Molecular Genetics & Genomic Medicine

Open Access

INVITED COMMENTARY

The evolution of comparative genomics

James C. Mullikin

Comparative Genomics Analysis Unit, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892

doi: 10.1002/mgg3.112

Introduction

The field of comparative genomics arose hand-in-hand with the ability to generate genomic sequence data. The first computer algorithms to compare amino acid sequences were developed over forty years ago (Fitch 1966; Needleman and Wunsch 1970)and improved upon as nucleic acid sequencing advanced (Sanger et al. 1977) with the application of improved statistical methods to the growing database of DNA sequence (Smith and Waterman 1981). This trend of exponentially increasing volumes of protein and DNA sequences has inspired a variety of algorithmic methods for DNA sequence comparison depending on the goal of a given investigation. BLAST (Altschul et al. 1990) is probably the best known of the alignment tools used today, but many others have been developed for specific comparative genomics studies, a few of which I will expand upon below.

Interspecies Comparative Genomics

As the human genome sequencing projects raced toward high-quality draft assemblies (Lander et al. 2001; Venter et al. 2001), the mouse genome sequencing project (Mouse Genome Sequencing Consortium 2002) was in high gear as well, because it was already understood that the power of comparing the genomes of these two species would be immensely informative for both understanding the human genome and for understanding the genome of one of the most studied laboratory animal species. One of the big mysteries of the human genome was: if the gene coding regions only make up about 1.5% of the human genomic DNA sequence and 50% is repetitive sequence, how much of remainder is functionally important as defined by excess sequence similarity between these two species? The answer required accurate alignment of the two genomes, and existing software algorithms at the time were either not sensitive enough or would have taken excessive compute time. To address this new challenge, a new software package, called BLASTZ was created. As the name of this specially developed program implies, BLASTZ (Schwartz et al. 2003) is based on the strategies of BLAST (Altschul et al. 1990), but optimized for whole genome alignments of diverged species. One optimization relied on having relatively high contiguity sequences, and even though the mouse and human genomes were called draft genomes, they were both of high enough quality to allow the program to assume that the matching regions occur in the same order and orientation in both sequences. The other optimization was to use a different scoring matrix for nucleotide substitutions and sequence gaps. These primary optimizations along with many other improved methodological approaches, all nicely detailed in BLASTZ manuscript, allowed these two genomes to be aligned in 481 central processing unit (CPU) days, and with 1024 CPUs available to the group, the wall clock time was less than a day. This essential comparative genomics step then allowed many others to start interpreting the results, with one being a statistical estimate of functionally constrained fraction of the human genome relative to the mouse genome, which when analyzed in 50 base-pair windows across the genomes totaled 5%, or 140 Mb of human genomic DNA.

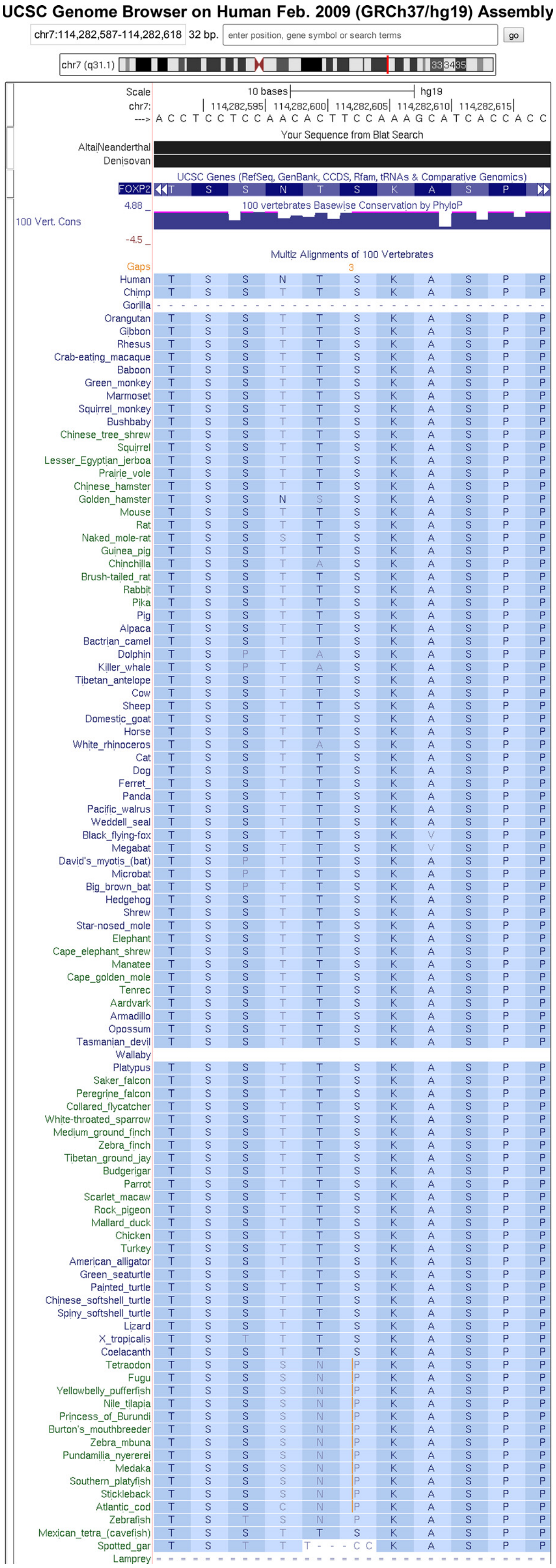
This number, 5%, was tantalizing in that we knew there were many more functionally important regions in the genome at the same level as coding sequence (CDS), but the locations of these regions were not as rigorously defined as CDSs. Thus in 2003, the **ENC**yclopedia **O**f **D**NA **E**lements (ENCODE) was launched to develop a variety of methods to "identify and precisely locate all of the protein-coding genes, non-protein coding genes and other sequence-based functional elements contained in the human DNA sequence. (http://www.genome.gov/10506706)" One of the key approaches was to use multi-species comparative genomics to improve the sensitivity and specificity of these elements. In the pilot phase of ENCODE Project Consortium (2007), 30 Mb (1%) of the human genome divided across 44 regions were selected for intense functional analyses including multispecies sequencing of orthologous regions in 28 other species.

Total sequence across all the species and orthologous regions was 546 Mb, and represented a new challenge for comparative genomic analyses. This time three different software packages (Brudno et al. 2003; Blanchette et al. 2004; Bray and Pachter 2004) were developed for alignment of the multispecies genomic sequences because the subsequent detection of the evolutionarily constrained regions was quite sensitive to the final alignments produced. Now with more species compared the resolution of the constrained regions improved to a median length of 19 bases and a minimum size of 8 bases, and overall, the total fraction of the human genome under evolutionary mammalian constraint remained at 5%, a testament to power of the original human-mouse comparative analysis result. However, the overlap of CDS (32%), UTRs (8%), and other ENCODE detected functional elements (20%) still left 40% of the genome identified as important using comparative genomics but with unknown function.

With the main phase of the ENCODE project now completed (Bernstein et al. 2012), we have a much more complete map of functional elements across the entire human genome. For this more recent genome-wide study, interspecies comparative genomics methods were applied to whole genomes of 29 mammals selected to maximize divergence across the four major mammalian clades (Lindblad-Toh et al. 2011). This resulted in a total effective branch length of 4.5 substitutions per site which, for example, translates into an incredibly infinitesimal probability of $<10^{-25}$ that a window of 12 nucleotides that are not under purifying selection will remain fixed across all 29 species. Today, the most resent compilation of genome-wide comparative genomic analyses includes 100 vertebrate species (www.genome.ucsc.edu), see Figure 1, and provides a tremendous resource to the community in interpreting the genome from an evolutionary foundation which was built upon decades of improvements in sequencing, computational, and statistical methods. Looking into the near future, the Genome 10K Project (https://genome10k.soe.ucsc.edu/) is coordinating the collection of samples from over 10,000 vertebrate species specifically designated for whole-genome sequencing to better understand vertebrate evolution (Genome 2009).

Intraspecies Comparative Genomics

In contrast to multispecies comparative genomics, intraspecies comparative genomics is used to find the variation across individuals of a given species. The first systematic effort to find large numbers of single-nucleotide polymorphisms (SNPs) in the human genome was through The SNP Consortium, which started generating data specifically for this effort in 1999 and completed in 2001. The original goal was to find at least 300,000 SNPs to give researchers landmarks across the genome to use for genetic association and linkage testing. This effort proved much more effective than originally planned, largely due to the acceleration of the Human Genome Project (Lander et al. 2001) during that time, with a final collection of over 1.4 million SNPs (Sachidanandam et al. 2001). The initial approach to discover SNPs did not require having the reference genome, because at the start of the project it was only 20% finished. Thus an approach called reduced representation shotgun sequencing (Altshuler et al. 2000) was developed, and proved to be an effective way to get enough sequencing reads to overlap to allow detection of variation from only a few hundreds of thousands of Sanger sequencing reads, instead of the then cost-prohibitive tens of millions of reads without this approach. However, by 2001, the draft human genome was nearly complete and random shotgun sequence from selected human genomic DNA samples proved to be much more cost effective. With these initial 1.4 million SNPs available, the focus turned toward understanding and mapping the haplotype structure of the human genome, however, other, more focused efforts, were indicating that many more SNPs were required to more completely resolve the haplotype map of the human genome (Mullikin et al. 2000). Thus, at the start of the human haplotype map project (HapMap) in 2003, focus continued on SNP discovery using random whole-genome shotgun sequences from individuals of European, Asian, and African ancestry and all compared to the improving human reference sequence. To map these Sanger reads, with lengths of 400–800 nucleotides in length, I developed and used the ssahaSNP algorithm (Ning et al. 2001) on the rapidly increasing number of reads generated by the genome sequencing centers. The optimizations of this algorithm used assumptions that the sequence of a given read would match with very few differences, so that the reference sequence could be indexed very efficiently in a large memory LINUX computer (over 12 gigabytes of random access memory) and the process of alignment became a memory lookup operation followed by a fast local alignment algorithm, making the speed of aligning a read to a reference genome essentially independent of the genome size. Even with the computers available in 2001, alignments raced along at 200 reads per second, which was three to four orders of magnitude faster than the version BLAST available then. The HapMap project contributed another six million SNPs to dbSNP, bringing the total in dbSNP to 9.2 million SNPs in October of 2005 (International HapMap Consortium 2005). With this SNP set available and a high-throughput genotype technology from Perlegen, phase II of HapMap proceeded quite quickly, culminating with a combined total of 3.1 million SNPs genotyped across 270 individuals from four geographically diverse populations (International HapMap Consortium 2007). The end result of these efforts and the continued improvements in genotyping technologies which utilized optimal subsets of SNPs based on the haplotype structures of the human genome populations enabled the huge expansion of genome-wide association studies (GWAS) which was reported in an earlier commentary in this journal (Adeyemo and Rotimi 2014). Furthermore, SNP discovery has been applied to hundreds of other species; see dbSNP (http://www.ncbi.nlm.nih.gov/SNP/index.html) for summaries of SNPs available across the kingdoms of life.



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr7:114,282,587-114,282,618 32 bp. enter position, gene symbol or search terms go

```
Scale          10 bases |------------------| hg19
chr7:          | 114,282,595| 114,282,600| 114,282,605| 114,282,610| 114,282,615|
--->  A C C T C C T C C A A C A C T T C C A A A G C A T C A C C A C C
                Your Sequence from Blat Search
AltaiNeanderthal
Denisovan
       UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
FOXP2   T   S   S   N   T   S   K   A   S   P
4.88 _        100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
-4.5 _
                Multiz Alignments of 100 Vertebrates
```

Gaps						3					
Human	T	S	S	N	T	S	K	A	S	P	P
Chimp	T	S	S	T	T	S	K	A	S	P	P
Gorilla	-	-	-	-	-	-	-	-	-	-	-
Orangutan	T	S	S	T	T	S	K	A	S	P	P
Gibbon	T	S	S	T	T	S	K	A	S	P	P
Rhesus	T	S	S	T	T	S	K	A	S	P	P
Crab-eating_macaque	T	S	S	T	T	S	K	A	S	P	P
Baboon	T	S	S	T	T	S	K	A	S	P	P
Green_monkey	T	S	S	T	T	S	K	A	S	P	P
Marmoset	T	S	S	T	T	S	K	A	S	P	P
Squirrel_monkey	T	S	S	T	T	S	K	A	S	P	P
Bushbaby	T	S	S	T	T	S	K	A	S	P	P
Chinese_tree_shrew	T	S	S	T	T	S	K	A	S	P	P
Squirrel	T	S	S	T	T	S	K	A	S	P	P
Lesser_Egyptian_jerboa	T	S	S	T	T	S	K	A	S	P	P
Prairie_vole	T	S	S	T	T	S	K	A	S	P	P
Chinese_hamster	T	S	S	T	T	S	K	A	S	P	P
Golden_hamster	T	S	S	N	S	S	K	A	S	P	P
Mouse	T	S	S	T	T	S	K	A	S	P	P
Rat	T	S	S	T	T	S	K	A	S	P	P
Naked_mole-rat	T	S	S	S	T	S	K	A	S	P	P
Guinea_pig	T	S	S	T	T	S	K	A	S	P	P
Chinchilla	T	S	S	T	A	S	K	A	S	P	P
Brush-tailed_rat	T	S	S	T	T	S	K	A	S	P	P
Rabbit	T	S	S	T	T	S	K	A	S	P	P
Pika	T	S	S	T	T	S	K	A	S	P	P
Pig	T	S	S	T	T	S	K	A	S	P	P
Alpaca	T	S	S	T	T	S	K	A	S	P	P
Bactrian_camel	T	S	S	T	T	S	K	A	S	P	P
Dolphin	T	S	P	T	A	S	K	A	S	P	P
Killer_whale	T	S	P	T	A	S	K	A	S	P	P
Tibetan_antelope	T	S	S	T	T	S	K	A	S	P	P
Cow	T	S	S	T	T	S	K	A	S	P	P
Sheep	T	S	S	T	T	S	K	A	S	P	P
Domestic_goat	T	S	S	T	T	S	K	A	S	P	P
Horse	T	S	S	T	T	S	K	A	S	P	P
White_rhinoceros	T	S	S	T	A	S	K	A	S	P	P
Cat	T	S	S	T	T	S	K	A	S	P	P
Dog	T	S	S	T	T	S	K	A	S	P	P
Ferret_	T	S	S	T	T	S	K	A	S	P	P
Panda	T	S	S	T	T	S	K	A	S	P	P
Pacific_walrus	T	S	S	T	T	S	K	A	S	P	P
Weddell_seal	T	S	S	T	T	S	K	A	S	P	P
Black_flying-fox	T	S	S	T	T	S	K	V	S	P	P
Megabat	T	S	S	T	T	S	K	V	S	P	P
David's_myotis_(bat)	T	S	P	T	T	S	K	A	S	P	P
Microbat	T	S	P	T	T	S	K	A	S	P	P
Big_brown_bat	T	S	P	T	T	S	K	A	S	P	P
Hedgehog	T	S	S	T	T	S	K	A	S	P	P
Shrew	T	S	S	T	T	S	K	A	S	P	P
Star-nosed_mole	T	S	S	T	T	S	K	A	S	P	P
Elephant	T	S	S	T	T	S	K	A	S	P	P
Cape_elephant_shrew	T	S	S	T	T	S	K	A	S	P	P
Manatee	T	S	S	T	T	S	K	A	S	P	P
Cape_golden_mole	T	S	S	T	T	S	K	A	S	P	P
Tenrec	T	S	S	T	T	S	K	A	S	P	P
Aardvark	T	S	S	T	T	S	K	A	S	P	P
Armadillo	T	S	S	T	T	S	K	A	S	P	P
Opossum	T	S	S	T	T	S	K	A	S	P	P
Tasmanian_devil	T	S	S	T	T	S	K	A	S	P	P
Wallaby											
Platypus	T	S	S	T	T	S	K	A	S	P	P
Saker_falcon	T	S	S	T	T	S	K	A	S	P	P
Peregrine_falcon	T	S	S	T	T	S	K	A	S	P	P
Collared_flycatcher	T	S	S	T	T	S	K	A	S	P	P
White-throated_sparrow	T	S	S	T	T	S	K	A	S	P	P
Medium_ground_finch	T	S	S	T	T	S	K	A	S	P	P
Zebra_finch	T	S	S	T	T	S	K	A	S	P	P
Tibetan_ground_jay	T	S	S	T	T	S	K	A	S	P	P
Budgerigar	T	S	S	T	T	S	K	A	S	P	P
Parrot	T	S	S	T	T	S	K	A	S	P	P
Scarlet_macaw	T	S	S	T	T	S	K	A	S	P	P
Rock_pigeon	T	S	S	T	T	S	K	A	S	P	P
Mallard_duck	T	S	S	T	T	S	K	A	S	P	P
Chicken	T	S	S	T	T	S	K	A	S	P	P
Turkey	T	S	S	T	T	S	K	A	S	P	P
American_alligator	T	S	S	T	T	S	K	A	S	P	P
Green_seaturtle	T	S	S	T	T	S	K	A	S	P	P
Painted_turtle	T	S	S	T	T	S	K	A	S	P	P
Chinese_softshell_turtle	T	S	S	T	T	S	K	A	S	P	P
Spiny_softshell_turtle	T	S	S	T	T	S	K	A	S	P	P
Lizard	T	S	S	T	T	S	K	A	S	P	P
X_tropicalis	T	S	T	T	T	S	K	A	S	P	P
Coelacanth	T	S	S	T	T	S	K	A	S	P	P
Tetraodon	T	S	S	S	N	P	K	A	S	P	P
Fugu	T	S	S	S	N	P	K	A	S	P	P
Yellowbelly_pufferfish	T	S	S	S	N	P	K	A	S	P	P
Nile_tilapia	T	S	S	S	N	P	K	A	S	P	P
Princess_of_Burundi	T	S	S	S	N	P	K	A	S	P	P
Burton's_mouthbreeder	T	S	S	S	N	P	K	A	S	P	P
Zebra_mbuna	T	S	S	S	N	P	K	A	S	P	P
Pundamilia_nyererei	T	S	S	S	N	P	K	A	S	P	P
Medaka	T	S	S	S	N	P	K	A	S	P	P
Southern_platyfish	T	S	S	S	N	P	K	A	S	P	P
Stickleback	T	S	S	S	N	P	K	A	S	P	P
Atlantic_cod	T	S	S	C	N	P	K	A	S	P	P
Zebrafish	T	S	T	S	N	P	K	A	S	P	P
Mexican_tetra_(cavefish)	T	S	S	T	T	S	K	A	S	P	P
Spotted_gar	T	S	T	T	T - -	- C C	K	A	S	P	P
Lamprey	=	=	=	=	=	=	=	=	=	=	=

Figure 1. This UCSC genome browser image of a 32 base-wide window of the FOXP2 gene overlaps with one of the two human "speech" amino acid adaptation alleles (Enard et al. 2002), from a threonine as the ancestral allele to an asparagine in the human genome. Note that most of the alleles in the fourth column of the 100-way multispecies alignment is threonine, thus highly conserved. However, along with human, the Altai Neanderthal and the Denisovan genomes agree at the nucleotide level, as show with the solid black bars from a UCSC blat alignment of the orthologous sequence from these hominins' genomes, indicating that this change happened after the split of hominins from the human–chimpanzee common ancestor and was fixed before the split of the human–neanderthal–denisovan common ancestor.

Comparative Genomics Insights into Hominin Evolution

Paleoanthropology over the last 150 years has built a tree of hominin evolution based on fossils that date back over the last 4–5 million years. Some recent and well preserved fossils of now extinct hominins dating back 30–100,000 years ago have been shown to contain enough endogenous DNA to allow us to sequence their genomes, and by comparing these archaic genomes to modern humans, gain new insights into human evolution. The first attempt to extract and sequence DNA from a Neanderthal bone targeted the hyper-variable region of the mitochondria (Krings et al. 1997). Using 13 overlapping PCR primer-pair amplification products, Dr. Pääbo's group was able to generate 379 bases of contiguous consensus sequence and compared this to modern human sequence and chimpanzee sequence of the same mitochondrial region, thus started the era of paleogenomics. As the sequencing technologies and methods advanced, first with the arrival of the 454 sequencing instrument and later with the Solexa, now Illumina, massively parallel sequencing instrument, sequencing the entire genome of the Neanderthal was completed (Green et al. 2010).

Subsequently, with the discovery of a very well preserved Neanderthal toe bone from the Denisova cave in Altai mountains along with advancements in archaic DNA extraction methods and sequencer throughput, a new and much improved Neanderthal genome was completed (Prufer et al. 2014).

One of the primary questions we hoped to find an answer to from the genomes of our closest archaic ancestors: is there any evidence, or not, of interbreeding when humans encountered Neanderthals as they left Africa and entered the domain that Neanderthals had occupied for the previous 400,000 years? The method to detect this required, in addition to the Neanderthal genome, whole genome sequences of modern humans from a variety of ancestral population locations. In the earlier work, these sequences were ascertained from five individuals, one San from Southern Africa, one Yoruba from West Africa, one Papua New Guinean, one Han Chinese, and one French from Western Europe. Each was sequenced to four- to sixfold coverage on the Illumina GAII platform. To avoid biased results by comparing these sequences to the reference human genome, which is a mosaic assembly of a few individuals of various ancestral population locations, the sequences were compared to the chimpanzee genome since the common ancestor to humans and chimpanzees predates the common ancestor to humans and Neanderthals, and the chimpanzee genome sequence is similar enough to the human and Neanderthal genomes sequence to allow robust alignments. If there had been interbreeding of Neanderthals with humans that left Africa between 30–80,000 years ago, and if their offspring remained in a geographical area, e.g., Europe, since that time, comparing pairs of modern human genomes, say a European and a Papua New Guinean, then one looks at all positions where there are differences between these present-day humans and count how many times the Neanderthal genome agrees with one versus the other. If the Neanderthal allele agreeing counts are statistically higher for the individual of European ancestry versus the individual of Papua New Guinea ancestry, that would show evidence of greater Neanderthal contribution to the European than to the Papua New Guinean. Putting this in a statistical framework, the "D" statistic developed for this very analysis, was able to determine evidence of interbreeding that occurred early in the migration of humans leaving Africa, about 50–80,000 years ago, since all three out-of-Africa ancestry individuals contained approximately the same skew of more Neanderthal alleles when compared to the Southern African or West African individuals. A similar analysis was repeated with the sequence of the Neanderthal individual from the Altai Mountains, along with an increased number of 25 present-day human genomes and the evidence for interbreeding remained, along with additional gene flow signatures, see figure 8 in (Prufer et al. 2014), which also incorporated a newly discovered hominin from the same Denisova cave (Krause et al. 2010; Meyer et al. 2012).

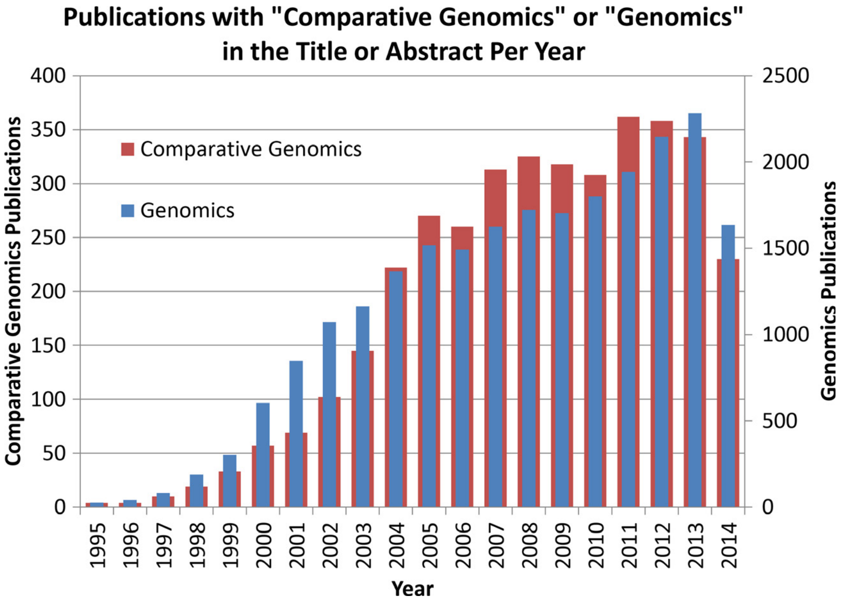


Figure 2. A PubMed search of publications with the exact combination and order of the words "comparative genomics" in the title or abstract identifies 3752 articles. This chart shows the growth of publications in this field year-by-year, and for comparison includes the same search for "genomics."

Summary

In this commentary I have only highlighted a few dimensions that comparative genomics has reached into. Looking at a PubMed search of publications with the exact combination and order of the words "comparative genomics" in the title or abstract identifies 3752 articles as of the date of this writing. The chart in Figure 2 shows the growth of this field, which at first lagged in growth relative to the same search for "genomics," but overall tracks this more general field of research. Other dimensions of comparative genomics, beyond the three areas I touched on above, include intraindividual comparative genomics (Cheng et al. 2012; Biesecker and Spinner 2013; Watson et al. 2013), human microbiome comparative genomics (Human Microbiome Project Consortium 2012) and how comparative genomics can shed light on a multidrug-resistant bacteria spread through a hospital (Snitkin et al. 2012). Clearly, as the field of genomics continues to expand, comparative genomics will always be an essential and central enabling component to help us discover and better understand the complexities, intricacies, and interrelatedness of the genomics of life.

Acknowledgments

This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Conflict of Interest

None declared.

References

Adeyemo, A., and C. Rotimi. 2014. What does genomic medicine mean for diverse populations? Mol Genet Genomic Med 2:3–6.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. van Etten, J. Baldwin, L. Linton, et al. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516.

Bernstein, B. E., E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Biesecker, L. G., and N. B. Spinner. 2013. A genomic view of mosaicism and human disease. Nat. Rev. Genet. 14:307–320.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Bray, N., and L. Pachter. 2004. MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 14:693–699.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. C. S. Program, et al. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721–731.

Cheng, L., N. F. Hansen, L. Zhao, Y. Du, C. Zou, F. X. Donovan, et al. 2012. Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. Cell Stem Cell 10:337–344.

Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe, T. Kitano, et al. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872.

ENCODE Project Consortium, E. Birney, Stamatoyannopoulos, J. A., A. Dutta, R. Guigo, T. R. Gingeras, et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447: 799–816.

Fitch, W. M. 1966. An improved method of testing for evolutionary homology. J. Mol. Biol. 16:9–16.

Genome, K. C. O. S. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100:659–674.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature, 486: 207–214.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature, 437: 1299–1320.

International HapMap Consortium; K. A. Frazer, K. A. Ballinger, D. R. Cox, D. A. Hinds, L. Stuve, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature, 449,851–861.

Krause, J., Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, A. P. Derevianko, et al. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464:894–897.

Krings, M., A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Paabo. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature: 478, 476–482.

Meyer, M., M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226.

Mouse Genome Sequencing Consortium; R. H. Waterston, Lindblad-Toh, K., E. Birney, J. Rogers, J. F. Abril, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, 420: 520–562.

Mullikin, J. C., S. E. Hunt, C. G. Cole, B. J. Mortimore, C. M. Rice, J. Burton, et al. 2000. An SNP map of human chromosome 22. Nature 407:516–520.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

Ning, Z., A. J. Cox, and J. C. Mullikin. 2001. SSAHA: a fast search method for large DNA databases. Genome Res. 11:1725–1729.

Prufer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Sherry, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl Acad. Sci. 74:5463–5467.

Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, et al. 2003. Human-mouse alignments with BLASTZ. Genome Res. 13:103–107.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197.

Snitkin, E. S., A. M. Zelazny, P. J. Thomas, F. Stock, NISC Comparative Sequencing Program Group; D. K. Henderson, et al. 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med, 4, 148ra116.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al. 2001. The sequence of the human genome. Science 291:1304–1351.

Watson, I. R., K. Takahashi, P. A. Futreal, and L. Chin. 2013. Emerging patterns of somatic mutations in cancer. Nat. Rev. Genet. 14:703–718.