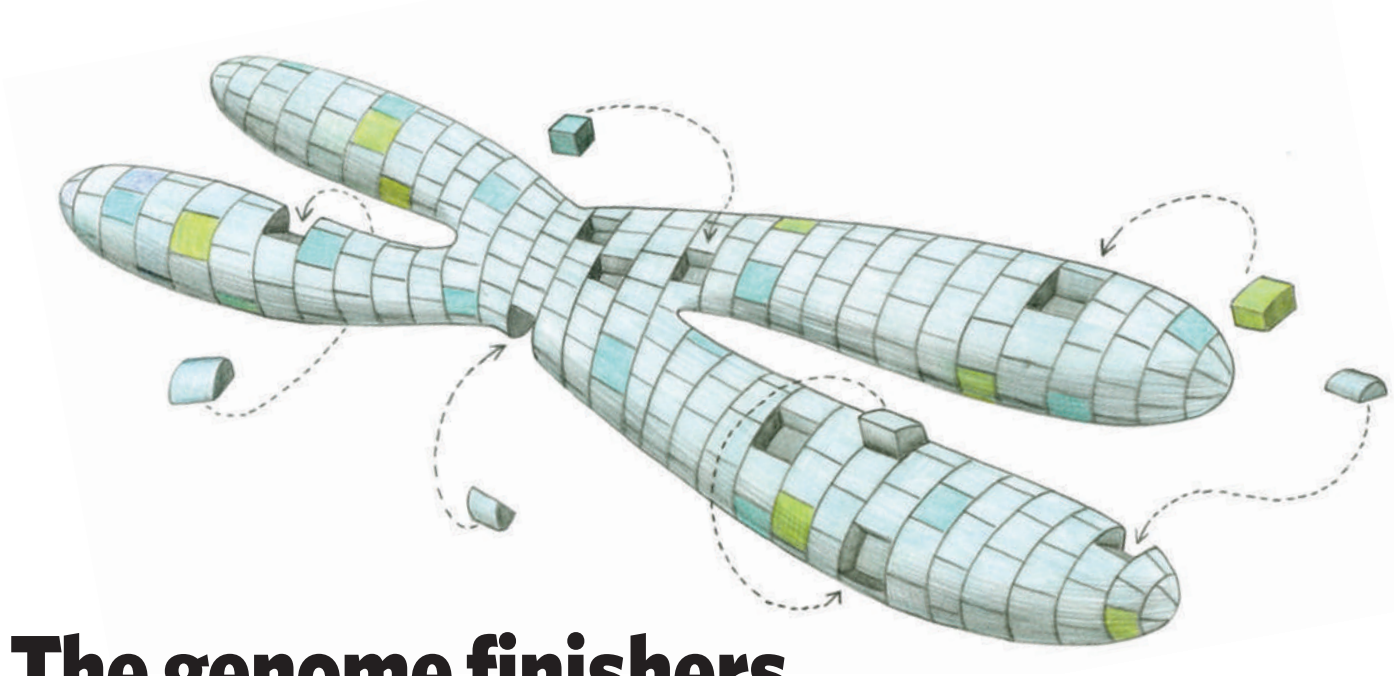


The genome finishers

Dedicated scientists are working hard to close the gaps, fix the errors and finally complete the human genome sequence. **Elie Dolgin** looks at how close they are.

From her windowless fifth-floor office at the US National Institutes of Health in Bethesda, Maryland, Deanna Church has few distractions from the job that lies before her. On her computer sit 888 open 'tickets', or outstanding problems with the human genome sequence. Although that number fluctuates, it's a not-so-subtle reminder that she and her team at the National Center for Biotechnology Information (NCBI) have a long way to go to finish the job started nearly two decades ago by the Human Genome Project.

This is the same project that an international team of scientists spent close to US$3 billion on to complete. In 2000, the scientists announced, to much fanfare at a White House ceremony, that they had finished the draft sequence of the human genome. They waxed poetic about opening 'evolution's lab notebook' when they published the draft the next year[1]. And they uncorked champagne bottles again in 2003 when the sequence was officially deemed finished[2]. By then, media outlets were reporting the developments with a twinge of fatigue. "This time it is the real thing, scientists promise," *New Scientist* reported. Another year passed before the final analyses were published[3], and two more went by before the paper detailing the last, fully polished chromosome came out in 2006 (ref. 4).

Still, three years later, Church is hunched over her computer, clicking away at her mouse, quietly clearing up the lingering troubles with the iconic sequence. Some of her tickets, submitted by her collaborators and users from around the world, are reports of missing bits. Others describe stretches in which someone thinks the sequence is mistaken. Still others are unique and unexpected challenges, such as complex DNA rearrangements, that could take years to sort out.

"It's a frustration," says Richard Gibbs, director of the Human Genome Sequencing Center at Baylor College of Medicine in Houston, Texas. "It's an extremely high-quality genome. It's the best there is, period. The problem is that a very small percentage of uncertainties still translates into a significant number of problems."

Church and her colleagues are working to build a solid, accurate reference, but their efforts have revealed how slippery that concept can be. The sequence, for instance, does not represent any one person's genome. It is an amalgam of DNA from different people, both male and female. It was put together this way to maintain anonymity for those who contributed the DNA and to ensure that the sequence represented all humanity — "our shared inheritance", as then-head of the project, Francis Collins, said.

"The work's not sexy. But it's important." — Deanna Church

But that shared inheritance is hard to capture. The genomes of two individuals look less alike than many had originally assumed. Rather than following a linear path of 3 billion base pairs with a letter changed now and then along the way, human genomes detour into hundreds of vastly different stretches in which, depending on the individual, millions of base pairs can be deleted, inserted, repeated or inverted.

A finished reference genome — if attainable — will therefore look very different from the project's first renditions. That's where Church and her team of finishers come in. They are striving to smooth out the differences and to develop a more dynamic platform that can capture much of humanity's commonalities and uniqueness. Some say it's a wasted effort now that individual human genomes can be sequenced at a fraction of what it cost ten years ago, but most say the reference is invaluable as a bedrock to support the sequencing of future human genomes.

Resolving the problems in the sequence will not win Church many accolades. She won't meet the president or land any papers in high-impact journals as those who "finished" the genome before her did. And once she puts a ticket to rest, there's always another one waiting. "It's not sexy," she says. "But it's important."

A coalition of the responsible

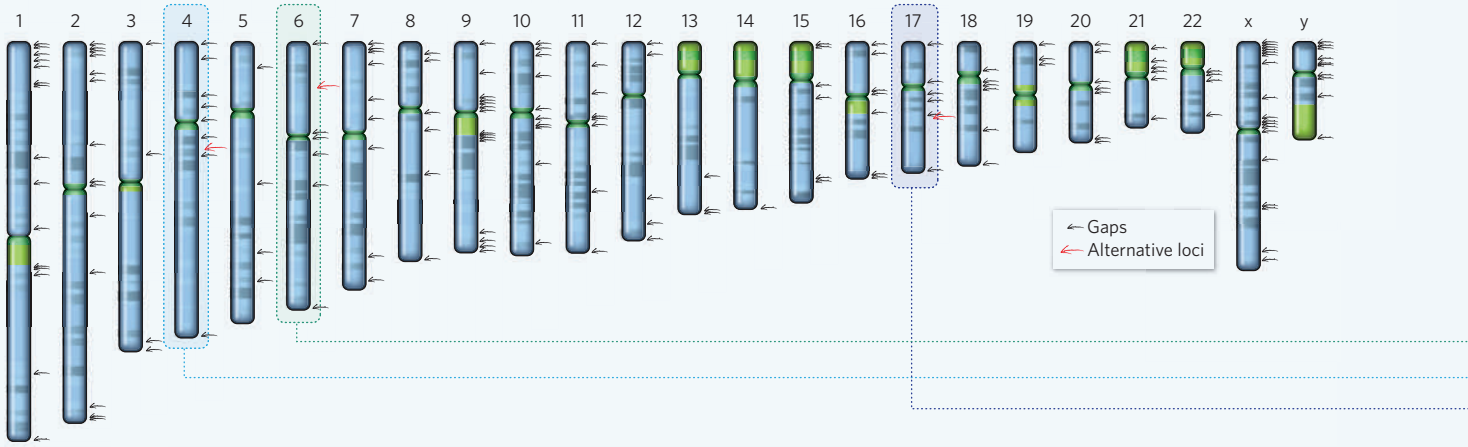
By April 2003, the sequencing had surpassed the international project's technical definition of completion — the sequence contained fewer than 1 error per 10,000 nucleotides and covered 95% of the gene-containing parts of the genome. But there were still errors — around 350 gaps in the sequence — and much of the structural variation was not included.

In 2004, Church and a few dozen researchers met to discuss genomics and structural variation at the Wellcome Trust Sanger Institute in Hinxton, UK. One complaint was echoed repeatedly: there was no easy way to fix or update the genome with new data. In the 1990s, when sequencing was in full

ILLUSTRATIONS BY CHRISTINE BERRIE

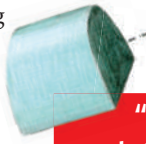
THE REMAINING GAPS

In March 2009, the Genome Reference Consortium released its first assembly of the human genome that had originally been completed by the Human Genome Project. The sequence closed 25 gaps and added alternative versions at three complex regions (red arrows). The sequence still contains nearly 300 gaps (general locations indicated by black arrows) that range from 700 to 30 million base pairs long, not including parts within the telomeres and centromeres (green) that are intractable to sequencing.



swing, researchers could contact the specific chromosome curators at each of the major sequencing centres involved with the project to report any sequencing slip-ups. But by 2004, few of the centres involved were actively monitoring their slices of the genome and there was little scientific impetus to revisit old work. This posed a problem. "Someone needs to have responsibility for the genome so that if errors are found, improvements can be made," says Adam Felsenfeld, a programme director at the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland.

Together with Ewan Birney of the European Bioinformatics Institute (EBI) in Hinxton, Church appealed to the NHGRI and the Wellcome Trust for funding. It took more than two years of meetings and deliberations, but eventually the NHGRI agreed to set aside up to US$1 million in operating funds from an annual large-scale sequencing award (more than $30 million per year) to Washington University in St Louis, Missouri. Sanger and the Wellcome Trust agreed to a similar amount, and the EBI and NCBI handle the informatics as part of their normal operations. The collaboration, known as the Genome Reference Consortium (GRC), is now the epicentre for genome improvement.



"Someone needs to have responsibility for the genome."
—Adam Felsenfeld

To improve the reference, the GRC is concentrating on three main goals: to correct assembly errors; to fill in the genome's remaining gaps; and to produce alternative sequences for regions of the genome with extensive variability.

Researchers have had the first two objectives on their agendas since the Human Genome Project was concluded, and they have been chipping away at them ever since. Some of the regions have been particularly difficult to polish off. For some repetitive stretches, for example, researchers struggled to make multiple copies in bacteria — a necessary part of the sequencing process. But newer methods are now allowing them to fill in these pieces of the genome. Earlier this year, a team led by Chad Nusbaum, co-director of the Genome Sequencing and Analysis programme at the Broad Institute in Cambridge, Massachusetts, used a next-generation sequencing technology that does not require bacteria to amplify the DNA[5]. Nusbaum's team then handed the sequences over to the GRC, which incorporated them into the reference assembly.

The third goal reflects something that has only recently come to light. At first, researchers assumed that genetic variation between people largely consisted of differences in single DNA letters. Now, however, they better understand the extent of structural variations — including deletions, insertions, duplications and inversions. Although some of these variants are involved in heritable disease, they are much more difficult to keep track of than single-letter differences because they often don't map easily to the reference. So instead of representing the genome as a single path of three billion letters, the GRC is introducing alternative paths to reflect its diversity.

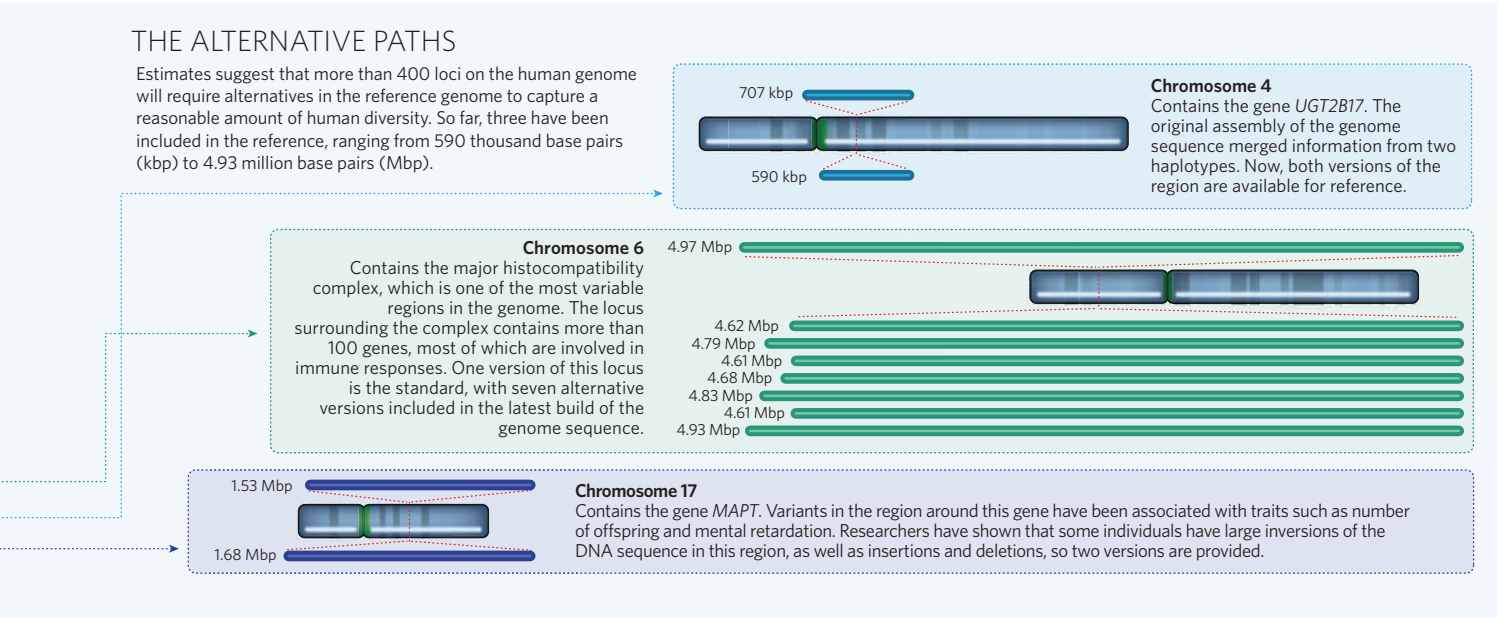
Twists and turns

One such region is the major histocompatibility complex (MHC) — a 4-million-base-pair stretch of chromosome 6 that contains many immunity-related genes and is recognized as one of the most variable slices of the human genome. The original reference sequence was a hotchpotch of multiple blocks of DNA, called haplotypes, taken from several donors, so the sequence that resulted didn't actually exist in any real person. To create a reference with clearer origins, a team led by Stephan Beck of the University College London Cancer Institute sequenced a single MHC haplotype. They then compared it against seven other common European haplotypes and discovered more than 37,000 single DNA letter differences and around 7,000 structural variations — a level of genetic diversity about an order of magnitude greater than the genomic average[6]. Beck's team's reference has now been swapped into the GRC's default sequence and the other seven haplotypes are included as alternative pathways.

Two other regions also have substitute haplotypes. One sits on chromosome 4 around the gene encoding the UGT2B17 enzyme, which metabolizes steroid hormones and many drugs. The 'finished' reference had misassembled two haplotypes, introducing a false gap. A corrected assembly found that the 'gap' was actually a deletion found only in some people, flanked by large duplications. That section is now included as an alternative pathway in the GRC reference.

The other region, on chromosome 17, encompasses the *MAPT* gene, and provides a case study of the limits of the original reference sequence, which consisted of only one haplotype. An alternative haplotype, a complex inversion of the first that is found in 20% of Europeans, was shown in 2005 to correlate with larger family sizes, suggesting that this second haplotype was under some form of positive selection[7]. But in 2006, Evan Eichler[8], a geneticist at the University of Washington in Seattle, and two other groups[9,10] showed that the inverted region was also prone to frequent spontaneous deletions that led to mental retardation. The inverted haplotype seemed to be both adaptive and the source of debilitating deletions. "You have a yin for the yang here determined by genomic structure," Eichler says. "The question was, 'What's going on?'"

To answer that question, Eichler needed the sequences. He teamed up with Michael Zody, chief technologist of the Genome Biology

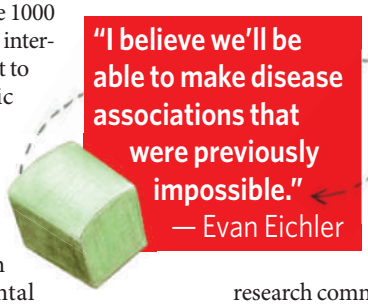


Program at the Broad Institute, to resequence the whole region and showed that the architecture of the inverted haplotype predisposed the sequence to undergo deletions associated with mental retardation[11]. By the time Eichler and Zody published their results, in 2008, the GRC was already in full swing and gearing up to release the next build of the genome. The researchers handed over their sequences to the consortium, and both haplotypes were included in the reference sequence. "The GRC provided a central clearing house for us to go through," Zody says.

Given the clinical relevance of these and other complex regions, providing multiple references is essential to detect the mutations underlying many diseases, says Eichler. "Once we get the alternative structures worked out, I believe we'll be able to make disease associations that were previously impossible," he says. Eichler estimates that around 5% of the genome — corresponding to around 400 specific locations — will need alternative sequences to provide a platform that adequately captures the spectrum of human diversity. These regions include more than 1,000 genes that affect a wide range of physiological processes, including immune responses, drug detoxification and reproductive ability, he says.

A common task

The GRC's first public offering — a more accurate version of the human genome — rolled out online in March 2009, updated the three parts of the genome with the alternative assemblies, corrected more than 150 alignment problems and closed 25 sequencing gaps. But that still leaves more than 300 gaps. In September, 20 of the GRC's core members gathered in Hinxton for the group's twice-yearly meeting to discuss what steps to take next. While lab workers were clacking away on their keyboards, bioinformaticians were working through one of the consortium's most contentious issues — how to change the reference genome to display only the 'common' gene variants. The GRC's nine-member scientific advisory board, which includes Eichler and Gibbs, has recommended that, wherever possible, the genome should include the common versions of the DNA sequence. But it hasn't defined what 'common' means. Should it be the highest-frequency variant or something that is shared by a reasonable proportion of the population? Should it be calculated across the world's six billion-plus residents or just in a particular ethnic or geographic group? Results emerging from the 1000 Genomes Project, a major international sequencing effort to catalogue human genetic variation in around 2,000 people from across four continents, should inform their decisions.



Some of the GRC members disagree with making such fundamental changes to the reference sequence. "I don't think we should be going through and flipping single bases throughout the genome," says Paul Flicek, who leads the vertebrate genomics group at the EBI. "Informatically, it just doesn't matter. As long as it works, I think it's okay."

Others outside the GRC question whether the entire project is justified. Why bother tinkering with a decade-old reference, asks Lincoln Stein, a bioinformatician at the Ontario Institute for Cancer Research in Toronto, Canada. He calls the effort "more of an abstract exercise than one that's going to have a practical impact". Church, for her part, waves off such criticisms as being from those preoccupied with large-scale genomics. As a detail-oriented person, she knows that the little things count. Individual investigators love their pet genes. That's one reason her queue of tickets is always full. And as genomics increasingly moves to the forefront of personalized medicine, many regions of clinical utility might slip through the cracks. For researchers interested in a particular disease-relevant locus "it doesn't matter that the genome may be 99% complete", she says. "If they're [working with] a region that's incomplete and wrong, they're screwed."

And so the GRC continues with its quiet quest, crossing Ts and sometimes changing them into As, Cs or Gs. Until a reference is no longer needed to assemble the DNA coming from current sequencing technologies, it will continue to document the evolving understanding of human variability in the reference genome. The GRC has also taken on the mouse sequence and will take responsibility for the zebrafish sequence in 2010. Although it may not capture headlines, most in the research community recognize its worth. "The GRC has exactly the right idea," says Jonathan Sebat, a geneticist who studies structural variation at Cold Spring Harbor Laboratory in New York. "It's a no-brainer that someone would be needed to clean up the mess." ■

Elie Dolgin is assistant news editor for *Nature Medicine* in New York.

1. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
2. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. *Nature* **422,** 835–847 (2003).
3. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
4. Gregory, S. G. *et al. Nature* **441,** 315–321 (2006).
5. Garber, M. *et al. Genome Biol.* **10,** R60 (2009).
6. Horton, R. *et al. Immunogenetics* **60,** 1–18 (2008).
7. Stefannson, H. *et al. Nature Genet.* **37,** 129–137 (2005).
8. Sharp, A. J. *et al. Nature Genet.* **38,** 1038–1042 (2006).
9. Shaw-Smith, C. *et al. Nature Genet.* **38,** 1032–1037 (2006).
10. Koolen, D. A. *et al. Nature Genet.* **38,** 999–1001 (2006).
11. Zody, M. C. *et al. Nature Genet.* **40,** 1076–1083 (2008).

See Editorial, page 825.

Families of young girls with Rett syndrome want barriers lifted to speed research on the disease.

colleagues said that they would have liked to but couldn't because of the terms of Novartis's licence on EGFP, which it obtained from GE Healthcare.

Novartis and GE have been unable to negotiate a way to share the mice, says Jeff Lockwood, spokesman for the Novartis Institutes for Biomedical Research — even though Novartis has ended its research project on the mice.

When Monica Coenraads, executive director of the Rett Syndrome Research Trust in Trumbull, Connecticut, tried to broker an agreement to share the mice, GE and Novartis asked the US National Institutes of Health (NIH) in Bethesda, Maryland, to distribute the mice through its Mutant Mouse Regional Resource Centers. But Lili Portilla, senior adviser for technology transfer at the NIH National Center for Research Resources, which funds the resource centre, says that GE placed such burdensome terms on the sharing that the NIH eventually gave up. For instance, researchers would not have been allowed to share the results of their research with the NIH, says Portilla.

GE spokesman Conor McKechnie blames the "third parties" from which GE gained the rights to the EGFP protein for the onerous licensing requirements. But David Einhorn, house counsel at the Jackson Laboratory in Bar Harbor, Maine, which distributes mice to researchers around the world, questions GE's contention. He points out that many other mouse models that incorporate the gene for EGFP have been made and shared without objection from GE or from the institutions that originally discovered and licensed the EGFP patents.

Researchers have had trouble sharing resources for decades, but the situation seems to be getting worse. A 2007 study, for instance, found that 18% of academics' requests for research materials from other academic labs were not fulfilled (see 'Limited access') — almost twice as many as found in a survey taken during the 1990s. For materials requested from industry, the 2007 study found, one-third of academics' requests were declined (J. P. Walsh, W. M. Cohen and C. Cho *Res. Pol.* **36,** 1184–1203; 2007).

Companies that don't want to share their resources don't usually publish papers describing them, says lawyer Tania Bubela of the University of Alberta School of Public Health in Edmonton, Canada. A publication changes the picture, she says. "The obligation of publication is to make your data and reagents available, so that people can replicate the results."

With no sign of a resolution, other labs have resorted to remaking the mouse model. Adrian Bird, director of the University of Edinburgh's Wellcome Trust Centre for Cell Biology, UK, says that his lab has re-engineered the mice and will distribute them through a repository, such as the Jackson Laboratory, as soon as his colony is large enough.

Bird and others say that it is unfortunate that scientists have had to delay research on the syndrome and spend money to regenerate a model that could already be in use.

"If you were to ask the families of people affected by this disease, they would say that every minute counts," says Bird. ■

➲ NATURE.COM
For a Commentary on the impact of licensing rules see:
go.nature.com/sgcjih

CORRECTION

The News Feature 'The Genome Finishers' (*Nature* **462,** 843–845; 2009) incorrectly states that a gap in the reference sequence of chromosome 4 was a deletion flanked by large gene duplications. The gap was an assembly error caused by attempting (and failing) to merge two alternative versions of gene sequence, which then erroneously appeared in the reference as large duplications in the UGT2B17 region.