after
that as well.
The goals of the HapMap were to
use just the density of SNPs
that you needed to find
associations between the
SNPs and the diseases,
and we'll talk about how
one does that, and trying not to
miss regions that had disease
associations but to produce a
tool that would help in finding
genes that affect health and
disease and recognizing that
one needs to use SNPs for
more complete genome --
you need more SNPs, sorry,
for complete genome coverage
of populations, particularly of
populations of African ancestry,
recent African ancestry since
we're all of African ancestry,
but that's because those
populations are older
and there's been more time for
the relationships between the
SNPs to break up, so you need
to measure more of them.
Along with the HapMap, and
probably stimulated by it,
genotyping technology has
improved dramatically and
the costs have
gone way down.
So in 2001, as the slide from my
colleague Steven Chanock shows,
we thought we were getting a
really good deal if we got
a genotype done by ABI's TaqMan
method for a cost of about
a dollar.
You can see the cost along
the Y-axis there in
cents per genotype.
And those costs have come
down really, you know,
almost linearly into 2005,
as shown here, with various
different platforms also
typing more and more SNPs.
And this continued, the slide
is now two years old,
but, you know, the same trends
continue, where the costs have
just fallen and fallen and
fallen and the numbers of
SNPs on the platforms

have increased as well.
And this is has allowed us then
to do these kinds of studies.
So, what is it exactly that you
test when you're doing this?
Well, say you have a bunch
of people who have had
a myocardial infarction or heart
attack and a bunch of people
who haven't and you'd like
to know how they differ.
And in traditional epidemiology
you would look at things like
their weight or their smoking
history or as time went by
their cholosterol levels or
their blood pressure, et cetera.
Well, one can do the same thing
with genetic factors and just
ask, you know, is a particular
gene or SNP, in this case,
RS1333049, as shown at the top
here, whether the different
forms of that SNP are associated
with being a case of myocardial
infarction or a control without
having myocardial infarction.
And as you can see, the C allele
of this particular SNP is more
common in the cases,
55 percent of the
cases have that SNP compared to
only 47 percent of the controls,
so that suggests -- oh, sorry,
have that allele rather than
the controls.
So that actually one can do a
statistical test on it called
a chi square test and estimate
how likely it is that you would
get -- you would see that
extreme value of a
chi square if there was actually
absolutely no association
and you just saw
that by chance.
And if this was just due to
chance alone, it would be a
very unlikely thing
to have happen.
It would happen only once
in 10 to the -13th times,
so much fewer than a billion
times would you ever see a
result as extreme
as that.
And the odds ratio is sort of
the risk associated with that,
so people who happen to carry
this allele are about
1.38 times more likely to have
a heart attack than the people
who don't carry this allele,
or 38 percent more likely
to have a heart attack.
One could also look at this by
genotype, because each of us
carries two copies of almost
every variant in the body,
except for men who are missing
some of those on the
X chromosome because they only
carry one X chromosome.
But in looking at the genotypes
for this particular SNP,
you can also see that the cases,
31 percent of the cases had
the CC genotype at this SNP
compared to only 23 percent
of the controls.
And then looking at the GG,
gene heterozygotes were
about the same but the GG
genotype is much more common
in the controls than
in the cases.
And again, one can calculate
a chi square value and a
probability associated with
that, and then the heterozygote
odds ratio would be what is
basically the odds on having
disease if you carry one copy of
the variant compared to carrying
no copies, and that's 1.47.
And then for the homozygote it's
1.90, which means you're nearly
twice as likely not to have
disease if you carry
two copies than
if you don't.
The challenge with these studies
is that you basically are doing
this same test 100,000 or
500,000 or a million times,
and the challenge is in
interpreting that massive
data are what make genome-wide
associations so interesting.
So shown here is the very first
truly genome-wide study,
this Klein study that I had
mentioned in looking at
macular degeneration that

was published in 2005.
And they tested 100,000 SNPs
and they set a level --
because they were looking
at so many SNPs they said,
we have to sort of control for
the fact that if we just looked
at, you know, things that
happened one in 20 times
would, you know, be
an unusual occurrence,
you're going to see an awful lot
of those things and those would
be false positives.
So one would want to set a very
sort of stringent level.
We only want to see something
that might happen by chance
one in a million times,
or in this case,
4.8 in 10 million times in
order to be concerned that
it might actually be an
unusual occurrence.
And that was where that arrow
is on the slide here is
chromosome 1 because these are
just lined up along from the
chromosome, the beginning
to the end of the genome,
essentially chromosome 1
to the X chromosome.
And there was a very
strong association.
There's another association
that's plotted along with,
you know, basically the
height of this line here,
and you can see around the
middle of the plot there's
another association that's
almost as strong as that one.
And it turned out that that
one was a genotyping error,
and when they went back and
looked at it very carefully
it was decided not to
be a true association,
and this can be a problem
with these studies.
You can make these -- you know,
show these in all kinds
of different
fancy colors.
Here's a red one looking
at nicotine dependence.
And again, the height of the
points here just shows how
strong the association is,
how unlikely it is to be
due to chance
essentially.
This is a nice multicolored
one of diabetes.
There's one in gray here that
shows each of the chromosomes
sort of separated out for you
and in red the things that
really kind of popped out
and were strongly related.
And here, a blue one, this
one has multiple diseases,
so this was a very extensive
study of seven different
common diseases and they showed
all of their associations in
one plot.
They like to call it the
10 million pound plot,
but at any rate.
This is one where they're sort
of falling from the sky.
This one was done over
Christmas time and that
was sort of what they
had on their mind.
But if one looks a little more
closely at one of these
associations, and this is one,
again, that I mentioned
previously for myocardial
infarction, you can see
that in blue here there is an
area that shows really very
strong association all the
way up to 10 to the -14th.
So one in 10 to the 14th chance
that this could have happened
by chance alone, and that
was that SNP that I showed
you before.
One can take this area on
chromosome 9 and sort of
stretch it out, and that's
this area here that I'm just
highlighting, and if you
sort of stretch it out,
this is the same region and it's
now just looking at chromosome 9
and just focusing on the blue
dots or -- the red dots were
a replication sample.
But this was the finding that
was reported by these authors

and it's in chromosome 9.
And then one can look again
at our old friend,
the red triangles and looking
for how the SNPs that have been
tested in this particular study
are related to each other.
Do they travel together
or don't they?
And as you can see from that
middle panel where you remember
the really dark things were the
Boston to Providence ones,
so those are ones that travel
very closely together.
And there are a number in, say,
the left-hand side of this
ellipse or maybe 10 of them or
so that are kind of clumped
in that region, and they seem to
be in this group of triangles
that's labeled one, this
triangle that's labeled one,
which is one kind
of linkage block,
a block that
moves together.
So, those are probably among
those you might not need
to test all of them,
all these authors did.
But there are other places
within this ellipse that are
not in that linkage block,
and so you would want to
test those other
areas as well.
And sometimes these linkage
plots can tell you a lot
about what might be
the causative gene.
So in this plot looking at
inflammatory bowel disease,
in the middle you can see,
again, these association
statistics and you see there's
sort of a mountain of them
around the 10 to the 10th
to 10 to the 12th p-value,
minus log 10 p-value level
right over the X axis that
says 67,400,000.
And in this region there
are actually three genes.
You can see that there's --
sorry, there's this IL12RB2,
the IL23R and a
hypothetical protein.
And all three of these might be
possibilities as being related
to this disease.
But if one looks at the linkage
patterns, you can see that
these darker triangles now just
shown in black and gray here,
they're really only about two
blocks that are strongly
associated with the disease and
those pretty much narrow you
in to looking at this
interleukin 23 receptor,
so that's how those can sort
of help point the way to a
particular disease that might
be -- a particular gene that
might be causing
the disease.
Unique aspects of these studies,
they really allow examination
of inherited variability
at an unprecedented level
of resolution.
And they allow you to look at
the genome really without
having prior hypotheses.
Because we know so little about
how the genome functions,
in some ways it may be better
just to say let's set aside
all our previous notions and
just look and see what we find.
And it's amazing what
we have found.
For example, and as another
sort of positive to this,
once you measure the genome in
this way you can really relate
it to any trait that is
consistent with the
informed consent that's been
provided by participants.
So, interestingly, most of the
really strong associations that
have been replicated a lot in
these kinds of studies have not
been with genes that anyone
would have suspected of being
associated with the
disease in question.
So they weren't really on
anybody's list of things
that probably would be
associated and so they
would have missed in prior

studies where you had
to rely on a prior
hypothesis.
And some associations have been
in regions that weren't even
known to harbor genes and no
one is quite sure what that
means and that's an area of very
active research right now.
But as Hunter and Kraft point
out, the chief strength of this
approach is also its chief
problem, because when you
make more than 500,000
comparisons per study,
the potential for
false positives is
really unprecedented.
I'm a big Gary
Larson fan.
This is "God, Collings, I hate
to start a Monday with a case
like this," and the annual
Butlers of the World banquet
with a knife sticking out
of one of the butlers,
and God knows who all these,
you know, false positives
there are along with the
possible true positive.
And so something that's been
recognized for a long time in
genetic studies is that false
positives are really quite
possible, even before we had
genome-wide association studies.
And this sort of now classic
review by Joel Hirschhorn
pointed out the large number of
genetic associations that had
been reported with diseases
and you can see them climbing
really dramatically
after about 1994.
But in looking at the 600 or so
studies that he reviewed there,
really only six of the
associations were significant
in a consistent way in more than
three-quarters of the studies
he looked at.
And these are the six that are
shown, that are shown here.
So this is not a
very good record.
It was really something that was
quite concerning to people.
We did much of the same
thing in atherosclerosis,
but I won't go over
this due to time.
And this led to calls among
editors and journals and
publishers for replication that
probably the most important way
to be sure that an association
was real was to demonstrate
that it had been
replicated elsewhere.
There weren't really good
criteria for what constituted
replication, so there was a lot
of discussion about that.
Then we ended up having a
workshop here with our
colleagues at the Cancer
Institute to come up with
a series of criteria essentially
for what truly is replication
and what the criteria
for it should be.
We all, I think, agree that
replication is probably the
three most important
things in confirming
a genome-wide
association.
But it was important that the
initial study be described in
sufficient detail so that you
could even try to replicate it,
because you needed to know where
the cases and controls came from
so you could have similar
kinds of cases and controls.
You needed to know things about
participation rates and how
they were selected into the
study and how affected status
or case status was defined
and a number of other things
shown here.
And then in the replication
study you wanted to be sure
that a similar population,
if not exactly the same
population, had been used,
that the phenotype was very
similar so they weren't studying
height in one study and weight
in another but really using
much of the same phenotype,
and that they used the same
sort of inheritance model,

the same SNP, the same direction and that they were adequately powered to detect the possible effects, the sample size was large enough really to be able to detect effect if it truly was there.

Strategy for doing this was described by Bob Hoover, again, at the Cancer Institute, suggesting that one approach, and this has been taken by many of these studies, is to begin with, say, a reasonably large sample, 1,150 cases and 1,150 controls with a large number of tag SNPs, 500,000 or more, and then a replication study that might be even larger than that but that would only test a subset of those, maybe 5 percent of those that were associated.

And then a second replication study, again of large size that tested an even smaller number that replicated in multiple studies.

And then getting down, you know, sort of at the bottom of this funnel to even a smaller number and hopefully coming out at the end with maybe 25 to 50 loci, in this case, for prostate cancer.

And this is very much what was done in prostate cancer and led -- I think, it's only been about five or six loci for prostate cancer, but there have been other diseases in which more loci have been found.

And this is the approach that was used in breast cancer. Easton et al published this in 2007, and they used a much smaller initial set of cases and controls and a moderate number of SNPs, 267,000, but then a tenfold greater size for the replication sample that tested 13,000 SNPs.

Then 24,000 cases and 24,000 controls to test 30 SNPs and then sort of came -- ended up with six at the end of that study.

And this involved over 50,000 women with and without breast cancer, and these were all of the cohorts that were studied and able to -- enabled this finding to be come up with.

So these are really big, big collaborations; they're real challenges to put together.

You can also have problems with false negatives, so here -- "And now Edgar's gone… something's going on around here," even when the false negatives might be really pretty obvious.

And this was the prostate cancer study I referred to previously with 1,100 cases, 1,100 controls, then dropping down to -- then increasing, sorry, to 4,000 cases and 4,000 controls with their top 27,000 SNPs selected at this particular p-value.

And what was interesting about this when they tested the two stages together, there were four SNPs that were really very strongly associated from the p-value here.

This MSNB, the SNP MSNB associated seven times 10 to the -13th and so on, but when that was just looked at in stage one the initial rank was actually number 24,223, so its p-value was not very impressive at all, it was really way down in the ranking.

And similarly, even this second SNP that ended up at two times 10 to the -9th was only the 2,400th SNP or so with p-values that would have not have knocked anybody's socks off.

So, this is a challenge in being sure that your replication sample is large enough not only to pick up the false positives but not to miss any kind of false negatives.

It's been a real challenge trying to keep up with this literature.
The number of published reports has increased nearly exponentially.
There were 191 as of September, at the end of September of 2008.
And at the Genome Institute we're trying to keep track of these through what we call the catalog of genome-wide association studies, which is available on our Web site.
If you can't remember the URL, if you just Google "GWAS catalog," it should come up as the first hit.
And what we have tried to do here is to give a comprehensive listing of all of the published genome-wide association studies, including information on the author, the date, the journal, the trait that's being studied, the sample sizes, both initial and replication, the region of the genome, whether it's on chromosome 22 or chromosome 3, the gene that has been implicated, the strongest SNP in the risk allele that have been suggested to be associated, and the frequencies of those p-values, as you can see here from the catalog.
So a fair amount of effort to pull out all of this, and really the objectives were to identify and track all of these publications, extract key information about the associations, and make this widely available as a scientific resource for the community.
And it includes a downloadable data file, so if people want to get on and download this into an Excel file and use it for other research, they are welcome to do that.
We see commonalities across associations, genome-wide rather than disease by disease, and I'll show you some of the things that we can draw, you know, conclusions we can draw about these SNPs.
And we want to describe the approach clearly so that others can replicate or expand on it and we can maintain consistency in the approach.
And we pulled these out basically from published databases and various electronic clipping services that we have of news and, as I described, what kinds of information we pull off previously.
And we are looking here at about 180 published papers, excluding a few of them that didn't report the specific SNP.
There were 145 reports involving nearly 800 unique SNPs, and then there were about 3,800 that were perfectly linked to them so they also would carry some important information, so about 4,600 SNPs total.
Eighty-three of the SNPs in these reports had been reported two to seven times, some of them in association with traits that we wouldn't really have thought were necessarily related to each other.
And just giving some examples of those -- sorry, before that, functional classifications of these index SNPs, whether they were in regions of a gene or of the genome that might be coding for proteins, and if they code for proteins do they lead to a missense change, so a change in the structure of that protein.
There were only 37 of these 782, or only about 4 percent of those, that were in those particular regions even though those were the things that everybody sort of thought for sure are what are going

to be causative of disease.
There were 11, or
about 2 percent,
of them that were in the coding
region and made a change but
they really didn't change the
protein that was coded for.
340 that were intronic and then
a number, a smaller number,
in various other parts
that might be related to
regulation of gene
expression.
And then a good 350 of them,
more than 45 percent,
that were intergenic,
that really weren't
in any genes at all and,
again, are stimulating
a lot of research
as to why that is.
I'll skip over this
one, I think.
The odds ratios, or basically
the probability,
essentially the risk of having
disease in people who carry
one of these variants compared
to those who don't carry the
variants are typically
fairly small.
As you can see, most of them
tend to cluster around the
1.2 to 1.4 range.
And half of these associations
the median is 1.28, so half of
them are actually less than
an odds ratio of 1.28,
and half are more,
obviously.
And this is very similar to
what's been seen in Crohn's
disease and the same kinds of
distributions of variants
explained or odds
associated with disease,
roughly the
same idea.
And what's shown in this dotted
line is the power to detect
these risk loci, so probably
there are many more that
have even smaller odds ratios
but they're very difficult
to detect unless you have
massive sample sizes,
so that may be why they're
not being seen.
And there are some that have
very large odds ratios.
Those may be of some interest
and something that would be
worth looking into
in more detail.
I'm going to skip through these
because it's kind of a pretty
picture, but I'm just showing
you here what some of the
very high odds ratios,
strong odds ratios have
been associated with in
the allele frequency of
those associations.
And these are shown in a little
more detail here with these
various diseases that all have
odds ratios greater than about
4.5-fold, and those might be
genes that would be of great
importance on a public
health basis, but again,
need to be looked at
in much more detail.
We have also looked at
differences across
populations as to how different
the frequencies are in people
of, say, European ancestry
or Asian ancestry or
recent African ancestry.
And for the most part they're
really pretty similar.
And again, just focus on
the light blue here,
but the pink is pretty
similar as well.
And for the most part, you know,
more than half of these are
under a genetic distance,
which is a calculation
of how different they are in
populations, of less than .7,
but there are a few that have
much greater variability across
populations than that,
and those might be of
some interest
as well.
And in fact, in looking at them,
many of them are traits for
both -- traits related to
immunity and traits related
to pigmentation, which we know
are highly differentiated

across populations.
So just looking here at the
top 5 percent of FST values,
so those that are
.49 or greater,
which is a pretty extreme
difference among populations.
In the blue, those tend to
cluster among immune-related
traits, pigment traits,
obesity traits,
and then some neurological and
height findings and that.
And the top 1 percent,
so the really extreme ones,
real pretty much focused in
immunity and pigmentation,
which are, again, probably
things that are quite distinct
by geographic origin and allow
you to survive in the particular
environment that
you find yourself.
Some interesting findings that
have been in genes that were
not previously expected
to be related to disease:
I already mentioned the macular
degeneration finding and
compliment factor H.
Macular degeneration was thought
to be a degenerative disease
or maybe an ischemic disease
related to blood flow,
but no one really thought it
was related to inflammation
and yet this gene shows
up very, very strongly.
Some others in coronary disease,
asthma, type 2 diabetes really
weren't on anybody's
candidate gene list.
Gene deserts, areas were there
really aren't any genes at all,
have been very strong
associations of prostate
cancer with the tip of
chromosome 8 and there
don't seem to be any genes for
500,000 megabases or more.
So, what does that mean in terms
of causation of disease?
Crohn's disease similarly in
various areas without a lot
of genes.
And interestingly, some of
these associations have been
in common with diseases that
really weren't thought to be
related to each other.
So even though diabetes and
coronary disease can be risk
factors, diabetes particularly
a risk factor for coronary
disease, even when you
control for that,
there seems to be this
association with two
otherwise quite
different diseases.
And melanoma, I don't think
anybody would have expected
that to share a pathogenesis
with coronary disease
or diabetes.
Crohn's disease wasn't thought
to be all that related to
childhood asthma and yet
they share this association.
Is this real, is
it replicable?
It seems
to be.
What does it mean for
disease pathogenesis,
we don't know and that's
something that's an area
of active research.
And multiple cancers related to
this prostate cancer signal and
other signals in common in
multiple sclerosis and
type 1 diabetes, again,
perhaps pointing a way
to common -- sort of a common
etiology of these diseases.
Something that may leap out at
you is that Crohn's disease
shows up a
lot here.
In fact, one of the lessons I've
learned from this is if you want
to find genes for common
diseases, you should study
Crohn's disease because here
are all these more than
30 associations that have
been reported for this,
more than any
other disease.
So I think I'll wind up here
and just note that nearly half
of the SNPs that have been
identified in genome-wide

association studies as being
related to common diseases
are intergenic, so we don't know
what genes they're related to
and we need to
find that out.
Only about 8 percent of index
SNPs, or the SNPs that are
identified in these studies,
are in coding regions or
regulatory regions of
the genome, so, again,
needing to look at intergenic
and intronic SNPs.
We recognize there is some bias
in genotype SNPs for an excess
of missense variants,
that's one of the slides
I skipped over, but it's
essentially some bias
on the platform for what kinds
of SNPs they're looking for.
Most of the odds ratios
are really pretty small,
well less
than 1.5.
And risk allele frequencies
don't appear skewed either
toward rare alleles or toward
variants that vary a great deal
between populations, as
indicated by large FST values.
But the small number of SNPs
that do seem to be highly
differentiated across
populations seem to be
enriched for a trait
such as these.
And looking at loci at extremes
to these characteristics might
really teach us a lot about
things we don't know about
the genome.
So, I think I'll end with a
quote from Sir Tim Rice in
"Aida," "The more we find,
the more we see,
the more we come
to learn.
The more we explore,
the more we shall return."
And we certainly have a lot
to return to in the genome.
And, Greg, I think I'll stop
there and be happy to take
some questions.
[Dr. Greg Feero]
Great.
Thanks, Teri.
Dr. Manolio, this was a really
excellent presentation,
amazingly fascinating
results.
I would like to now open the
line for questions from
the audience.
Diane, I think
we're ready.
To reach the questions
you need to dial star 1.
[Diane]
Thank you.
We will now begin the
question/answer session.
If you would like to ask a
question, please press star 1.
Please unmute your phone and
record your name clearly
when prompted.
Your name is required to
introduce your question.
To withdraw your request
press star 2.
One moment, please, while we
wait for the first question.

[Diane]
[Unintelligible], your
line is now open.
[Male Speaker]
So, thank you for
a fascinating talk.
My question is, given all of the
association with Crohn's disease
and given the high frequency
of Crohn's disease in the
Ashkenazi Jewish population,
how are we assuring that
we're not actually seeing that
type of founder effect and
that we're really getting it
over diverse populations?
[Dr. Teri Manolio]
No, that's a
good question.
And many of these associations
were initially found in
Ashkenazi Jewish populations
but they have been extended
to populations that don't --
that are not of that descent
and we're seeing exactly
the same associations.

[Dr. Greg Feero]
While we're waiting, I actually have a question for you.
Given the large number of associations with Crohn's, it's a little curious to me, how frequently does ulcerative colitis show up on that?
I think clinicians think of those as sort of related, perhaps, disorders.
[Dr. Teri Manolio]
Sure.
Yeah, maybe about half of the loci that are seen in Crohn's disease are also seen in inflammatory -- well, in ulcerative colitis or inflammatory bowel disease in general.
And the reasons for that are not entirely clear because they are -- they can be difficult to distinguish both clinically and histopathologically but there are, you know, clearly some syndromic differences between them.
So it looks like about half of them are shared.
Now, whether that's a power issue that we just don't have enough cases to be able to detect them or not is not entirely clear.

[Dr. Greg Feero]
Diane, other questions from the audience?
[Diane]
A question came in from Sharon Jones.
Your line is now open.
[Sharon Jones]
Hi, I'm Sharon Jones with humangeneticsdisorders.com.
I wanted to know, in what ways can I incorporate this into genetics education awareness for the general public?
[Dr. Teri Manolio]
Yeah, I think it's reasonable at this point to say that this research is ongoing and it really has exploded in the past couple of years and this is what many geneticists are very, very excited about, that, you know, we've been looking and looking and looking with various tools and really hadn't found a lot that held up in lots of other studies, but this really has.
Unfortunately, at this point there's much more to be learned about this than there is to be taught about it, in that, you know, every answer we get raises 20 questions that we don't have good answers for yet.
So the fact that these associations are generally pretty darn small suggests that these aren't going to be useful really very soon for predicting disease.
They may be very useful in identifying treatments or pathways that might suggest approaches either to prevention or treatment.
But I think for the moment if we can convey the excitement of being able to find parts of the genome that everybody thought were silent and that really didn't do anything and we sort of arrogantly used to refer to junk DNA and that.
Well, these junk DNA areas are associated with disease and in a very, you know, sort of replicatable, duplicatable way, in ways that we don't understand, and, you know, it's a real challenge, I think, to all of us and a reason to get young people into science is to try to figure out these associations.
[Female Speaker]
[Unintelligible]
[Diane]
Next question comes from Becky McLane.
Your line is now open.
[Becky McLane]
Yes, thank you.
Do you have any ideas of why your associations are more

frequently found in
immuno-, pigment-,
and obesity-related
diseases?
[Dr. Teri Manolio]
Well, actually the ones that
I was showing you there were
the ones that differed
dramatically between
populations, so between
populations of recent
African ancestry versus
European ancestry or
Asian ancestry
populations.
And we suspect that -- we know
that pigmentation varies
dramatically by geography and
there seem to be, you know,
sort of plausible reasons
for why that would be,
and so that in a way kind of
reinforces the fact that,
yes, this makes sense.
The immune-related ones may be
a little bit more obscure but
probably -- as a matter of fact,
we do know that there are some
pathogens and bacteria and that
that only live in certain
climates or other, you know,
factors related to environment
or soil or plants or allergies
or whatever that are only
available in certain
climates.
And so when those climates or
geographic areas are acting
on a sub-population over
tens of thousands of years,
we evolve to sort of respond
to that, those environmental
stimuli, so that would probably
be why those are differentiated
as well.
The obesity ones I can't really
explain, or the neurology ones.
Those, again, are sort of
question marks we have
to pursue.

[Diane]
If you have any more
questions or comments,
again, please press star 1.
Again, please
press star 1.

[Dr. Greg Feero]
While we're waiting for their
questions to come in,
I'd like to draw your attention
to the slide that I failed to
put up at the beginning
of the webinar.
This is an additional e-mail
that you can use to reach
Laura Rodriguez regarding
data sharing policies for
genome-wide association
studies.
Diane, any further
questions coming in?
[Diane]
I have no questions
at this time.
[Dr. Greg Feero]
Fair enough.
Well, I would like to thank
all of you for participating
in this webinar.
We have enjoyed hearing
your questions.
Our next webinar will be held
in two months, on Thursday,
January the 8th at 1:00,
I think, Eastern time.
I think it will be a very
interesting topic,
the long and short of it,
"Finding Genes for Complex
Traits in the
Domestic Dog."
I have heard this talk before
and it is quite interesting.
So, I will leave you with the
fact that you'll be receiving
more information about this
upcoming webinar as the time
draws closer.
Again, thank you all
for attending.