The Human Genome Project

Ken H. Buetow*

First, I would like to thank the organizers for inviting me to participate in this conference. I always look forward to the opportunity to talk about the work that is being conducted through the human genome initiative, especially to audiences somewhat outside of our traditional area.

I will also ask your indulgence as I may rush through particular structures or terminology, so please feel free to raise questions or ask for clarifications, in the discussion period.

What I am going to talk about is one of the things in which we are very interested in the field of human genetics, which is a classic paradox that we face in human science right now. That is, we know a great deal about what causes disease or particular phenotypes, but we know almost nothing about which individuals are going to develop particular diseases.

To illustrate, let's focus on a very common behavior and relationship that we are all very familiar with — cigarette smoking and lung cancer.

We nearly definitively know that cigarette smoking is directly related to lung cancer with 80% to 90% of lung cancer in the United States related to cigarette smoking or exposure to cigarette smoke. However, the interesting paradox is that only one in 10 smokers develops lung cancer. So even though we know that smoking is important in the development of cancer, we also know that the majority of smokers actually do survive quite well without ever developing lung cancer symptoms.

In fact, this whole paradox is ubiquitous throughout the entire public health arena. You can't open a newspaper, read a magazine or listen to a local newscast without reading or hearing about the latest thing that's bad for you or that increases your health risk by 50% or 100% or 200%. But each of us sits back and says, "Well, O.K., but almost everybody I know does that and hardly any of them ever gets sick."

We can probably all name a person who drank three pints of whiskey and smoked four packs of cigarettes a day, caroused a lot, did all the horrible anti-health things that you could think of, and still lived to be 110 and died with his or her boots on, depending on how you view it.

What we actually do know of the human genetic evolution is that the key factor that differentiates who will develop disease and who does not is likely mediated both by exposure and by one's genetic constitution.

Unfortunately, it is not easy to just pop open somebody's chest, take a peek inside, and see what their genetic constitution is. We need to develop tools and have a strategy by which we can actually determine what this person has or what is different about that person. This is really the focus of this discussion today. How can we use the genetic tools that we are developing to allow us to figuratively (if not literally) crack open this person's chest, look inside the person and see why one person developed a disease while another person did not?

To understand the phenomenon and its difficulty, I think it is important and relevant to put into perspective what we are trying to do when we say that we want to find those genetic events that are important.

Cancer phenotypes that we are interested in are phenomena that occur at the cellular level. What we want to do is to be able to identify events at the gene level, or more precisely, at the nucleotide level (which means functionally). We need to traverse several orders of magnitude of resolution in order to be able to characterize these particular traits.

However, to find a gene that is involved with a particular disease or trait, it is similar to saying that we want to find a particular house in a specific town, given only the information that it is located somewhere on the planet earth—a somewhat daunting task!

What the human genome project is doing — and I believe the goal of much of the experimental work in agriculture — is actually developing a whole mapping strategy that will allow us to walk from the cellular level all the way down to the gene level and beyond, in a manner that is relatively straightforward, economical and efficient.

Initially, we describe things at the site or genetic or chromosomal level, but the initial localization of a genetic trait occurs through its establishment in a genetic linkage map.

Then, through a variety of increasingly higher resolution maps, we walk from this disease localization at a point in a genetic map, through physical maps that include a YAC contigs, cosmids and ultimately nucleotide sequence, to help us find what those specific genetic traits are.

The focus of my topic today really touches only a small percentage of the total effort associated with the human genome project, and will focus on this very top layer of genetic mapping. Then at a later time, hopefully, you will become more immersed in what these are. All I would like to do today is discuss one of the key early critical steps in this localization, because it puts us into the pathway that allows us to then move directly from a primary characterization to the identification of the specific genetic event that may be important in a disease or a trait.

I want to show you, essentially, where human gene mapping was about five years ago. In 1990, the state of human genetic mapping was quite primitive. While the human genetic map in 1990 was composed of about 250 to 300 loci, we

*K.H. Buetow, Fox Chase Cancer Center, 7701 Burholme, Philadelphia, PA 19111.

Reciprocal Meat Conference Proceedings, Volume 47, 1994.

have to put this into some sense of perspective. The human genome has somewhere between 50,000 and 100,000 genes, and it is composed of about 3 x 10^9 base pairs. So what we have are 300 "road signs" out of an organism that had several billion base pairs of information.

But those crude maps actually were serviceable maps that allowed us to do particularly interesting things in 1990. I don't want to suggest that genetics was invented then, but we could see that we had certain regions that were reasonably well characterized.

This is not unlike what the Scandinavians thought of the world in the 10th century. Certain areas of Europe were reasonably well described, in terms of both proportionality and boundaries. But there were huge vast regions that nobody knew about at all, or which were very poorly described. For instance, they knew that North America existed, but (at least from an American perspective) it was somewhat under-represented in its overall importance and magnitude. Likewise, Africa was poorly described, and Australia was still totally unknown.

This is representative of the state of the art of genetic mapping of humans about four years ago. As a consequence of this, the human genome initiative, as it was undertaken in 1991, recognized that if we were actually going to be able to undertake this large strategy, that first and foremost we must generate a complete fully-connected human genetic map of reasonably high density — a map that had markers every two to five "*centimorgans*."

A centimorgan (cM) is a sort of "mile marker" unit that is used to measure genomes, and the human genome is composed of 3,000 map units, or 3,000 cM. So the goal was to put a marker, or a "road sign," every two to five cM, which was felt to be a reasonable level at which we could begin that traversal to that higher level of physical mapping.

To undertake this effort, one of the centers that was funded was a group called the Cooperative Human Linkage Center (a group of which I am a member) founded at the University of Iowa and headed by Jeff Murray. The goal of this center of the National Center of Human Genome Research was to generate, maintain and distribute a high-resolution human genetic map composed of "user-friendly" genetic markers, high heterozygosity, and PCR-based genetic markers. The completion of the genetic map was constituted as our mission, as suggested in the goals of the genome project.

To do that, we undertook a reasonably novel structure as centers go, and formed a center without walls, the Cooperative Human Linkage Center. The Center's geographic distribution has performance sites at Harvard University in Boston, at Marshfield, Wisconsin, at Iowa City, Iowa, (the administrative home of the center), and Fox Chase Cancer Center in Philadelphia.

The goal of this Center was actually quite "heroic" at the time, that being to establish and map 4,200 highly polymorphic genetic markers. In addition to these 4,200 markers, we were going to generate 8,000 sequence tag sites; these are other non-polymorphic reference points that would punctuate the map. Onto this map of 4,200 markers, we were going to place 2,500 genes out of that total collection of 50,000 to 100,000. So we were going to generate a very high-resolution map, and punctuate it with a collection of genes that had been identified through other components of the human genome project.

One of the key traits of the markers that we were constructing was that we wanted them to be a new class of genetic polymorphisms, based on what is an increasingly ubiquitous laboratory procedure called PCR.

The PCR procedure starts with a very small amount of DNA, and from this simple information we can then amplify that small amount of DNA to huge quantities. The advantage of this technique is that it required us to have very little basic reagent. In other words, the DNA or the samples that we needed from individuals could be very small. The other advantage is that this is much less labor-intensive than many other approaches that were used.

To briefly describe the procedure, we start with a piece of target DNA. We then use reagents or "primers" that recognize the small piece of the target DNA. We melt the DNA, anneal the primers to the DNA, and then let the cell's machinery that normally replicates the DNA make an additional copy of that DNA. We repeat this time and time again, each reaction cycle doubling the amount of DNA. Within a few cycles — as few as 30 cycles — we come up with huge amounts of DNA from the original source.

The other approach we are taking is to use a new generation marker that is also PCR-based; these were called simple tandem repeat polymorphisms. These polymorphisms are based on an interesting observation made by one of my colleagues in the Center, that there are a high frequency of simple tandem repeats distributed throughout the entire human genome. This is a cytosine-adenosine (CA) repeat with the sequence motif — CA, CA, CA, CA, CA — repeated literally tens of thousands of times throughout the human genome. It is estimated that there may be as many as 30,000 copies of this simple repeat motif distributed throughout the human genome.

What we can do, then, is use our PCR methods to design pieces (or primers) of DNA that flank these repeats, and amplify this piece of DNA.

The most provocative observation at the time was that these repeat motifs actually vary dramatically in length between any two individuals. In one chromosome, the CA motif is repeated five times; in another chromosome, this CA motif is repeated 10 times. If we amplified across these and then sorted them, we actually got different sizes of fragments as a function of the total number of repeats present in a particular unit.

We can use this approach because humans, like most mammals, are diploid organisms. We can thus distinguish the individual allele composition for individual persons in a study. This is important in humans because we cannot construct crosses in humans to our desire, and we have to take advantage of the fact that polymorphism naturally exists. These polymorphisms result in a natural class of markers that are constitutionally useful in genetic studies.

Our group decided to use these simple motif repeats as the genetic basis of our markers, but also explore types of motif other than the CA phenomenon.

We looked at repeats other than the CA class that included tri-, tetra-, and pentanucleotide repeats. This approach has the advantage that when the polymerase in the PCR reaction amplifies the reagent, it is very faithful to the replication; it

Figure 1
Sex Averaged Skeletal Map of a Genome.

Marker	Map Interval (cM)	Map Position (cM)	Marker	Map Interval (cM)	Map Position (cM)
GATA4H04	12.6	0.0	D1S187	3.3	144.0
D1S228	19.5	12.6	D1S189	11.1	147.3
D1S199	7.6	32.1	D1S303	3.7	158.4
D1S234	2.6	39.8	SPTA1	3.3	162.0
D1S247	3.3	42.4	CRP	0.0	165.3
D1S233	1.2	45.8	CRP	6.9	165.3
D1S201	5.6	47.0	APOA2	0.0	172.2
D1S186	4.9	52.6	APOA2	4.5	172.2
D1S193	3.6	57.5	D1S104	2.6	176.7
D1S319	6.0	61.1	D1S194	3.5	179.2
D1S200	6.2	67.1	D1S318	8.7	182.8
D1S220	3.5	73.3	D1S210	5.1	191.5
D1S312	5.6	76.8	D1S215	2.7	196.5
D1S246	4.4	82.4	D1S399	1.4	199.3
D1S198	4.0	86.8	D1S240	0.7	200.6
D1S159	3.4	90.7	D1S191	1.1	201.3
D1S224	2.7	94.1	GATA7C01	0.0	202.4
GAAT1D9	8.3	96.8	GATA7C01	28.2	202.4
GATA109	2.7	105.1	D1S245	5.6	230.7
D1S207	12.5	107.8	D1S237	3.2	236.2
D1S188	0.0	120.3	D1S229	6.1	239.5
D1S188	3.1	120.3	D1S213	3.0	245.6
D1S236	6.1	123.4	D1S225	10.2	248.6
D1S223	4.3	129.5	ACTN2	13.5	258.8
D1S239	10.3	133.7	GATA4A09		272.4

doesn't get significantly confused when it has a triple or quadruple repeat, minimizing replication errors.

The downside of using these particular types of markers, though, is that they are not nearly as ubiquitous as the CA class. As I mentioned before, there are literally tens of thousands of the CA's. In our screening of these, the most frequently observed so far is the G-A-T-A class, represented in a couple of thousand copies in the genome, rather than in tens of thousands. Therefore, it has been much harder for us to find these.

We use a particular type of technique developed by Geoffrey Duyk, the investigator at Boston, who uses marker selection techniques to enrich these particular classes (I won't go into the details of this procedure, but I can provide further information for those interested). We generate the libraries of the markers from those enriched for these particular motifs.

We have now generated and mapped more than 500 of these markers in the human genome, a collection of tetra-, tri- and pentanucleotide repeats. The very highly polymorphic distribution of frequency of the markers differs between any two individuals due to heterozygosity. Looking at the probability that two individual ilials are different within an individual, the mean value of heterozygosity turns out to be about 70%. So if we took all the people in this audience, 7 out of 10 of you would have (for any specific marker we look at) differences between the two ilials, since you got one from your mother and one from your father.

We then took these markers, in conjunction with a collection of about 3,000 other markers we have, and constructed primary maps of the human genome. We use a very rigorous algorithm in order to reduce the amount of errors, because very early in our mapping efforts we discovered that even small amounts of gene typing errors or laboratory errors can result in very dramatic changes in the maps. These errors actually result in incorrect orders in the maps, resulting in maps that are grossly inflated. As you can see, we go through a great number of details to assure these maps are rigorously constructed. We start with a particular collection of points and actually build up the map one locus at a time, with data integrity checks built into the map at each point that we introduce a locus. These "checks" assure that we can statistically evaluate the map to provide as strong or as rigorous a point as we can possibly build.

We believe that by doing these steps, we can build maps that have virtually zero error associated with them in terms of laboratory error, and "myotic error" (error associated with violations of the assumptions of the map construction statistics).

Shown in Figure 1 is a map that is a product of this algo-

Figure 2
Sex Averaged Framework Map of a Genome.

Marker	*Map Interval (cM)*	*Map Position (cM)*
D1S77	15.5	0.0
D1S160	11.5	15.5
GATA4H04	7.7	34.7
D1S228	6.3	34.7
D1S170	12.1	41.1
D1S199	2.8	53.2
ALPL	0.0	56.0
ALPL	1.8	56.0
FUCA1	0.0	57.8
FUCA1	6.0	57.8
D1S234	3.0	63.7
D1S247	3.8	66.7
D1S233	1.4	70.5
D1S201	6.9	71.9
D1S186	3.7	78.8
MYCL1	0.9	82.4
D1S193	5.4	83.3
D1S168	5.6	88.7
D1S319	3.0	94.3
D1S161	6.8	97.3
D1S200	2.7	104.1
D1S21	3.9	106.8
D1S220	3.7	110.7
D1S312	4.5	114.3
D1S209	3.1	118.8
D1S246	4.4	121.9
D1S198	3.8	126.3
D1S159	3.4	130.0
D1S224	2.6	133.4
GAAT1D9	7.9	136.0
GATA109	3.2	143.9
D1S207	8.4	147.0
D1S167	5.5	155.4
D1S188	0.0	160.9
D1S188	3.7	160.9
D1S236	6.4	164.6
D1S223	5.0	170.9
D1S239	4.7	175.9
GSTM1	0.0	180.6
GSTM1	6.6	180.6
D1S73	4.9	187.2
D1S187	2.6	192.1
TSHB	0.0	194.7
TSHB	0.0	194.7
TSHB	2.3	194.7
D1S189	3.4	197.0
D1S36	4.7	200.4
GATA12A07	3.1	205.0
D1Z5	6.2	208.1
D1S67	4.0	214.3
D1S303	5.1	218.4
STPA1	3.6	223.5
CRP	0.0	227.1
CRP	3.2	227.1
ATP1A2	0.0	230.4
ATP1A2	3.5	230.4
APOA2	0.0	233.8
APOA2	4.3	233.8
D1S104	2.5	238.1
D1S194	3.4	240.6
D1S318	6.6	244.1
D1S210	2.2	250.6
AT3	0.0	252.9
AT3	4.8	252.9
D1S215	7.7	257.7
LAMB2	3.7	265.4
D1S399	1.4	269.0
D1S240	0.6	270.4
D1S191	0.5	271.0
D1S202	0.6	271.5
GATA7C01	0.0	272.1
GATA7C01	7.8	272.1
GATA10C02	7.9	279.9
D1S310	4.5	287.9
D1S52	3.2	292.4
REN	0.0	295.6
REN	1.9	295.6
D1S58	2.0	297.6
CR2	1.6	299.6
D1S54	5.6	301.2
D1S245	5.7	306.8
D1S237	3.7	312.5
D1S229	2.6	316.2
GATA4H09	4.1	318.8
D1S213	3.7	322.9
D1S225	0.0	326.6
D1S103	0.0	326.6
D1S103	3.5	326.6
D1S178	8.6	330.1
D1S179	3.3	338.7
ACTN2	3.0	342.1
D1S74	3.5	345.1
D1S8	4.8	348.6
D1S163	16.1	353.4
GATA4A09	5.0	369.5
D1S180	8.1	374.5
D1S102	0.0	382.6
D1S102	13.5	382.6
D1S68		396.2

Figure 3
Sex Averaged Scaffold Map Composed of CHLC Markers Ordered Using an Integrated Framework Map.

Marker	*Map Interval (cM)*	*Map Position (cM)*	*Marker*	*Map Interval (cM)*	*Map Position (cM)*
GATA4H04	12.2	0.0	GATA25B02	34.7	166.3
GATA27E01	10.4	12.2	ATA4E02	8.7	201.0
GAAT4D10	30.9	22.6	GATA31H02	1.6	209.7
GATA27F07	9.3	53.5	GATA7C01	0.0	211.2
GTAT1A7	16.9	62.8	GATA7C01	0.0	211.2
GATA26G09	12.4	79.7	GATA43D10	7.2	211.2
GGAA7C04	8.6	92.1	GATA10C02	6.1	218.5
GATA5G07	1.0	100.7	GATA46C02	4.5	224.6
GAAT1D9	7.7	101.7	REN	12.7	229.1
GATA109	3.2	109.4	GATA42A04	12.6	241.8
GATA6A05	10.3	112.6	GATA4H09	0.0	254.4
ATA2E04	13.0	123.0	GATA4H09	2.2	254.4
GATA25	20.3	136.0	GATA3D01	39.1	256.6
GCT1C9	3.5	156.3	GATA4A09	3.9	295.6
GATA12A07	6.5	159.8	ATA5E03		299.5

rithm. We call it a *skeletal map* because it is a map that is somewhat sparse, but absolutely rigorous and firm in its construction and statistical integrity. We have constructed for each map using that rigorous algorithm, with one map of each chromosome. We presently have 24 different chromosome maps with an average density of 6.5 cM's. They are not quite as dense as the projections of the human genome project required, but they are as firm in construction as can be estimated with a high degree of order.

We also have generated a series of maps that are somewhat more annotated than these skeletal maps called *framework maps* (Figure 2). These framework maps are built using the standard criteria that the human genetics community has established for the construction of maps. In other words, the first set has very rigorous requirements as to the loci, but can't inflate the map.

In framework maps, we actually "deconstrain" to the point that allows additional loci into the map. What we can see is that these maps then increase somewhat in density. The average density of this map now goes up to 4.5 cM, rather than 6.5 as seen in the skeletal map. We are very close to where we want to be now in the targeted closure, although "very close" is certainly argumentative. If we are at 4.5 and our desired resolution is 2.5, it just means we have to do twice as much mapping as we have done to date to get to full closure.

There are a couple interesting properties that I won't dwell on, but I want to note. Looking at the skeletal map, we see that there is a region at the top of the chromosome that is not very well mapped. Even when we deconstrain and allow more markers in, some of these regions still aren't very well mapped. We are not sure whether this represents an actual flaw in the distribution of the types of markers we're using, or whether they represent some interesting regions where the genetic map distance is actually quite large in proportion to the physical distance. Several laboratories (including ours) are conducting experiments to determine the true state of that anomaly.

To reiterate the difference between a skeletal map and a framework map, we use a diagnostic that measures map area by looking at how much inflation is associated with the introduction of a particular locus. So, if we introduce a locus, the map will actually get larger (or smaller when you remove it, depending on how you want to look at it). The skeletal map is virtually flat, which is what we want. It has less than 0.1% typing error associated with it, whereas if we look at a framework map built with the standard human genetics criteria, we have about a 0.5% typing error associated with it. So we believe the effort is worthwhile.

We have constructed a series of maps that we think will also be of great use to the human genetics community, called *scaffold maps* (Figure 3). These maps are actually subsets of the framework maps, and are composed solely of the tri-, tetra- and penta-nucleotide repeat markers which are very easy-to-use markers. These maps have an average density of 10 cM.

The advantage of these maps is that not only are they user friendly, but they actually permit a very straightforward and easy screen of the human genome at a coarse level. You will see later that this will be important, because we want to be able to have tools that we can use rapidly to find disease genes and similar characteristic traits of interest.

The other thing that we are doing is making an extensive use of electronic media for the publication and distribution of our work. I would encourage you to consider this approach to your work as well. We distribute our maps and our information long before publication via a variety of electronic sources — FTP, Gopher, and now through Worldwide Web.

We have established a system that allows people to add points to the maps and collaboratively map, even without our

participation, by using these electronic servers. One of the things I would like to emphasize is the introduction of a graphic user interface we maintain that allows literally "point and click" access by anyone to all of the genetic information (the mapping reagents, the primers, the map locations, the diagnostics). Anybody in this room who can log into Internet can access, read, take to your lab, and look at any and all of the information. It is all done literally through "clicking" any information points that are highlighted. You can pick up a map, any of the data, copies of our newsletter, etc.

As I noted previously, we have been mapping with approximately 3,000 markers. We have been somewhat challenged recently with the introduction of a collection of about 2,000 additional markers, which we hope will ultimately bring us to our closure goals, since we need to do about twice as much mapping as we are doing right now. This should bring us very close. However, when we make the jump from mapping 3,000 markers to nearly 5,000 markers, it amplifies the challenge somewhat and we have to devise new or alternative strategies for building maps, compared to what we had been doing previously. I would like to spend a moment focusing on this.

In this newest strategy, rather than *de novo* reconstructing the maps every time from a collection of 3,000 markers, we are starting with our original collection of markers, comprising the scaffold maps and generating what we call a "recombinant interval data base." All our maps are constructed using this same collection of families, called the CEPH Reference Family, which is distributed out of France. For our reference map, we identify only the subset of individuals that actually contain recombination events in the particular map that we are looking at.

The advantage to this approach is that if we are generating a 10 cM map (what 10 cM means is the frequency with which we see a recombination event in that interval), only one in 10 individuals actually has a recombination event. We can then reduce our genotyping, characterization and statistical analysis literally by 90% by using such a strategy. We then only focus on information on the recombinant intervals.

The key point is that we actually start with a fixed map, and find out where in the map a new locus is placed. Once we identify a new point in the map, it is placed in the appropriate order with the smaller intervals. Starting with a reference map, we determine where new markers would lie in this reference map with a minimal amount of typing typically using a 10 cM reference map. Then we simply order this (based on their combination events that we previously described in the parental origin of those events) until we find the minimum number of recombination events, (or the total number of recombination events) that were originally in the interval.

The other key advantage to this approach is that it allows us to use state-of-the-art computing techniques. We now break our map — instead of considering one continuous 3,000 cM map; we break it into 300 ten cM maps, and use distributed processing to dissect and analyze the map. To update the map, we distribute it to a whole variety of servers, who then redistribute it back and the map gets reassembled.

Thus, operations that used to take literally days to run now run in hours because we can simultaneously use multiple processors and make very efficient use of computing resources.

At this point, you may ask, "So you have these high density maps. What are you doing with them and why are they valuable? How can you go about applying them, and how can you make use of them?"

I want to reply to this with where we originally started our studies in human genetics, and what the main paradigm is in the application of genetic maps for the finding of disease genes or interesting traits.

Historically, and continuing today, the major approach that human geneticists have used in trying to find genetic traits was to identify interesting families that appear to be behaving within the parameters of simple Mendelian or statistical or genetic rules. We then trace the segregation of the trait through those families, and look for cosegregation of the trait with a marker in a particular genetic map.

Perhaps one of the more famous success stories of this (at least in terms of public health significance) is the localization of the human breast cancer gene. Over the last three to five years, there have been 50 genes that have been localized; and probably within the next five years, almost every gene that can be mapped by this process will likely be mapped, given the current maps that are available to us.

Basically, the technique is to take a family that appears to be transmitting a particular trait and then follow the transmission of that trait with the markers (or the individual loci) within the genetic maps. Therefore, within this family, everyone who ultimately develops breast cancer has inherited this ileal. What this technique does is provide us with a powerful tool in that it allows us to both characterize and identify the location of a disease gene within a genetic map so that we can ultimately go through the rest of that hierarchical process to identify and clone it. But it also provides us with an immediate diagnostic test to further characterize the epidemiology and the clinical nature of the disease itself.

For instance, in terms of diagnostics, once we establish that a specific ileal is the risk ileal, those women with the ileal are at virtually zero risk (or at least a much reduced risk of developing breast cancer) compared to their sisters, because they do not share the ileal. With this information, we could immediately reduce the anxiety relative to risk in this family, from concern that they had a 50/50% chance of developing breast cancer by the time they were 50, to saying that if you don't have the disease gene, you have essentially only the overall population risk of developing breast cancer.

In other instances, this approach permits us to look at women who actually carry the risk ileal, but who have not developed breast cancer, or conversely, look at the women in the same family who have developed what would be considered sporadic breast cancer, but do not carry the disease ileal, to assess why those women have developed cancer. To determine this, we statistically examine the cosegregation of the markers and the disease trait in a family to identify instances where there is statistically non-random transmission, or where the cosegregation of these events is non-random, based on statistical criteria.

To this point, I have used cancer as an example. This procedure has also been used to identify and localize a gene for Alzheimer's disease. It has been used to locate genes for congenital birth defects. It has been used to localize genes for a variety of different types of cancer. This technique is a very powerful and well-demonstrated technique at this point.

We could also use these genetic maps to look at events that are not necessarily simple genetic events, or events that do not involve genes being transmitted, but are events that are occurring somatically in an individual's tissue or constitution.

One area of discussion is how a tumor develops in an individual who has a genetic predisposition *vs.* an individual who doesn't. The current thinking is that tumors develop by loss of genetic material in individual cells, and that one has to lose both copies of a controlling or a suppressing locus in order to develop cancer.

We also know that the development of cancer requires us to lose material, not just at a single locus, but at multiple loci. Therefore, we need to characterize the genetic maps so we can look for these transitional events by comparing a tumor sample to the normal constitution, looking at the variant patterns in the normal tissue and the tumor tissue. By looking for the differences, it can tell us where the genes involved in cancer (or other somatic alterations) might occur. We can look at genetic markers on each of the individual chromosome arms, to see which chromosomal arms show high frequencies of alterations. This indicates to us that there is a cancer gene located on that particular arm.

We now have maps that allow us to actually identify simple traits that are "well-behaved" within families; but what can we do about the majority of cases? When we talk about the type of breast cancer or the type of most of the diseases we are looking at, they actually represent a very small proportion of the total burden of human disease that we are examining. For example, the BRAC1 locus, that will probably be cloned in the next three to four months, constitutes less than 2.5% of the total burden. So even though we will find the genetic basis of an individual type of breast cancer (and I don't want to belittle the significance of knowing some basic genetic etiology of breast cancer), we will still be left with having to determine what the other 97.5% of breast cancer is related to.

What I want to close with is how we might go about using these genetic maps and reagents to address traits that are more complicated than these simple traits, and to address the kind of traits that this audience would likely be interested in — complex traits, and traits that are the product of interactions between genes and environment that produce specific phenotypes.

Historically, there have been two approaches to finding human disease genes. The first is through the family studies I described previously. But the second (and more provocative) approach has been through candidate locus studies, where we look for association of a disease gene with a trait, rather than look for co-segregation of a disease gene and a particular marker trait.

What these studies assume is that any trait that has an underlying genetic basis should be consistent. Two people who have a given trait that is genetically determined should be more similar, genetically, than two people who do not share that trait. This is the fundamental premise of this approach. If we start with that premise, it is a very powerful tool, because we can now scan the genome and look for regions where people are more genetically similar than you would expect them to be by chance alone.

If we look at cases for a particular trait (versus controls) such as lung cancer and we assume that lung cancer has a genetic component associated with it, we can assemble people who have lung cancer and compare them to individuals who do not have the disease. When we look at the appropriate genetic location, we should see that the people with lung cancer are more similar than people who don't have lung cancer, using statistical tests to detect this.

Another important factor about using association studies is that we don't have to be perfect. We do not have to hit exactly the variant that a particular disease is associated with, but in fact only need to be reasonably close. There is a phenomenon called "linkage disequilibrium" that says that individual polymorphisms remain correlated over periods of time even if you're not looking exactly at the disease trait (or variance) itself. Markers (or variances that are very close to it) will show association. So we do not have to actually hit a home run on this; we only need to be within the ballpark to be able to identify these associations.

Given that we have a well-spaced, high-density genetic map, we can literally traverse the genome sequentially, looking for regions or locations within the genetic map where the variants are not randomly distributed between people who have a disease or trait, and people who do not.

To do this, we require a genetic map of about three cM's in density in order to have a reasonable chance of being able to be successful.

I would like to close with this thought: I think that what we are doing in humans is applicable and equally important to animal studies. I think that the next generation of genetic problems that we are going to address, both in humans and (I assume) in other experimental organisms, including those of agriculture importance, is the interface between environmental effects and the combination of genetic effects that ultimately result in the outcome of an important phenotype. In humans, it may be disease; but it may also be important traits for agriculture. With increasingly dense genetic maps, and the ability to characterize the environmental effects, we now can have a reasonable chance of identifying the aggregation or combination of genetic traits that will result in desirable phenotypes.

Editor's Note: This paper is based on a transcription of Dr. Buetow's presentation and was not edited or proofed by him. The work of Dr. Buetow and his colleagues in the Cooperative Human Linkage Center can be found on the Internet through the World-Wide Web at http://www.chlc.org/.