Interrogating the genome: approaches and tools for accessing genomic databases

Reference genome DNA sequences for thousands of species have now been determined, and deposited in freely-accessible public databases. The availability of these genomic data has revolutionized biology, enabling insights into the evolution of life on earth, genome-wide screens to characterize genotype-phenotype relationships, and investigations into disease-linked genetic variations, providing exciting new possibilities for personalized medicine.

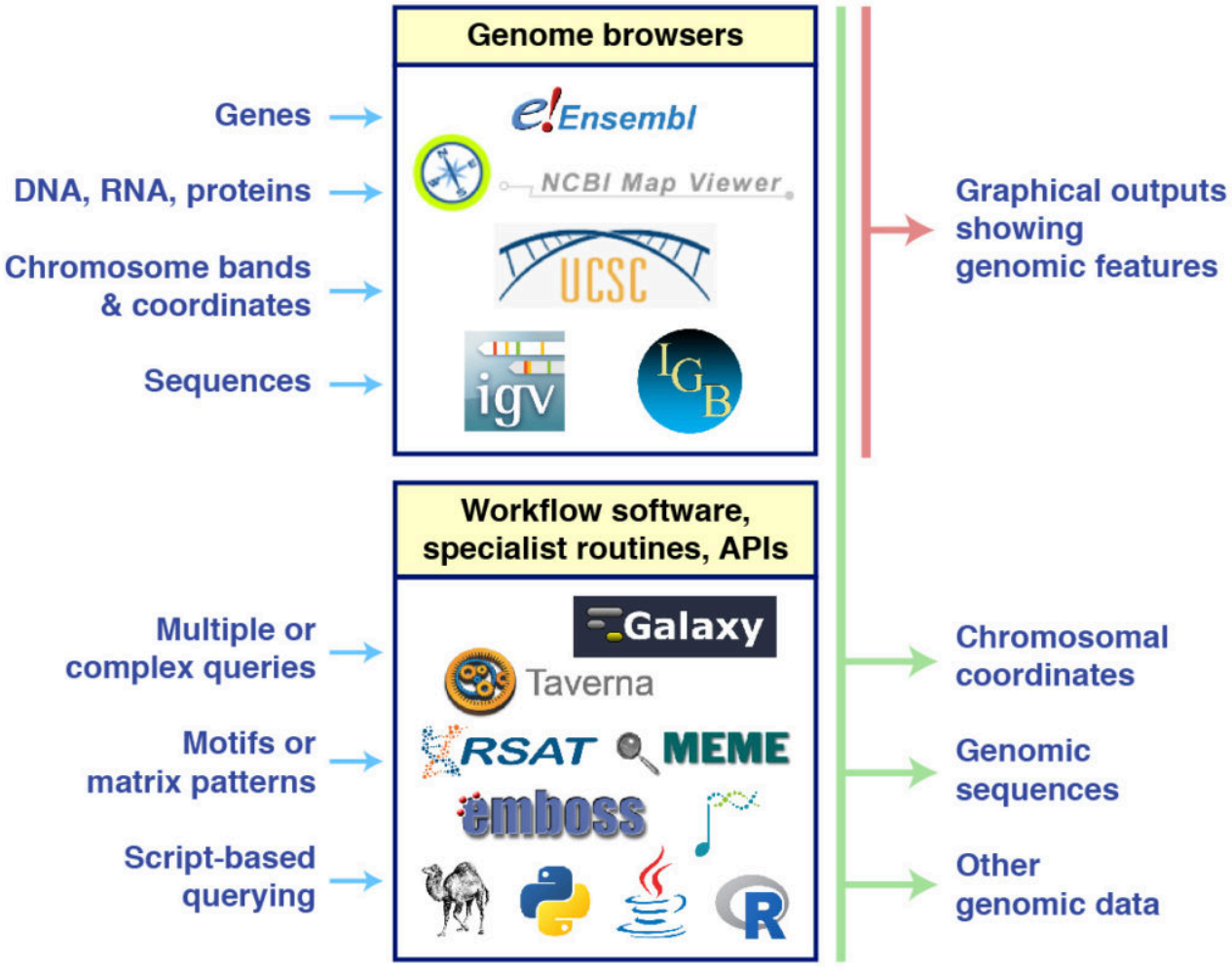

Fig. 1. Query and output types for interrogating genome databases.

Genomic sequences can be downloaded in their entirety from public databases, but a more powerful way of accessing these data is via genome browser software. These free programs are relatively easy to use despite their sophistication, and allow users to rapidly query the genome of

the species in question, access sequence data, and identify and visualize genomic loci in the context of other genomic features, generating publication-quality figures.

In parallel to genome sequences are gene sets. The definition of a gene now extends beyond a section of genomic DNA coding for a protein, to encompass regions whose transcripts are non-coding RNAs, which in turn regulate the expression of further genes. Gene sets such as GENCODE and RefSeq include predicted and expert-curated entries and are represented graphically in genome browsers. Epigenetic modification data, mapped by large-scale projects such as ENCODE and modENCODE, can also be visualized using such genome browsers.

Ensembl is a family of genome databases and browsers covering metazoan species, with expert-curated VEGA gene annotations. Ensembl Genomes covers small metazoa, plants, fungi, protists and bacteria. The Map Viewer browser from the National Center for Biotechnology Information (NCBI) provides a graphical overview of genomes in the context of NCBI sequence entries. The University of California Santa Cruz (UCSC) Genome Browser is a popular and flexible tool integrating a wide range of functionalities and output modes. Other stand-alone browsers include the Integrated Genome Browser (IGB) and Integrative Genomics Viewer (IGV). All of these browsers also allow users to upload and display experimental data alongside reference genomes.

Using these browsers, genomic databases can be accessed using a variety of query types (Fig. 1). Searches could be based on known genes, using their names, symbols, or identifier codes (IDs), or IDs of DNA molecules (e.g. cDNA clones), protein-coding and non-coding RNAs, or proteins. Searches can also be based on karyotype banding patterns (e.g. 17p13.1) and chromosomal coordinates (e.g. 13:72708357-72727687).

Nucleotide sequences of various lengths may be used as queries, using algorithms such as BLAST or BLAT to identify regions on the genome of the chosen species matching the query sequence; tools for launching such searches are integrated into most genome browsers.

Searching genome sequences to identify loci that match motifs (which may include sequence degeneracy and variable-length spaces) and matrix-based patterns (which reflects differing frequencies of occurrence of bases within each position of the motif) requires more specialist software, such as routines from the EMBOSS, MEME and RSAT suites.

Resource / Tool	Website
Ensembl	http://www.ensembl.org/
Ensembl Genomes	http://ensemblgenomes.org/
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/
NCBI Viral Genomes	http://www.ncbi.nlm.nih.gov/genome/viruses/
UCSC Genome Browser	https://genome.ucsc.edu/
Integrated Genome Browser	http://bioviz.org/igb/
Integrative Genomics Viewer	http://software.broadinstitute.org/software/igv/
Galaxy	https://galaxyproject.org/
Taverna	https://taverna.incubator.apache.org/
RSAT suite	http://www.rsat.eu/
MEME Suite	http://meme-suite.org/
EMBOSS suite	http://emboss.sourceforge.net/
Bio* Toolkits	https://www.open-bio.org/wiki/Main_Page
Bioconductor	https://www.bioconductor.org/
Ensembl REST API	https://rest.ensembl.org/

Fig. 2. Web addresses for the resources and tools described in the text.

Multiple genomic queries can also be performed using facilities such as Ensembl's BioMart. More complex queries can be performed using UCSC's Table Browser, and analytical pipelines can be established using software such as Taverna and Galaxy. The latter includes routines that for Next-Generation Sequencing (NGS) projects allows the upload and mapping of hundreds of millions of reads of experimental data for alignment to the genome.

For users with programming experience, libraries such as the Bio*Toolkits and Bioconductor, and application programming interfaces (APIs) exist for popular languages such as Perl, Python, R and Java, allowing querying and retrieval of genomic data; the REpresentational State Transfer (RESTful) API from Ensembl enables data access using any language.

These freely-available resources (Fig. 2) are valuable and powerful additions to the molecular biologist's toolkit, allowing full access to the fruits of the genomic revolution.

James R. A. Hutchins
Institute of Human Genetics (IGH), Université de Montpellier, France

Publication

Genomic Database Searching.
Hutchins JR
Methods Mol Biol. 2017