

The Scientist 14[15]:1, Jul. 24, 2000

NEWS

The Human Genome

Armed with a working draft, scientist, ponder the meaning of all 23 human chromosome pairs and more

By Arielle Emmett

Life sciences took center stage virtually around the world June 26. President **Bill Clinton**, flanked on the left by Celera Genomics Group president **J. Craig Venter** and on the right by National Human Genome Research Institute director **Francis S. Collins**, announced the completion of "the first survey of the entire human genome." Among others present for the announcement in the White House East Room were ambassadors from the United Kingdom, Japan, Germany, and France. British Prime Minister **Tony Blair** attended via satellite.

That "survey," of course, is the "working draft" of the human genome produced by the publicly funded international consortium Human Genome Project (HGP) and the "first assembly of the human genome" produced by privately funded Celera Genomics. "Without a doubt, this is the most important, most wondrous map ever produced by humankind," Clinton said of the human genome. He went on to praise **Francis Crick** and **James Watson**, who discovered the structure of DNA. "Dr. Watson, the way you announced your discovery in the journal *Nature* was one of the great understatements of all time. This structure has novel features, which are of considerable biological interest," Clinton said directly to Watson--with Clinton's own understatement producing laughter in the room. He was referring to the 1953 Watson and Crick paper "Molecular structure of nucleic acids--a structure for deoxyribose nucleic acid" (*Nature*, 171:737-738), which according to Philadelphia's Institute for Scientific Information's *Web of Science* has been cited in thousands of papers.

After the White House announcement, TV news shows and print media around the world proclaimed the completion of a major scientific milestone that promised great changes in the future of medicine and health. Indeed, it was a major scientific milestone. However, for life scientists, the milestone is more of a *beginning* than an *end*.

Now scientists must decipher the human genome's true pattern, functions, and meaning. The drive to decipher will not be limited to the mere chromosome unit, nor to individual genes. Increasingly, scientists will home in on the broader evolutionary patterns and interactions of large genes and nongenes--the "metadata" of the genome.

The more complex elements of interpretation may take decades or even centuries to resolve. But already, those intimately involved say it goes beyond identifying the interactions of a few thousand "disease genes"--normal genes whose mutation, deletion, or rearrangement spells the onset of birth defects, adult syndromes, or even subclinical health problems. The research focus will move rapidly toward the interactions of multiple genes, very large genes across multiple chromosomes, and global analysis of DNA activity--that is, the number of

RNA molecules transcribed from each gene in particular cell types, and the nature of the encoded protein products. Blocks of human DNA will be compared to those of other species to understand not only sequencing gaps but also genetic conservation--which parts of the genome are conserved during evolution as physiologically necessary for all life. Further, scientists will look at the interactions of environment, genetic makeup, and toxic exposures, including the ability of certain "beneficial" genes to detoxify the body and resist disease. Researchers will also examine the role of noncoding elements of the genome--introns--whose differing sizes across species, it turns out, may play a direct role in modulating gene expression levels, even shaping individual human differences of thought, morphology, and personality.

"There was clearly a time when the genome was so 'big' that the technologies concentrated individual attention on one chromosome or another, depending on the significance of its biology," explains **Aravinda Chakravarti**, James Jewell Professor of Genetics at Case Western Reserve University and a member of the Genome Institute Advisory Council. "For example, we studied the X chromosome for X-linked mental retardation or chromosome 21 for Down syndrome. We have focused on single gene-related illnesses, such as Pendred's syndrome on chromosome 7. But now people are moving from single-gene defects and beginning to study disorders with multiple genes involved, so the focus on chromosomes [as a unit] is no longer [as] meaningful."

Scientists will look at the structures of chromosomes as part of the whole genome mystery, Chakravarti asserts. "For example: How is the DNA in a chromosome organized the way it is?" he asks. "We know it's not random; we know there's a centromere and a telomere with defined sequences. We know the nature of repeats within each band, and the gene density by position [within] the chromosome. But if we study one chromosome, realizing it's one [structural] way that evolution has played out, there's nothing to say that another chromosome will follow exactly the same rules. Clearly there's much more to understand."

Analysis and sequence mapping of the genome has been a first necessary step in a longer road that will identify chromosome landmarks and the features of genes, especially large genes spread across hundreds, even thousands of kilobases. Up until now, scientists have identified relatively modest-size genes of about 30,000 base pairs--in most cases by positional cloning. Much more attention will be paid to "large genes split up into multiple exons, ... the structures on the chromosome [e.g., centromeres], and the nature and position of repeated sequences, which contain a marvelous history and record of evolution," Chakravarti declares.

Role of Serendipity

Already, the publicly available genome sequencing data has led to identification of "at least 13 or 14 disease genes," says **Eric Green**, chief of the National Institutes of Health Genome Technology Branch and director of the NIH Intramural Sequencing Center. Green's laboratory has been investigating chromosome 7 for over a decade. When the public Human Genome Project started in 1990, Green chose chromosome 7 "partly because of its intermediate size, partly because researchers had just identified the cystic fibrosis gene on chromosome 7. And the chromosome itself constituted five percent of the genome, which was a challenge." But Green admits it was a bit of serendipity--and the help of the detailed human chromosome 7 map and sequence that his group helped to generate--that led him to discover the abnormal base-pair configuration that gives rise to Pendred's syndrome, a congenital defect associated with deafness and goiter.

"Patients with Pendred's syndrome do not have cytogenetic abnormalities, but they do have small defects in a specific gene on chromosome 7, sometimes as small as a three-base-pair

deletion--and that's in a genome of three billion letters," Green says. By examining sequence maps and comparing the polymorphisms of patients affected by Pendred's with genetic maps of normal patients, Green's lab was eventually able to pin down the defective gene (the eighth one they tested on chromosome 7) in just over a year. "Before the super advances of the human genome sequence, this could have taken dozens of persons years to identify," he claims. "But because we had really good maps and sequence data, this was just one of the beautiful flowers that popped out [of the research]."

Complexities of Chromosome 22

Some of the astonishing new complexities of genetics are suggested by work on diminutive chromosome 22, the first to be decoded in draft form last December by researchers at the Sanger Centre in England, the University of Oklahoma, Washington University in St. Louis, and Keio University in Japan.

When the sequence was announced,¹ chromosome 22 was depicted as a gene-dense tangle. Unlike chromosome 21,^{2,3} which was considered genetically sparse and characterized by dark bands, chromosome 22 is "pale," meaning jam-packed with genes and repetitive base pair sequences near the centromere, variations of which are thought to be associated with at least 27 human disorders. Mutations here include deletion and translocation disorders; disease types include brain cancers and schizophrenia. Sequencing, which originally left 11 gaps, has revealed a total of 545 genes and 134 pseudogenes (an additional 300 or more genes are suspected) ranging in size from 1,000 to 583,000 bases of DNA. A total of 39 percent of the chromosome is copied into RNA, while only 3 percent of the chromosome encodes for protein. One of the sequence gaps has been closed since the December announcements. Scientists anticipate they will be able to close additional gaps using more statistical samples and new chemistries, enzymes, and libraries of clones targeting the missing pieces of DNA across the gaps.

Further, computer analyses reveal a total of 247 genes on chromosome 22 that are identical to previously identified human genes or protein sequences. Another 150 genes were discovered with DNA sequences similar to known genes, and 148 predicted genes contain sequences homologous to known genetic markers, expressed sequence tags (ESTs). Further, several gene families appear to have arisen by tandem duplication. "Families of genes ... are interspersed among other genes and distributed over large chromosome regions," reports a news summary from NIH.⁴ "There is unexpected long-range complexity of the chromosome with an elaborate array of repeat sequences near the centromere of the chromosome. The existence of so much repetitive DNA information could help explain how this chromosome rearranges or reshuffles its DNA, leading to human disorders such as DiGeorge's syndrome, which includes a form of mental retardation, and how chromosome structure changes over time."

Beverly Emanuel, chief of the human genetics and molecular biology branch of Children's Hospital of Philadelphia, says that positional cloning, the use of large clones, and other techniques have enabled her department to map deletion syndromes more precisely than before. "In 1980 there were two papers in short order that demonstrated chromosome 22 deletions. But they weren't the type studied now; these were deletions that were obvious because a whole part of the chromosome was missing. The affected patients only had 45 chromosomes; the rest of 22 was translocated onto another chromosome [11]."

Emanuel, however, began with the hypothesis that there must be "less obvious deletions" on chromosome 22 accounting for DiGeorge syndrome's multiple birth defect manifestations. By 1991 a decision was made to map chromosome 22 completely. "The technologies for doing so began to escalate," she says. Emanuel's lab, cooperating with the Sanger Centre and

Bruce Roe, a leading geneticist at the University of Oklahoma, was among the first to provide mapped clones to be sequenced. "We generated the first marker maps and yeast artificial chromosome (YAC) maps." YACs are large insert clones tricked into replicating the same base pair patterns of a piece of human genomic DNA, she explains. While figuring out the jigsaw puzzle of DNA fragments flanking chromosome 22 gaps, Emanuel's group published a paper arguing that "chromosome 22, as well as others ... has large blocks of duplicated sequence located at various positions [that apparently] permit illegitimate recombinations to take place and deletions to occur."⁵

Microdeletions--analyzed through molecular cytogenetic studies and sequenced genome maps--have shown up on the proximal part of chromosome 22 and chromosome 15. Emanuel has thus been able to rule out environmental or in utero factors as being primarily responsible for the chromosome 22 deletions that occur in children born with DiGeorge syndrome (the estimate now is that one in 3,000 to 4,000 babies has a deletion). In the July 1, 2000 issue of *Human Molecular Genetics*,⁶ Emanuel coauthors a paper describing how and why chromosome 22 and 11 translocation carriers (who are completely normal) can pass along an abnormal batch of 47 chromosomes to their offspring during meiosis producing a child with 47 chromosomes. Emanuel's terse conclusion: "Chromosome 22 is overinvolved genetically for its size." That overinvolvement may have to do with the willy-nilly replication of dense regions of almost identical sequences. This, coupled with the infidelity of the DNA replication machinery, permits illegitimate recombinations within the genome, causing the normal genomic organization to go awry.

Could these revelations be used to develop preimplantation genetic testing? Perhaps, Emanuel acknowledges. "But in many cases it might be quite difficult to know early enough or prevent the damage. However, knowledge of these disorders will lead to much better modalities of treatment, better diagnostics, and an approach toward prevention." In some cases, early diagnosis of the microdeletion disorders can help doctors and parents manage children with new therapies and ward off some of the later complications of the disease.

The Mysteries of C-G

"There is a significant fraction of 'junk DNA' that is self-replicating," according to **Elbert Branscomb**, director of the Joint Genome Institute at the Department of Energy (see sidebar). The purpose of these weirdly self-replicated parts of the genome, where deletions and changes often occur, have become a research agenda for scientists. Many are fascinated by the function of cytosine-guanine (C-G) base pairs that seem to crowd out some of the chromosomal telomeres. According to **Ian Dunham**, a senior research fellow at the Sanger Centre and head of the chromosome 22 project, the parts of the human genome that have unusually high concentrations of C-G pairs also have the most gaps--and are difficult to clone.

"These sequences are generally short but rich in C-G--and the enzymes for cloning don't go through them easily; you need to use specific tricks ... to get through those sequences. There's an extra level of polishing on the sequence you need to do to get an accurate determination." What that grinds down to is "taking the clone and cutting out specific DNA that's causing the problem and then creating a short insert library of the pieces," he says. "If you sequence it lots of times, it allows you to step through the difficult sequence [step by step], and that allows you to read it more accurately, to break the problem down into smaller pieces." Most, but not all, of the genome gaps have been resolved by subcloning or using new chemistries and enzymes. Others, including those on 22, especially gaps on the telomeres, remain resistant.

In Dunham's view, "I think chromosome 22 is particularly bad. When you look at 21, it's very

different in that 22 has a higher density of genes and higher levels of base pairs of C-G. Generally these self-replicating regions of C-G fall near the telomeres, where many of the deletions occur."

"It may be that the genome divides up where there are gaps," he speculates. There may even be genes "hidden" in the gap. "There's some evidence in *E. coli* that there

are runs of C-G dinucleotide that are not stable," Dunham observes. "Although there isn't any clear correlation between the types of rearrangements that occur in these gaps as far as we can tell, the gaps are associated with those repeats." What might this tell us about the seeming role of the repeat sequences? Dunham isn't sure, but evolution may play a role in a somewhat randomized process. "I think there are patterns to the genes; they reflect relatively recent evolution where you can see the genes having duplicated, and neighboring each other. There are regions of one to three megabases where the overall base content is different from others, usually different in the levels of C-G, and those areas seem to be especially rich in genes and in small genes. You get a different array of genes that may be doing different things. There's a random processing going on, tending to drive insertions. We don't know what it is."

Comparisons across mouse and other animal genomes may eventually yield some answers. But to Roe, the George Lynn Cross Research Professor of chemistry and biochemistry at the University of Oklahoma, self-replicating zones already suggest a certain logic--a logic that underpins the function of noncoding regions (so-called junk DNA), regions that may directly contribute to evolution of new forms of life and ultimately, genomic individuality.

According to Roe, chromosome 22 has divulged many intriguing secrets. Not only is it the site of flip-flops and translocations (Roe discovered the translocation of the q arm of chromosome 9 and chromosome 22, whose flip-flop results in the "Philadelphia chromosome" that causes the two major forms of leukemia), "it's a pain to sequence because it has a gaggle of repeated sequences. When DNA gets replicated," he continues, "the polymerasing enzyme skips railroad tracks, as it were; it skips that loop, and it gets that region deleted. You get DiGeorge syndrome, and also regions involved in brain tumors."

The study of noncoding regions on the chromosome may reveal something about descent from common ancestry. In Roe's view, a primordial genome, perhaps 5 percent of the size of the human genome today, began rearranging and recombining "a bazillion times." Accidents, mutations, exposure to radiation, comets, and other phenomena resulted in new genomes and new life forms. The genome took on characteristic structures for protein coding and noncoding regions. "If you look at a given gene in humans and corn, the coding regions are 40 percent identical," he observes. "But the difference is that the noncoding intron sizes in animals such as humans, mice, and monkeys are large compared to what they are in corn. Yet the genomes are roughly the same size. When you look at the corn genome, that each gene is smaller, it's because it has smaller introns," he says. Corn may have only 10 percent of its genome coding for something; the rest is noncoding regions between the genes. A lily, by contrast, is 10 times the size of the human genome because "it has a lot of stuff in noncoding regions."

There's more. "Let's say that, forget the X and Y sex chromosome stuff, you and I are 99.8 percent identical from a genomic point of view. Our chromosomes are 99.8 percent identical; about 4 to 5 percent are exons, coding regions of those genes. In that instance, you and I are 99.999 identical, and your hemoglobin and mine are identical. The protein that makes melanin is 99.9999 percent identical between you and me, except that part of my heritage comes from southern Italy, so the difference in my skin color has to do with the levels of gene expression--I express more melanin than you do. The genes that contribute to our nose shape and size [are also a product of what gets expressed more or less]--say, some noses

are longer and skinnier," Roe says.

"So what our differences are, what gives us our individuality, is not the coding regions, but the noncoding regions, which include the 45 percent of our genome that's introns. ...[I]t turns out that the difference between you and me is one in 500 to 1,000 bases in the noncoding regions." This includes introns and intergenic components, he says.

Roe's conclusion: "What this means in real life is that these noncoding regions are modulating or regulating the gene expression of coding regions. We knew that before. We thought it was promoters doing the work. But it now turns out that it's the size of the introns that matters." This could have evolutionary significance--intron size varies from species to species--although the meanings are still obscure. "Our differences lie in the noncoding regions; those are the differences that give us individuality," he claims.

Roe adds, "There are [loads of] single nucleotide polymorphisms (SNPs) [in the genome], and we are getting ready to submit a paper from all our groups together describing 11,000 SNPs in human chromosome 22. The overwhelming majority occur in noncoding regions; the number of SNPs causes a different codon to be used in an exon. So by sequencing the whole chromosome, we can now ask questions about what makes individuals different on a whole-chromosome scale." Roe suspects the role of introns in modulating genes may say something about the role of genetic accidents, mutations, and serendipity in ushering evolution to new, more survivable life forms. His chief conclusion, though: Humans are more alike than they are different. "We are 99.8 percent alike. Long ago, when the preachers began their sermons with 'Brother and Sister,' no one knew how right they really were."

While Roe's assertions have yet to be substantiated definitively, they offer an intriguing picture of the possibilities of genomic science. "The medical study of the genome is just the tip of the iceberg," he says. "We only know about a third of the genes on chromosomes. Now we've got to figure out what additional genes are there, what proteins they're making, and how they're being expressed in which tissues. The scary thing is that [scientific] people come to presentations and meetings and make claims [that cannot be demonstrated at this point]. We get more questions than answers, but that's the exciting thing." S

Arielle Emmett is a contributing editor for The Scientist.

References

1. I. Dunham et al., "The DNA sequence of human chromosome 22," *Nature*, 402:489-95, Dec. 2, 1999.
2. M. Hattori et al., "The DNA sequence of chromosome 21," *Nature*, 405:311-9, May 18, 2000.
3. R. Lewis, "[Chromosome 21 reveals sparse gene content](#)," *The Scientist*, 14[2]:1, June 12, 2000.
4. www.ornl.gov/hgmis/project/chr22.html
5. T.H. Shaikh et al., "Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis," *Human Molecular Genetics*, 9:489-501, March 1, 2000.
6. H. Kurahashi et al., "Regions of genomic instability on 22q11 and 11q23 as the etiology for

the recurrent constitutional t(11;22)," *Human Molecular Genetics*, 9:1665-70, July 1, 2000.

|