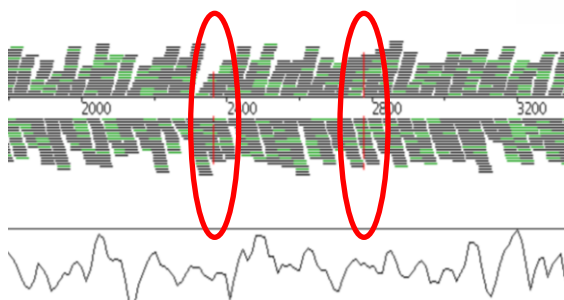
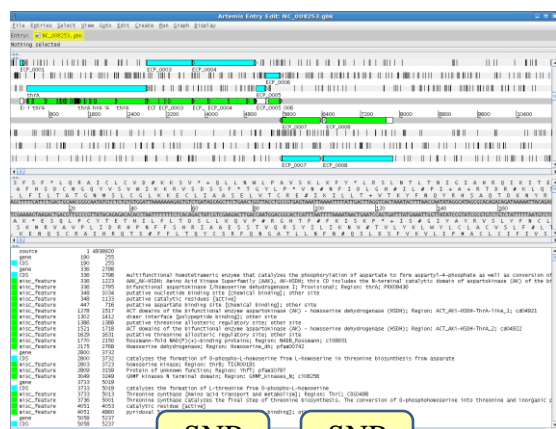


Workshop on Genome Annotation

29–30 September 2011



INFORMATICS FACILITY

Workshop Manual

International Centre for Genetic Engineering and Biotechnology (ICGEB)

Aruna Asaf Ali Marg, New Delhi 110067 India

Workshop on Genome Annotation

29-30 September 2011

Bioinformatics Laboratory
Structural and Computational Biology Group
International Center for Genetic Engineering and
Biotechnology (ICGEB)
New Delhi, India

Sponsored By:

Biotechnology Information System Network (BTIS)
Department of Biotechnology
Ministry of Science & Technology, Government of India

Staff

Principal Coordinator

Dr. Dinesh Gupta
Staff Research Scientist
E-mail: dinesh@icgeb.res.in

Guest Speaker

Dr. Mukesh Jain
Staff Scientist
E-mail: mjain@nipgr.res.in

Course Instructors

Abhinav Kaushik
Research Fellow
E-mail: abhinav@icgeb.res.in

Achal Rastogi
Research Fellow
E-mail: achal@icgeb.res.in

Sangeetha Subramaniam
Ph.D. Student
E-mail: sangi@icgeb.res.in

Zeenia Jagga
Ph.D. Student
E-mail: zeenia@icgeb.res.in

With grateful thanks, acknowledging Mayank Gupta, Anil and Babita Singh for their assistance in organizing this workshop.

Workshop Schedule

	Thursday September 29, 2011	Friday September 30, 2011
10:00	Introduction of Participants	Guest Lecture By Dr. Mukesh Jain (NIPGR)
10:15	Introduction to Genome Annotation By Dr. Dinesh Gupta (ICGEB)	
10:30		
10:45		
11:00	Coffee-Break	Coffee-Break
11:15	Basic Module: Unix, BLAST, EMBOSS (with hands-on Session) By Dr. Dinesh Gupta (ICGEB)	
11:30		
11:45		
12:00		
12:15		
12:30		
12:45		[AR]
13:00	Lunch	Lunch
14:00	CLC-Workbench	Module 2: Mapping Sequence Data (with hands-on Session) [Z]
14:15		
14:30		
14:45		
15:00	Coffee-Break	Coffee-Break
15:15		
15:30	CLC-Workbench continues...	Module 3: ACT (with hands-on Session) [AK]
15:45		
16:00		
16:15		
16:30		
16:45		
17:00	Mop-up Session	Mop-up Session

Workshop on Genome Annotation

29-30 September 2011

ICGEB, New Delhi, India

The two-day workshop aims to give researchers with a working knowledge of computational sequence analysis, a firm grounding in the use of the latest genome analysis software (Artemis and ACT) developed at the Wellcome Trust Institute Pathogen Sequencing Unit (PSU) and an insight into in-silico Next Generation Sequence (NGS) data analysis.

[Artemis](#) is a powerful annotation tool and DNA viewer that allows the user to analyze sequence data from databases such as EMBL or Genbank. [ACT](#) is a comparative genomic tool that allows direct, and interactive, comparisons of multiple genomes/read sequences. This enables the user to exploit the growing number of genomes and NGS data from closely related organisms to look at genome architecture and evolution.

The course will be taught by members of the Bioinformatics Laboratory, Structural and Computational Biology Group and will take the form of a series of modules covering most aspects of sequence analysis and exploitation. Each module will be introduced with a short talk followed by 'hands on', to illustrate points in whole genome analysis.

Module 1 Artemis

(using *S. typhi*)

Introduction

Artemis is a free DNA viewer and annotation tool written by Kim Rutherford (Rutherford *et al.*, 2000). It is routinely used by the Sanger Institute Pathogen Sequencing Unit for annotation and analysis of both prokaryotic and eukaryotic genomes. The program allows the user to view simple sequence files, EMBL/Genbank entries and the results of sequence analyses in a highly interactive and intuitive graphical format. Artemis is designed to present multiple sets/types of information within a single context. This manifests itself as the ability to zoom in to inspect DNA sequence motifs and zoom out to view local gene architecture (e.g. operons), several kilobases of a genome or even an entire genome in one screen. It is also possible to perform some analyses within Artemis with the output stored for later access.

Aims

The aim of this part of the Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and crannies of Artemis that are not featured in the exercises in this manual. Like all the Modules in this workshop, the key is 'if you don't understand please ask'.

Artemis Exercise 1 Part I

1. Starting up the Artemis software

Navigate your way into the correct directory for this module

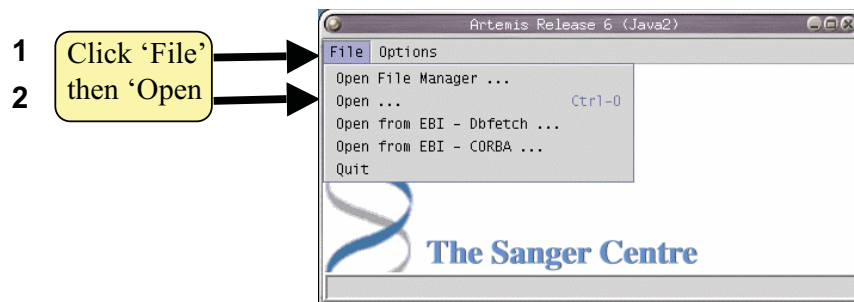
Then type:

```
art & [return]
```

A small start-up window will appear (see below).

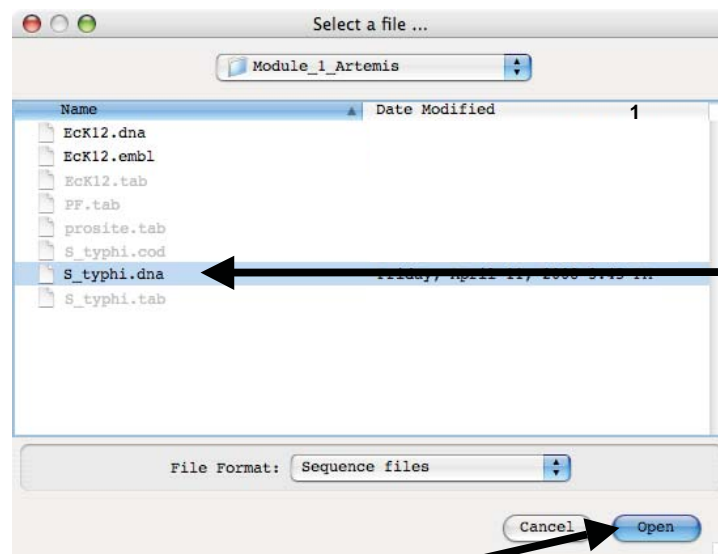
Now follow the sequence of numbers to load up the *Salmonella* Typhi chromosome sequence.

Ask a demonstrator for help if you have any problems.



For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.

In the 'Options' menu you can switch between prokaryotic and eukaryotic mode.



3

Single click to select DNA file

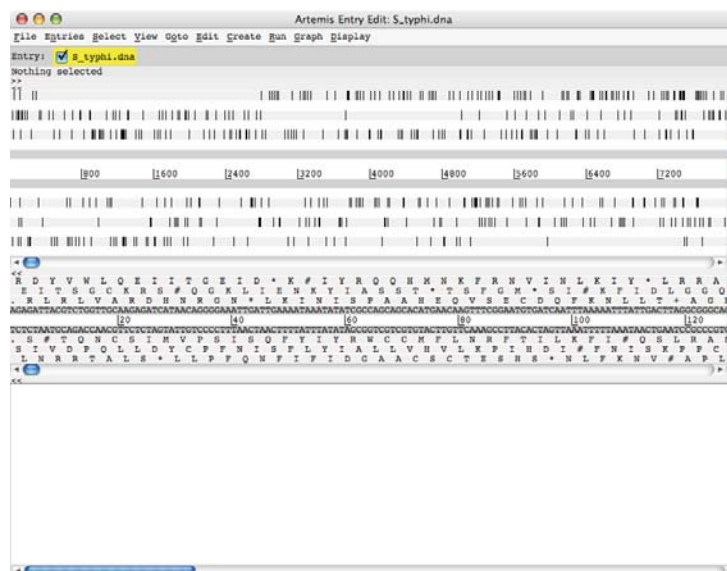
4

Single click to open file in Artemis then wait

DNA sequence files will have the suffix '.dna'. Annotation files end with '.tab'. Use this feature to select the type of file displayed in this panel.

2. Loading annotation files (entries) into Artemis

Hopefully you will now have an Artemis window like this! If not, ask a demonstrator for assistance.



Now follow the numbers to load up the annotation file for the *Salmonella* Typhi chromosome.

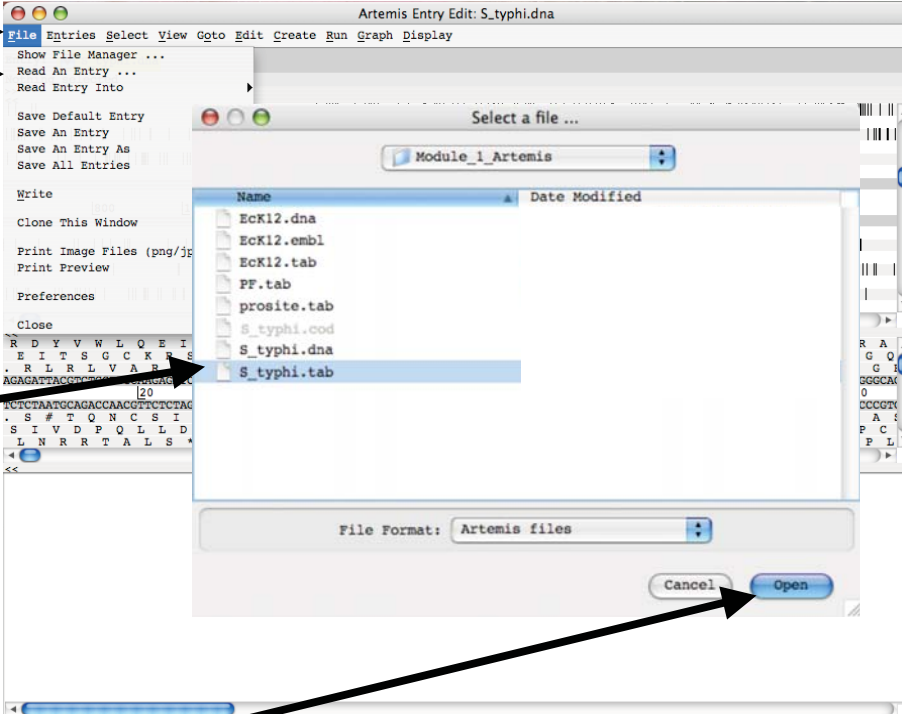
- 1

Click 'File' then 'Read an Entry'

Entry = file
- 2

Single click to select tab file
- 3

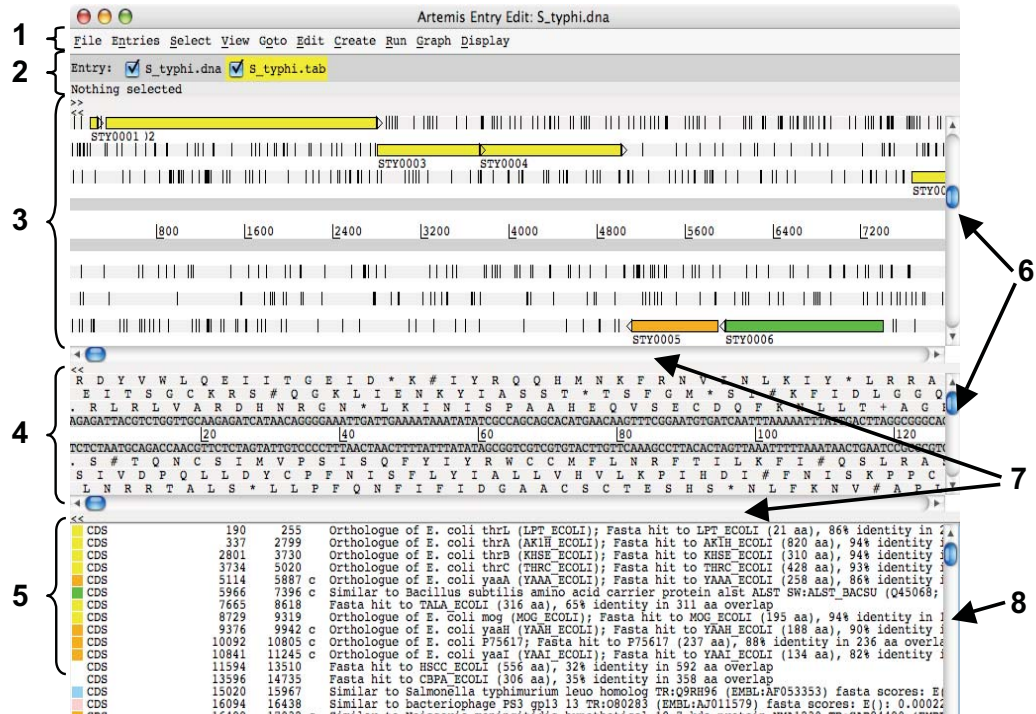
Single click to open file in Artemis then wait



What's an "Entry"? It's a file of DNA and/or features which can be overlaid onto the sequence information displayed in the main Artemis view panel.

3. The basics of Artemis

Now you have an Artemis window open let's look at what's in there.



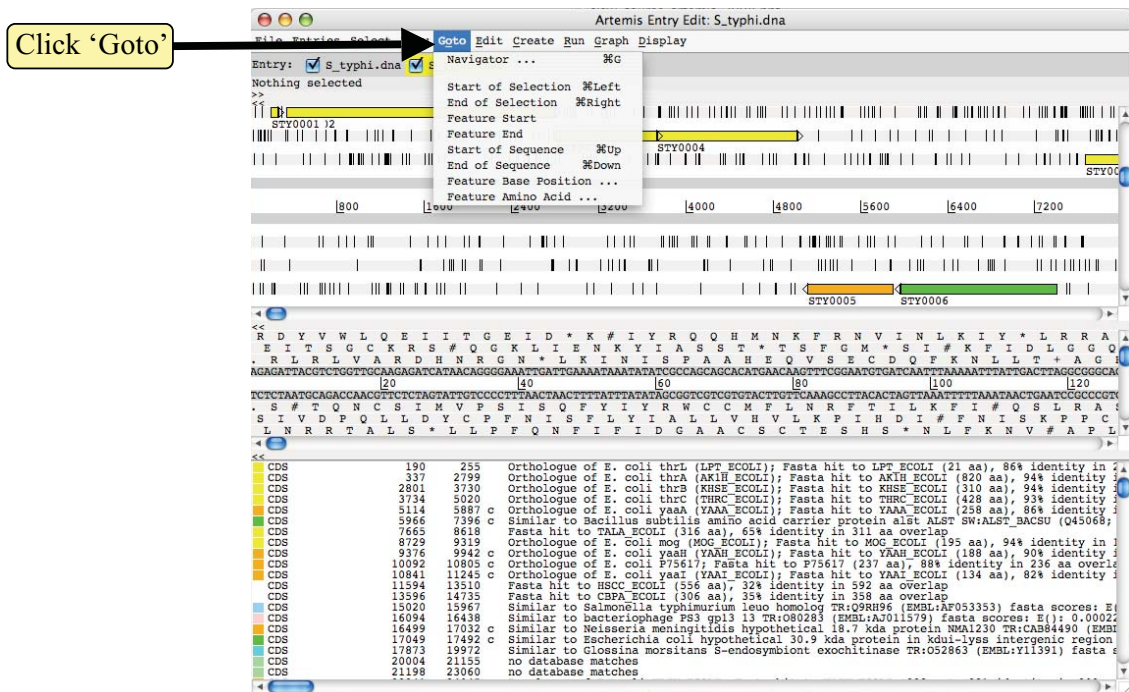
1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case gene STY0003 (top line).
3. This is the main sequence view panel. The central 2 grey lines represent the forward (top) and reverse (bottom) DNA strands. Above and below those are the 3 forward and 3 reverse reading frames. Stop codons are marked as black vertical bars. Genes and other features (eg. Pfam and Prosite matches) are displayed as coloured boxes. We will refer to genes as coding sequences or CDSs from now on.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene highlighted. The list can be scrolled (8, below).
6. Sliders for zooming view panels.
7. Sliders for scrolling along the DNA.
8. Slider for scrolling feature list.

4. Getting around in Artemis

The 3 main ways of getting to where you want to be in Artemis are the Goto drop-down menu, the Navigator and the Feature Selector. The best method depends on what you're trying to do and knowing which one to use comes with practice.

4.1 The 'Goto' menu

The functions on this menu (ignore the Navigator for now) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. This one's really intuitive so give it a try!



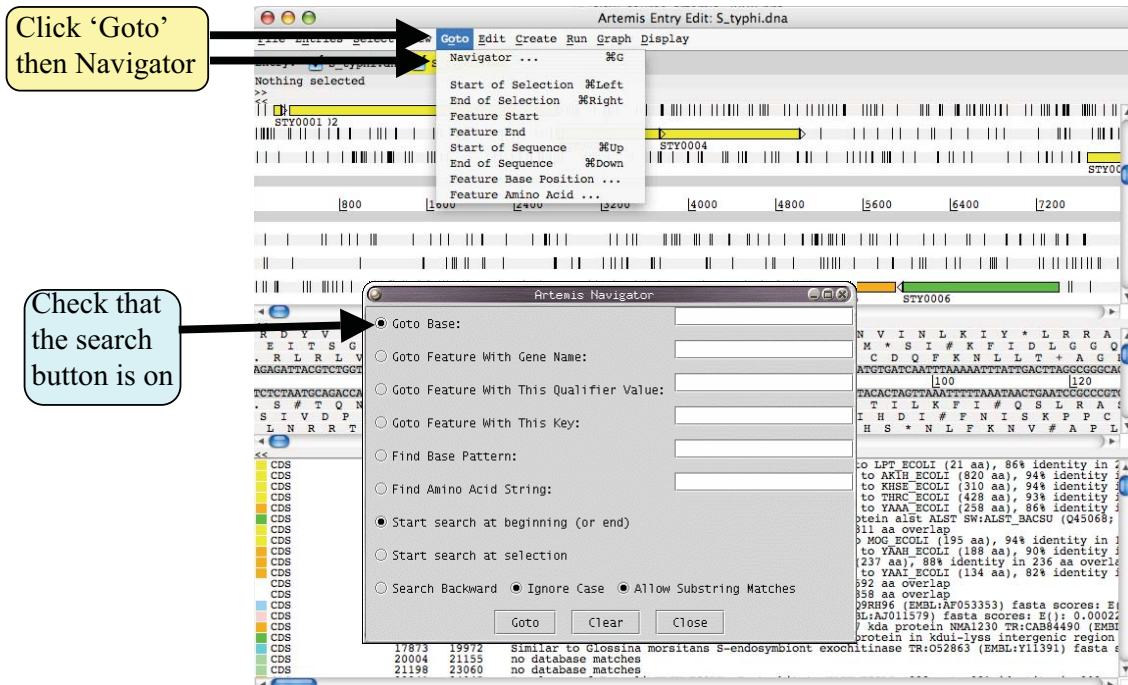
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try!

Suggested tasks:

1. Zoom out, highlight a large region of sequence by clicking the left hand button and dragging the cursor then go to the start and end of the highlighted region.
2. Select a gene then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a gene. Within it, go to a base (nucleotide) and/or amino acid of your choice.

4.2 Navigator

The Navigator panel is fairly intuitive so open it up and give it a try.



Suggestions of where to go:

1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try 'fts').
3. Use 'Goto Feature With This Qualifier value' to search the contents of all qualifiers for a particular term. For example using the word 'pseudogene' will take you to the next feature with the word 'pseudogene' in any of its qualifiers. Note how repeated clicking of the 'Goto' button takes you through the pseudogenes as they occur on the chromosome.
4. tRNA genes. Type 'tRNA' in the 'Goto Feature With This Key'.
5. Regulator-binding DNA consensus sequence (real or made up!). Note that degenerate base values can be used (Appendix III).
6. Amino acid consensus sequences (real or made up!). You can use 'X's. Note that it searches all six reading frames regardless of whether the amino acids are encoded or not.

Clearly there are many more features of Artemis which we will not have time to explain in detail. Before getting on with this next section it might be worth browsing the menus. Hopefully you will find most of them easy to understand.

Artemis Exercise 1 Part II

This part of the exercise uses the files and data you already have loaded into Artemis from Part I. By a method of your choice go to the region located between bases 1625084 to 1664823 on the DNA sequence. This region is bordered by *ribE* gene which codes for riboflavin synthase alpha chain. You can use either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.

Artemis Entry Edit: S.typhi.dna

File Entries Select View Goto Edit Create Run Graph Display

Entry: S.typhi.dna S.typhi.tab

Selected feature: bases 77 tRNA (/colour=4 /note="tRNA Val anticodon GAC, Cove score 96.48")

STY1695

STY1696

STY1697

STY1698

STY1699

STY1700

STY1701

STY1702

STY1703

STY1704

STY1705

STY1706

STY1707

STY1708

STY1709

STY1710

STY1711

STY1712

STY1713

STY1714

STY1715

STY1716

STY1717

STY1718

STY1719

STY1720

STY1721

STY1722

STY1723

STY1724

STY1725

STY1726

STY1727

STY1728

STY1729

STY1730

STY1731

STY1732

STY1733

STY1734

STY1735

STY1736

STY1737

STY1738

STY1739

STY1740

STY1741

STY1742

STY1743

STY1744

STY1745

STY1746

STY1747

STY1748

STY1749

STY1750

STY1751

STY1752

STY1753

STY1754

STY1755

STY1756

STY1757

STY1758

STY1759

STY1760

STY1761

STY1762

STY1763

STY1764

STY1765

STY1766

STY1767

STY1768

STY1769

STY1770

STY1771

STY1772

STY1773

STY1774

STY1775

STY1776

STY1777

STY1778

STY1779

STY1780

STY1781

STY1782

STY1783

STY1784

STY1785

STY1786

STY1787

STY1788

STY1789

STY1790

STY1791

STY1792

STY1793

STY1794

STY1795

STY1796

STY1797

STY1798

STY1799

STY1800

STY1801

STY1802

STY1803

STY1804

STY1805

STY1806

STY1807

STY1808

STY1809

STY1810

STY1811

STY1812

STY1813

STY1814

STY1815

STY1816

STY1817

STY1818

STY1819

STY1820

STY1821

STY1822

STY1823

STY1824

STY1825

STY1826

STY1827

STY1828

STY1829

STY1830

STY1831

STY1832

STY1833

STY1834

STY1835

STY1836

STY1837

STY1838

STY1839

STY1840

STY1841

STY1842

STY1843

STY1844

STY1845

STY1846

STY1847

STY1848

STY1849

STY1850

STY1851

STY1852

STY1853

STY1854

STY1855

STY1856

STY1857

STY1858

STY1859

STY1860

STY1861

STY1862

STY1863

STY1864

STY1865

STY1866

STY1867

STY1868

STY1869

STY1870

STY1871

STY1872

STY1873

STY1874

STY1875

STY1876

STY1877

STY1878

STY1879

STY1880

STY1881

STY1882

STY1883

STY1884

STY1885

STY1886

STY1887

STY1888

STY1889

STY1890

STY1891

STY1892

STY1893

STY1894

STY1895

STY1896

STY1897

STY1898

STY1899

STY1900

STY1901

STY1902

STY1903

STY1904

STY1905

STY1906

STY1907

STY1908

STY1909

STY1910

STY1911

STY1912

STY1913

STY1914

STY1915

STY1916

STY1917

STY1918

STY1919

STY1920

STY1921

STY1922

STY1923

STY1924

STY1925

STY1926

STY1927

STY1928

STY1929

STY1930

STY1931

STY1932

STY1933

STY1934

STY1935

STY1936

STY1937

STY1938

STY1939

STY1940

STY1941

STY1942

STY1943

STY1944

STY1945

STY1946

STY1947

STY1948

STY1949

STY1950

STY1951

STY1952

STY1953

STY1954

STY1955

STY1956

STY1957

STY1958

STY1959

STY1960

STY1961

STY1962

STY1963

STY1964

STY1965

STY1966

STY1967

STY1968

STY1969

STY1970

STY1971

STY1972

STY1973

STY1974

STY1975

STY1976

STY1977

STY1978

STY1979

STY1980

STY1981

STY1982

STY1983

STY1984

STY1985

STY1986

STY1987

STY1988

STY1989

STY1990

STY1991

STY1992

STY1993

STY1994

STY1995

STY1996

STY1997

STY1998

STY1999

STY2000

STY2001

STY2002

STY2003

STY2004

STY2005

STY2006

STY2007

STY2008

STY2009

STY2010

STY2011

STY2012

STY2013

STY2014

STY2015

STY2016

STY2017

STY2018

STY2019

STY2020

STY2021

STY2022

STY2023

STY2024

STY2025

STY2026

STY2027

STY2028

STY2029

STY2030

STY2031

STY2032

STY2033

STY2034

STY2035

STY2036

STY2037

STY2038

STY2039

STY2040

STY2041

STY2042

STY2043

STY2044

STY2045

STY2046

STY2047

STY2048

STY2049

STY2050

STY2051

STY2052

STY2053

STY2054

STY2055

STY2056

STY2057

STY2058

STY2059

STY2060

STY2061

STY2062

STY2063

STY2064

STY2065

STY2066

STY2067

STY2068

STY2069

STY2070

STY2071

STY2072

STY2073

STY2074

STY2075

STY2076

STY2077

STY2078

STY2079

STY2080

STY2081

STY2082

STY2083

STY2084

STY2085

STY2086

STY2087

STY2088

STY2089

STY2090

STY2091

STY2092

STY2093

STY2094

STY2095

STY2096

STY2097

STY2098

STY2099

STY2100

STY2101

STY2102

STY2103

STY2104

STY2105

STY2106

STY2107

STY2108

STY2109

STY2110

STY2111

STY2112

STY2113

STY2114

STY2115

STY2116

STY2117

STY2118

STY2119

STY2120

STY2121

STY2122

STY2123

STY2124

STY2125

STY2126

STY2127

STY2128

STY2129

STY2130

STY2131

STY2132

STY2133

STY2134

STY2135

STY2136

STY2137

STY2138

STY2139

STY2140

STY2141

STY2142

STY2143

STY2144

STY2145

STY2146

STY2147

STY2148

STY2149

STY2150

STY2151

STY2152

STY2153

STY2154

STY2155

STY2156

STY2157

STY2158

STY2159

STY2160

STY2161

STY2162

STY2163

STY2164

STY2165

STY2166

STY2167

STY2168

STY2169

STY2170

STY2171

STY2172

STY2173

STY2174

STY2175

STY2176

STY2177

STY2178

STY2179

STY2180

STY2181

STY2182

STY2183

STY2184

STY2185

STY2186

STY2187

STY2188

STY2189

STY2190

STY2191

STY2192

STY2193

STY2194

STY2195

STY2196

STY2197

STY2198

STY2199

STY2200

STY2201

STY2202

STY2203

STY2204

STY2205

STY2206

STY2207

STY2208

STY2209

STY2210

STY2211

STY2212

STY2213

STY2214

STY2215

STY2216

STY2217

STY2218

STY2219

STY2220

STY2221

STY2222

STY2223

STY2224

STY2225

STY2226

STY2227

STY2228

STY2229

STY2230

STY2231

STY2232

STY2233

STY2234

STY2235

STY2236

STY2237

STY2238

STY2239

STY2240

STY2241

STY2242

STY2243

STY2244

STY2245

STY2246

STY2247

STY2248

STY2249

STY2250

STY2251

STY2252

STY2253

STY2254

STY2255

STY2256

STY2257

STY2258

STY2259

STY2260

STY2261

STY2262

STY2263

STY2264

STY2265

STY2266

STY2267

STY2268

STY2269

STY2270

STY2271

STY2272

STY2273

STY2274

STY2275

STY2276

STY2277

STY2278

STY2279

STY2280

STY2281

STY2282

STY2283

STY2284

STY2285

STY2286

STY2287

STY2288

STY2289

STY2290

STY2291

STY2292

STY2293

STY2294

STY2295

STY2296

STY2297

STY2298

STY2299

STY2300

STY2301

STY2302

STY2303

STY2304

STY2305

STY2306

STY2307

STY2308

STY2309

STY2310

STY2311

STY2312

STY2313

STY2314

STY2315

STY2316

STY2317

STY2318

STY2319

STY2320

STY2321

STY2322

STY2323

STY2324

STY2325

STY2326

STY2327

STY2328

STY2329

STY2330

STY2331

STY2332

STY2333

STY2334

STY2335

STY2336

STY2337

STY2338

STY2339

STY2340

STY2341

STY2342

STY2343

STY2344

STY2345

STY2346

STY2347

STY2348

STY2349

STY2350

STY2351

STY2352

STY2353

STY2354

STY2355

STY2356

STY2357

STY2358

STY2359

STY2360

STY2361

STY2362

STY2363

STY2364

STY2365

STY2366

STY2367

STY2368

STY2369

STY2370

STY2371

STY2372

STY2373

STY2374

STY2375

STY2376

STY2377

STY2378

STY2379

STY2380

STY2381

STY2382

STY2383

STY2384

STY2385

STY2386

STY2387

STY2388

STY2389

STY2390

STY2391

STY2392

STY2393

STY2394

STY2395

STY2396

STY2397

STY2398

STY2399

STY2400

STY2401

STY2402

STY2403

STY2404

STY2405

STY2406

STY2407

STY2408

STY2409

STY2410

STY2411

STY2412

STY2413

STY2414

STY2415

STY2416

STY2417

STY2418

STY2419

STY2420

STY2421

STY2422

STY2423

STY2424

STY2425

STY2426

STY2427

STY2428

STY2429

STY2430

STY2431

STY2432

STY2433

STY2434

STY2435

STY2436

STY2437

STY2438

STY2439

STY2440

STY2441

STY2442

STY2443

STY2444

STY2445

STY2446

STY2447

STY2448

STY2449

STY2450

STY2451

STY2452

STY2453

STY2454

STY2455

STY2456

STY2457

STY2458

STY2459

STY2460

STY2461

STY2462

STY2463

STY2464

STY2465

STY2466

STY2467

STY2468

STY2469

STY2470

STY2471

STY2472

STY2473

STY2474

STY2475

STY2476

STY2477

STY2478

STY2479

STY2480

STY2481

STY2482

STY2483

STY2484

STY2485

STY2486

STY2487

STY2488

STY2489

STY2490

STY2491

STY2492

STY2493

STY2494

STY2495

STY2496

STY2497

STY2498

STY2499

STY2500

STY2501

STY2502

STY2503

STY2504

STY2505

STY2506

STY2507

STY2508

STY2509

STY2510

STY2511

STY2512

STY2513

STY2514

STY2515

STY2516

STY2517

STY2518

STY2519

STY2520

STY2521

STY2522

STY2523

STY2524

STY2525

STY2526

STY2527

STY2528

STY2529

STY2530

STY2531

STY2532

STY2533

STY2534

STY2535

STY2536

STY2537

STY2538

STY2539

STY2540

STY2541

STY2542

STY2543

STY2544

STY2545

STY2546

STY2547

STY2548

STY2549

STY2550

STY2551

STY2552

STY2553

STY2554

STY2555

STY2556

STY2557

STY2558

STY2559

STY2560

STY2561

STY2562

STY2563

STY2564

STY2565

STY2566

STY2567

STY2568

STY2569

STY2570

STY2571

STY2572

STY2573

STY2574

STY2575

STY2576

STY2577

STY2578

STY2579

STY2580

STY2581

STY2582

STY2583

STY2584

STY2585

STY2586

STY2587

STY2588

STY2589

STY2590

STY2591

STY2592

STY2593

STY2594

STY2595

STY2596

STY2597

STY2598

STY2599

STY2600

STY2601

STY2602

STY2603

STY2604

STY2605

STY2606

STY2607

STY2608

STY2609

STY2610

STY2611

STY2612

STY2613

STY2614

STY2615

STY2616

STY2617

STY2618

STY2619

STY2620

STY2621

STY2622

STY2623

STY2624

STY2625

STY2626

STY2627

STY2628

STY2629

STY2630

STY2631

STY2632

STY2633

STY2634

STY2635

STY2636

STY2637

STY2638

STY2639

STY2640

STY2641

STY2642

STY2643

STY2644

STY2645

STY2646

STY2647

STY2648

STY2649

STY2650

STY2651

STY2652

STY2653

STY2654

STY2655

STY2656

STY2657

STY2658

STY2659

STY2660

STY2661

STY2662

STY2663

STY2664

STY2665

STY2666

STY2667

Once you have found this region have a look at some of the information that is available to you:-

Information to view:

Annotation

If you click on a particular feature you can view the annotation attached to it: select a CDS feature (or any other feature) and click on the Edit menu and select Edit Selected Feature. A window will appear containing all the annotation that is associated with that CDS. The format for this information is constrained by that which can be submitted to the EMBL database where it is stored within 'keys' and 'Qualifiers' see Appendix II.

Viewing amino acid or protein sequence

Click on the view menu and you will see various options for viewing the bases or amino acids of the feature you have selected, in two formats i.e. EMBL or FASTA. This can be very useful when using other programs that are not integrated into Artemis e.g. those available on the Web that require you to cut and paste sequence into them.

Plots/Graphs

Feature plots can be displayed by selecting a CDS feature then clicking 'View' and 'Show Feature Plots'. The window which appears shows plots predicting hydrophobicity, hydrophilicity and coiled-coil regions for the protein product of the selected CDS.

Load additional files

The results from Prosite searches run on the translation of each CDS should already be on display as pale-green boxes on the grey DNA lines. The results from the Pfam protein motif searches are not shown, but can be viewed by loading the appropriate file. Click on 'File' then 'Read an Entry' and select the file PF.tab. Each Pfam match will appear as a coloured blue feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'View Selection' or click 'Edit' then Edit Selected Features'. Please ask if you are unsure about Prosite and Pfam.

Further information on specific Prosite or Pfam entries can be found on the web at <http://ca.expasy.org/prosite> and <http://www.sanger.ac.uk/software/Pfam/tsearch.shtml>

In addition to looking at the fine detail of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding in to the display various plots showing different characteristics of the DNA. This information is generated dynamically by Artemis and although this is a relatively speedy exercise for a small region of DNA, on a whole genome view (we will move onto this later) this may take a little time so be patient.

To view the graphs:

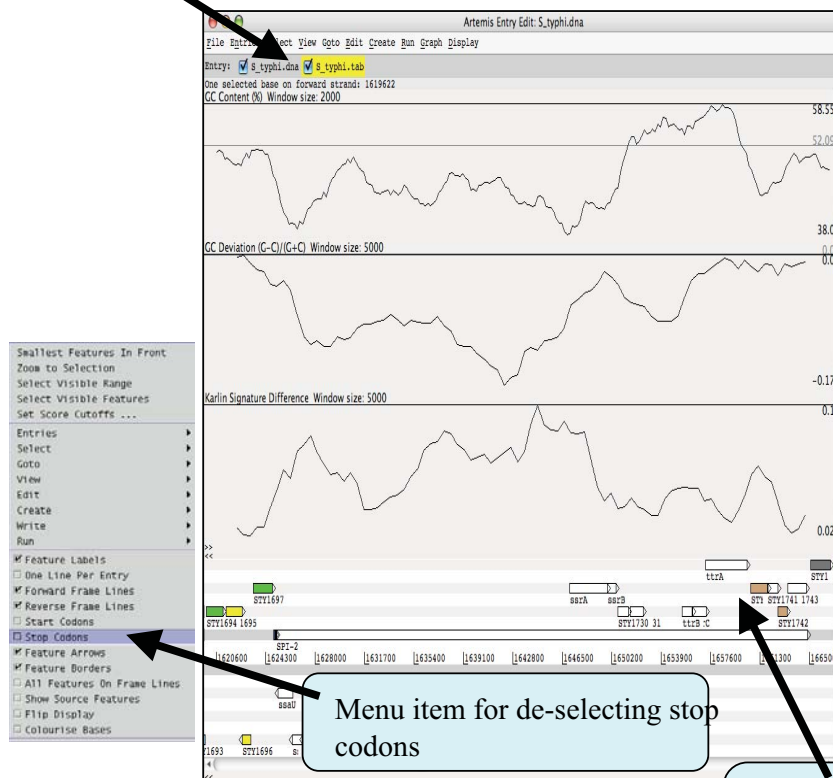
Click on the 'Graph' menu to see all those available. Perhaps some of the most useful plots are the 'GC Content (%)' (1) 'GC Deviation' (2) and 'Karlin signature plots' (3) as shown below. To adjust the smoothing of the graph you change the window size over which the points on the graph are calculated, using the sliders shown below. If you are not familiar with any of these please ask.



Notice how several of the plots show a marked deviation around the region you are currently looking at. To fully appreciate how anomalous this region is move the genome view by scrolling to the left and right of this region. Notice also that the nucleotide profile of SPI-2 appears to split the region into two segments.

As well as looking at the characteristics of small regions of the genome, it is possible to zoom out and look at the characteristics of the genome as a whole. To view the entire genome use the sliders indicated below. However, be careful zooming out quickly with all the features being displayed, as this may temporarily lock up the computer. To make this process faster, and clearer, switch off stop codons by clicking with the right mouse button in the main view panel. A menu will appear with an option to de-select stop codons (see below). If you have any problems ask a demonstrator.

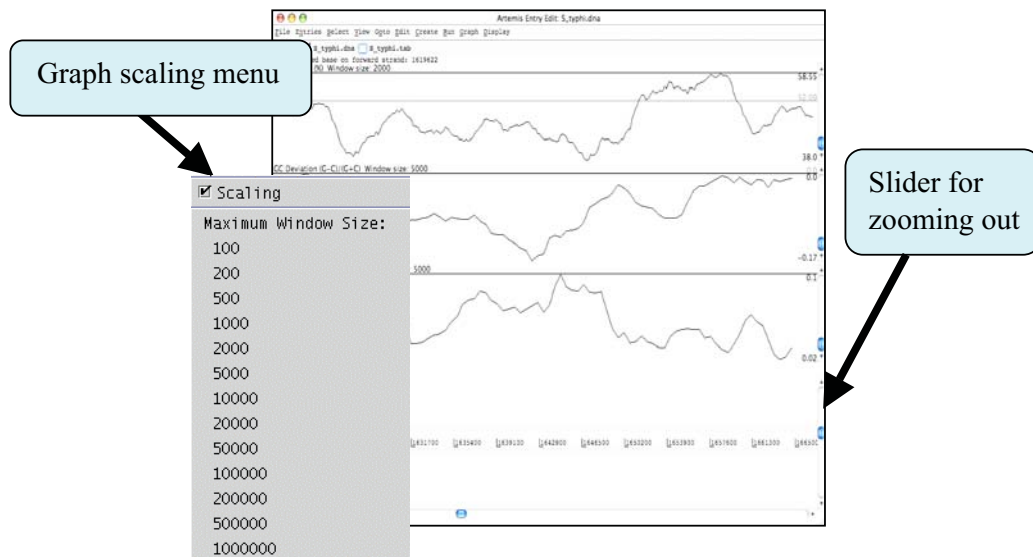
To de-select the annotation click here.



No stop codons shown on frame lines

You will also need to temporarily remove all of the annotated features from the Artemis display window. In fact if you leave them on, which you can, they would be too small to see when you zoomed out to display the entire genome. To remove the annotation click on the S_typhi.tab entry button on the grey entry line of the Artemis window shown above.

Your Artemis window should now look similar to the one shown below.



One final tip is to adjust the scaling for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window. A menu will appear with a series of values for the maximum window size (see above), select 20000. You should do this for each graph displayed.

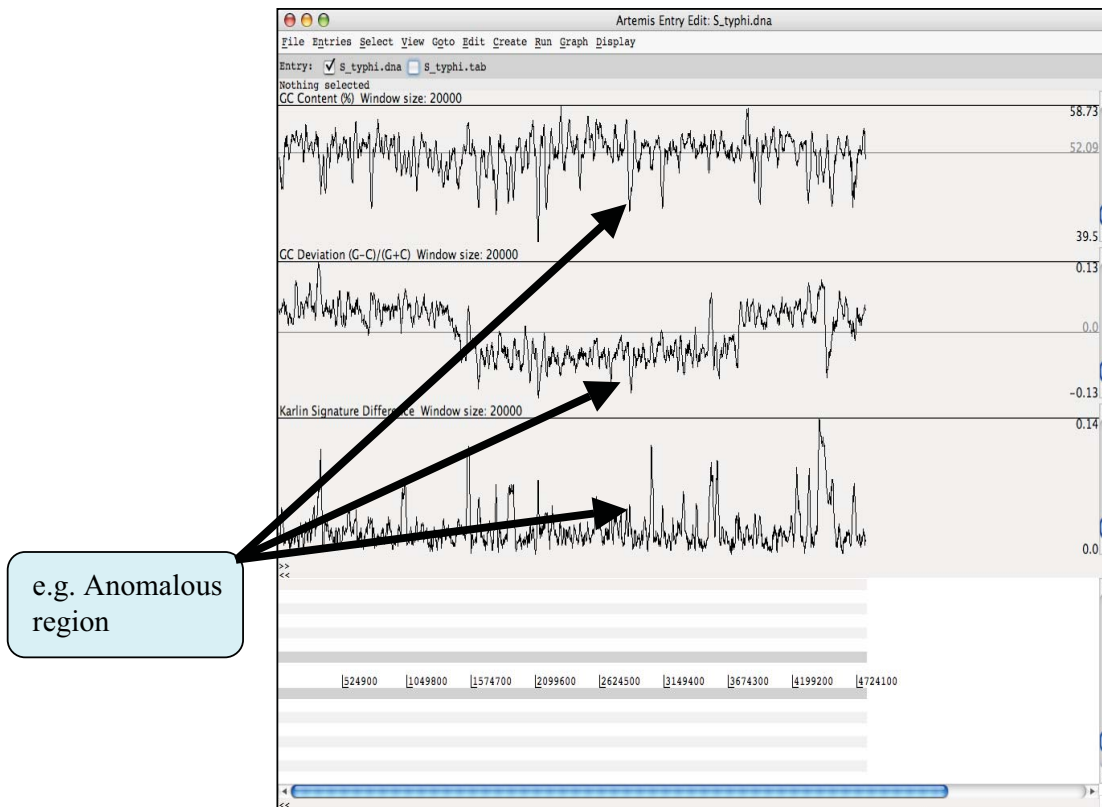
You are now ready to zoom out by dragging or clicking the slider indicated above. Once you have zoomed out fully to see the entire genome you will need to adjust the smoothing of the graphs using the vertical sliders as before to have a similar view to that shown below.



Click with the left mouse button in a graph window. A line and a number will appear. The number is the relative position within the genome (bps).

Click and drag to highlight a region on the main DNA line. Notice that the boundaries of this region should now be marked in the graph windows that you previously clicked in.

Artemis Exercise 1 Part III



The graphs can be used to look at other regions within the genome that stand out by having an atypical G+C content or Karlin signature (di-nucleotide frequency). You will see from the whole genome view of *S. Typhi* that there are many other examples of anomalous regions of DNA within a genome, many of which will have been laterally acquired. Since it has been shown in many bacteria that laterally acquired DNA carries important genes for virulence and lifestyle it is worth spending a little time exploring some of these regions. You should identify a region which you feel is interesting based on the graphs and zoom into look at the genes encoded within this region (one example is shown above).

So firstly zoom back into the genome to look in more detail at the first of these three peaks. Zoom into this position by first clicking on the DNA line at approximately the correct location. If you then use the vertical side slider to zoom back in, Artemis will go to the location you selected. Remember that in order to see the CDS features lying within this region you will need to turn the annotation (*S_typhi.tab*) entry back on.

It is also worth looking for other markers of lateral acquisition such as integrase genes and tRNA integration sites.

If you have time, and as a cautionary note, you should also go and have a look at Karlin and G+C plots for the region centred around genome position 4231500.

Module 2

Mapping Sequence Data

Introduction

Improvements in DNA sequencing technology have led to new opportunities for the studying organism at the genomic and transcriptomic levels. Applications include studies of the genomic variation within species and gene identification. In this module we will be using simulated data from *E.coli* genome, although all though the techniques you will learn are applicable all the technologies (e.g., Illumina Genome Analyzer, 454 GS FLX and ABI SOLiD). A single machine can produce around 20 Gigabases of sequence data from the Illumina machine comes as relatively short stretches of 35-100 base pairs (bp) of DNA- around 300 million of them. These individual sequences are called sequencing reads. The older capillary sequencing methods produces longer reads of ~500bp, but is much slower and more expensive.

One of the greatest challenges of sequencing a genome is determining how to arrange sequencing reads into chromosomes. This process of determining how the reads fit together by looking for the overlaps between them is called **genome assembly**. Capillary sequencing reads (~500bp) are considered long enough for the genome assembly. Genome assembly using sequence reads of >100bp is not possible in many cases due to the high frequency of repeats longer than the read length. Therefore, new sequencing technologies are mostly used where a **reference genome** already exists is called **resequencing**.

When resequencing, instead of assembling the reads to produce a new genome sequence and then comparing the two genomes sequences, we map the new sequence data to the reference genome. We can then identify **Single Nucleotide Polymorphism (SNPs)**, insertions and deletions (indels) and Copy Number Variants (CNVs) between two similar organisms.

In this module

The example used in this module is a set of 1000, 35bp simulated reads from throughout the *E.coli* genome and for finding SNPs is a set of simulated reads to cover 10,000 bases of the *E.coli* genome.

Module summary

- A. File formats
- B. Making index files
- C. Mapping the data
- D. Converting bowtie output to BAM
- E. Visualizing the mapped reads in Artemis
- F. Identifying the Single Nucleotide Polymorphism

A. File formats

You have the reference file of *E. coli* strain 536 (NC_008253.fna), a strain known to cause urinary tract infections. You also have two files of sequence reads simulated from that genome (a set of 1000, 35-bp reads). Look in both the reference file and read files.

Open up a terminal and navigate to Module_2_MSD directory, then type:

```
$ cd reads/  
$ head NC_008253.fna
```

Compare the reference file above to the files of sequencing reads:

```
$ more e_coli_1000.fastq
```

Each sequence read is represented by four lines.

1. @r1
2. CCGAACTGGATGTCTCATGGGATAAAAATCATCCG
3. +
4. EDCCCBAAAA@@@?>===<;9:99987776554

1. Sequence Header

3. Sequence/Quality Line Separator

4. Sequence Quality. There is one character for each nucleotide. The characters relate to a sequence quality score e.g. how likely is the nucleotide correct? '>' is higher quality than '6'. Sequence reads tend to have more error at the end than at the start.

2. The Read Sequence

```
File Edit View Terminal Tabs Help  
@r0  
GAACGATACCCACCCAACTATCGCCATTCCAGCAT  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r1  
CCGAACTGGATGTCTCATGGGATAAAAATCATCCG  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r4  
GCAGAGCAGTTGCTAGAAANNNTTGAAGAGGTT  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r6  
GGCAGTGATGCAACTGCCGTTATCAACAGNCT  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r7  
GCATATTGCCAATTTTCGCTTCGGGGATCAGGTTA  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r8  
GGTTCAGTTCAGTATACGCCTTATCCGGCCTACGG  
+  
@r1  
CCGAACTGGATGTCTCATGGGATAAAAATCATCCG  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r12  
AATCACAGGCGGTGAGCAGTAACGATAATTCGGCT  
+  
EDCCCBAAAA@@@?>===<;9:99987776554  
@r13
```


B. Making index files

Now we will make index from the *E.coli* genome sequence for the strain 536, which is known to cause urinary tract infections using short reading mapping program Bowtie (Langmead *et al.*, 2009).

Bowtie is an ultra fast, memory efficient short read aligner for aligning large sets of short DNA sequences (reads) to large genome. Bowtie indexes the genome with a burrows-wheeler index to keep its memory footprint small.

Navigate to module 3 reference genome directory, then type:

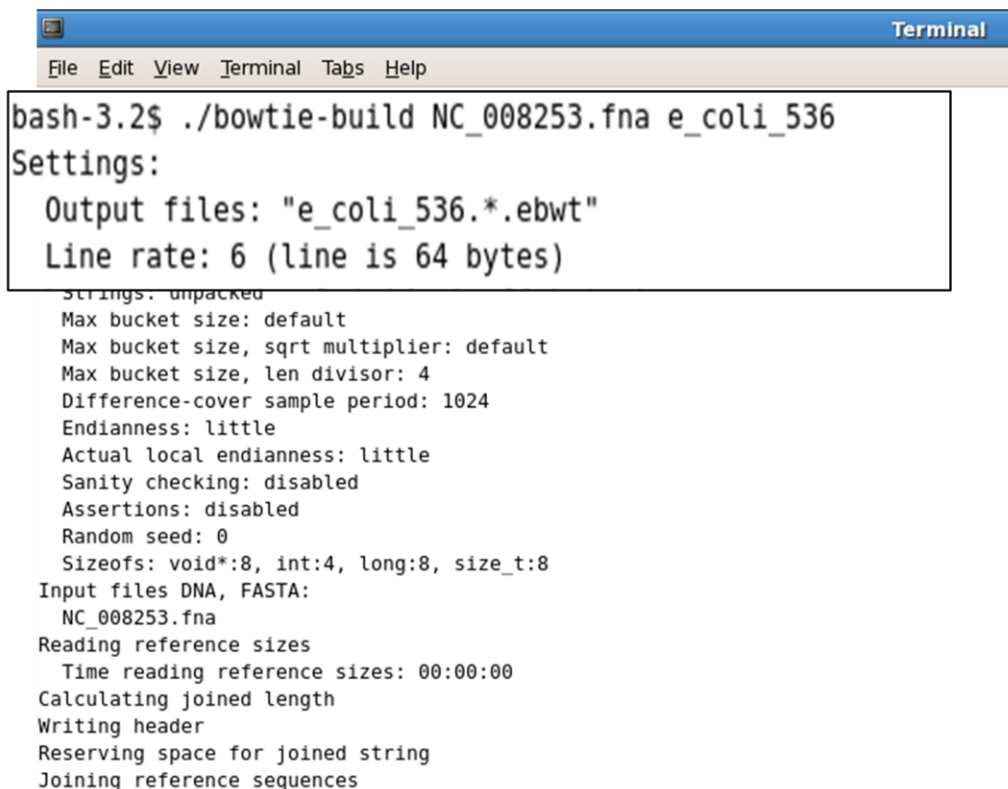
```
$/bowtie-build NC_008253.fna e_coli_536
```

The command should finish quickly, and print several lines of status messages. When command has completed, note that current directory contains 4 new files named *e_coli_536.1.ebwt*, *e_coli_536.2.ebwt*, *e_coli_536.3.ebwt*, *e_coli_536.4.ebwt*, *e_coli_536.1.rev.1.ebwt* and *e_coli_536.1.rev.2.ebwt*.

These files constitute the index. Move these files to the indexes subdirectory to install it. To test that the index is installed properly, on the terminal type:

```
$/bowtie -c e_coli_536 GCCTGAGCTATGAGAAAGCGCCACGCTTCC
```

If the index is installed properly, this command should print a single alignment and then exit.



```
Terminal
File Edit View Terminal Tabs Help

bash-3.2$ ./bowtie-build NC_008253.fna e_coli_536
Settings:
  Output files: "e_coli_536.*.ebwt"
  Line rate: 6 (line is 64 bytes)
Strings: unpacked
Max bucket size: default
Max bucket size, sqrt multiplier: default
Max bucket size, len divisor: 4
Difference-cover sample period: 1024
Endianness: little
Actual local endianness: little
Sanity checking: disabled
Assertions: disabled
Random seed: 0
Sizeofs: void*:8, int:4, long:8, size_t:8
Input files DNA, FASTA:
  NC_008253.fna
Reading reference sizes
  Time reading reference sizes: 00:00:00
Calculating joined length
Writing header
Reserving space for joined string
Joining reference sequences
```

C. Mapping the data

Now we will map reads to the *E.coli* reference genome using Bowtie (Langmead *et al.*, 2009).

On the command line of the aligned read appears in the left, type:

```
$ ./bowtie reference_genome/indexes/e_coli_536 reads/e_coli_1000.fq
```

The first argument to bowtie is the base name of the index for the genome to be searched. Second argument is the name of FASTQ files containing the reads. You will see bowtie print many lines of output. Each line is an alignment for the read.

1. Sequence Header

2. Reference Strand [Forward (+)/Reverse (-)]

3. Reference Sequence ID

4. 0 – based offset into the forward reference strand.

5. Read Sequence [reverse complimented if orientation is reverse (-)]

6. ASCII encode read qualities [reverse if orientation is reverse (-)]

7. Instances of the read Alignment .

8. Comma separated list of mismatch descriptors.

```
G>A,33:G>T
# reads processed: 1000
# reads with at least one reported alignment: 699 (69.90%)
# reads that failed to align: 301 (30.10%)
Reported 699 alignments to 1 output stream(s)
```

C. Mapping the data..continued...

Next, on the command line type :

```
$ ./bowtie -t reference_genome/indexes/e_coli_536 reads/e_coli_1000.fq e_coli.map
```

This run calculates the same alignments as the previous run, but the alignments are written to *e_coli.map*(the final argument) rather than to screen.

-t option instructs Bowtie to the print timing statistics.

The image shows a terminal window titled "Terminal" with a menu bar (File, Edit, View, Terminal, Tabs, Help). The command executed is `bash-3.2$./bowtie -t reference_genome/indexes/e_coli_536 reads/e_coli_1000.fq e_coli.map`. The output is as follows:

```
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
Seeded quality full-index search: 00:00:00
# reads processed: 1000
# reads with at least one reported alignment: 699 (69.90%)
# reads that failed to align: 301 (30.10%)
Reported 699 alignments to 1 output stream(s)
Time searching: 00:00:00
Overall time: 00:00:00
bash-3.2$
```

Three yellow callout boxes with arrows point to specific parts of the output:

- 1. Loading reference indexes time statistics.** Points to the first three lines of output: "Time loading forward index: 00:00:00", "Time loading mirror index: 00:00:00", and "Seeded quality full-index search: 00:00:00".
- 2. Mapped reads statistics.** Points to the middle three lines of output: "# reads processed: 1000", "# reads with at least one reported alignment: 699 (69.90%)", and "# reads that failed to align: 301 (30.10%)".
- 3. Alignment time statistics** Points to the last two lines of output: "Reported 699 alignments to 1 output stream(s)" and "Time searching: 00:00:00".

D. Converting Bowtie output to BAM

We are going to view the map reads in Artemis using Artemis BAM view. However the mapping result is not currently in BAM format. Therefore, we will convert Bowtie output to BAM using SAMtools (Li *et al.*, 2009).

SAM tools is a suit of tools for storing, manipulating, and analyzing alignment such as those output by Bowtie. SAM understands alignment in either of the two complementary formats: the human readable SAM format, or the binary BAM format.

Bowtie can output SAM (using `-S/--sam` option), and SAM can be converted to BAM using SAMtools .

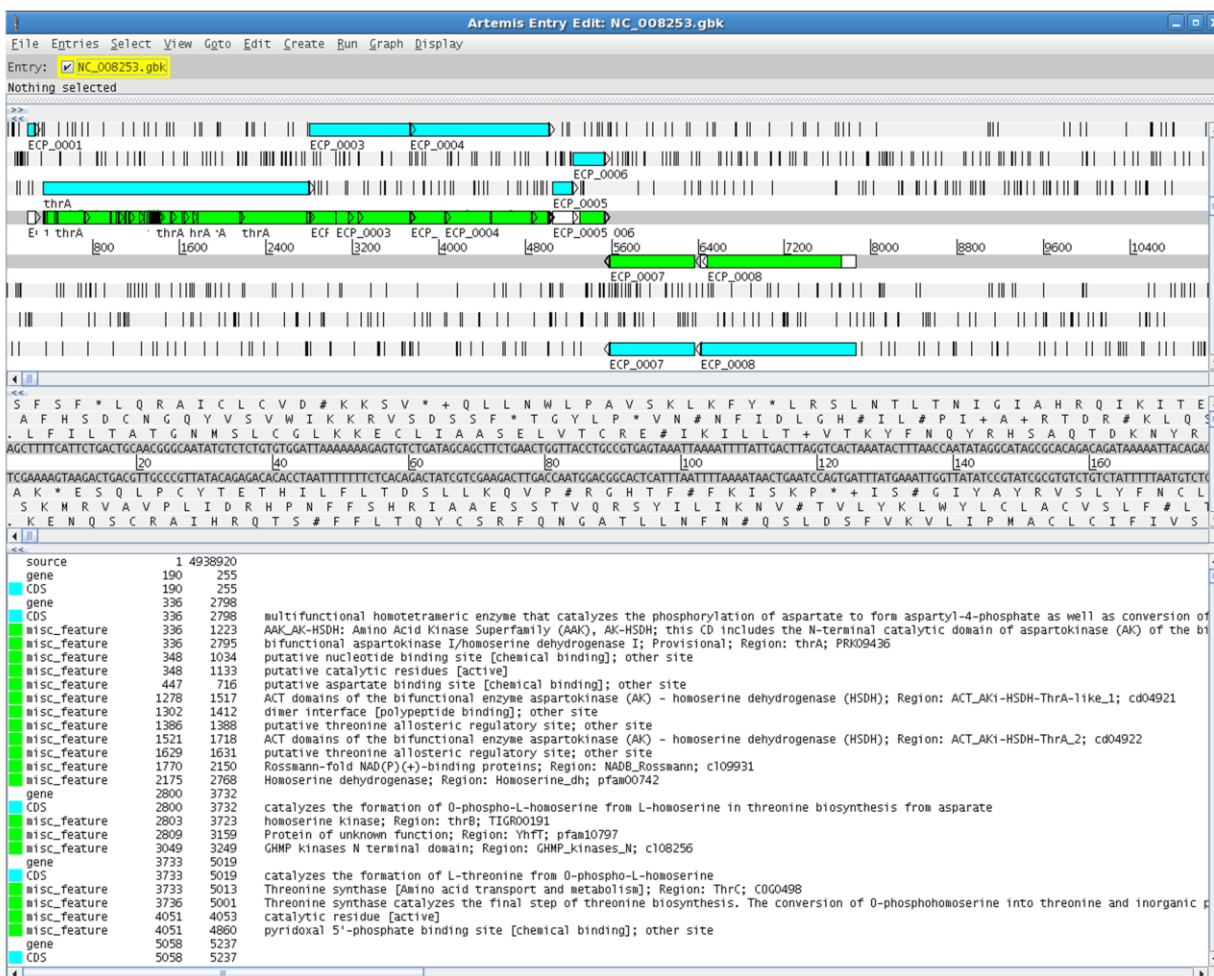
```
$ ./bowtie -S reference_genome/indexes/e_coli_536 reads/e_coli10000snp.fq  
e_coli_snp.sam  
  
$ ./samtools view -bt reference_genome/NC_008253.fna.fai e_coli_snp.sam >  
e_coli_snp.bam  
  
$ ./samtools sort e_coli_snp.bam e_coli_snp.sorted  
  
$ ./samtools index e_coli_snp.sorted.bam
```

E. Visualizing the mapped reads in Artemis

We will now examine the read mapping in Artemis using the BAMview feature.

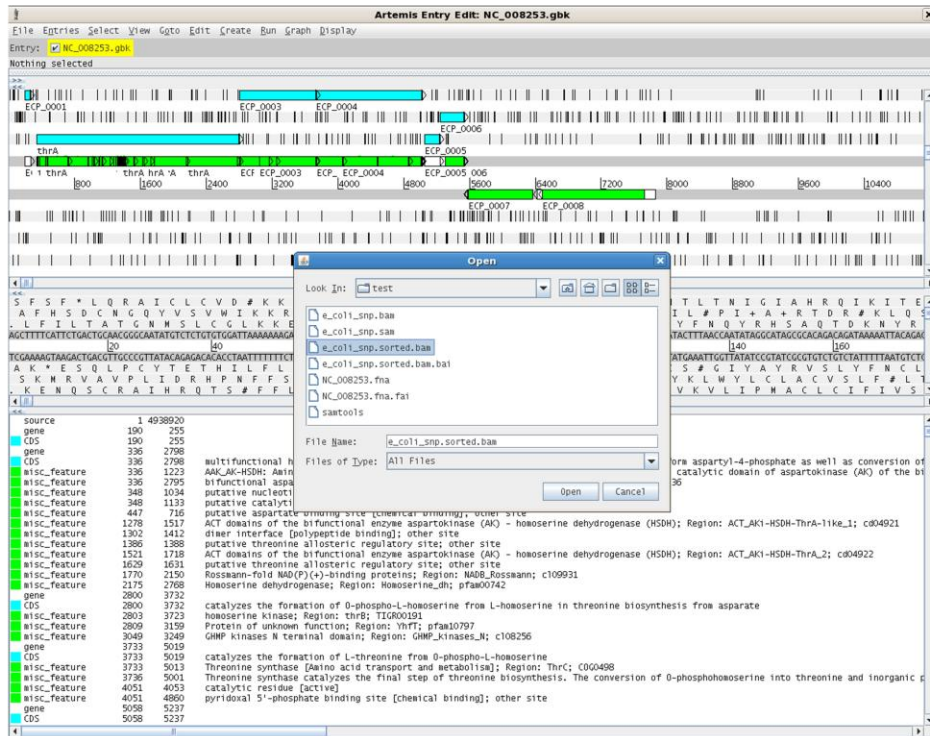
Open Artemis and load NC_008253.gbk from reference_genome directory. This contains exactly the same sequence as NC_008253.fna, but also has genome annotation so we can see the gene models.

You should see the Artemis window appear as in the screenshot below.

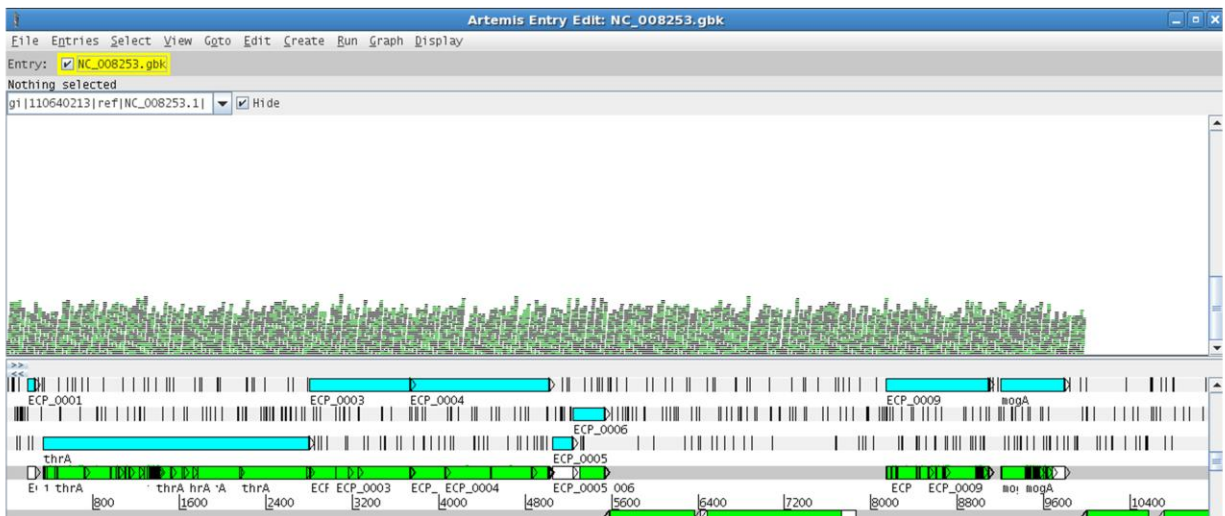


E. Visualizing the mapped reads in Artemis..continued...

From the Artemis file menu, select 'Read BAM', then locate the file *e_coli_snp.sorted.bam* from the module 3 data directory.

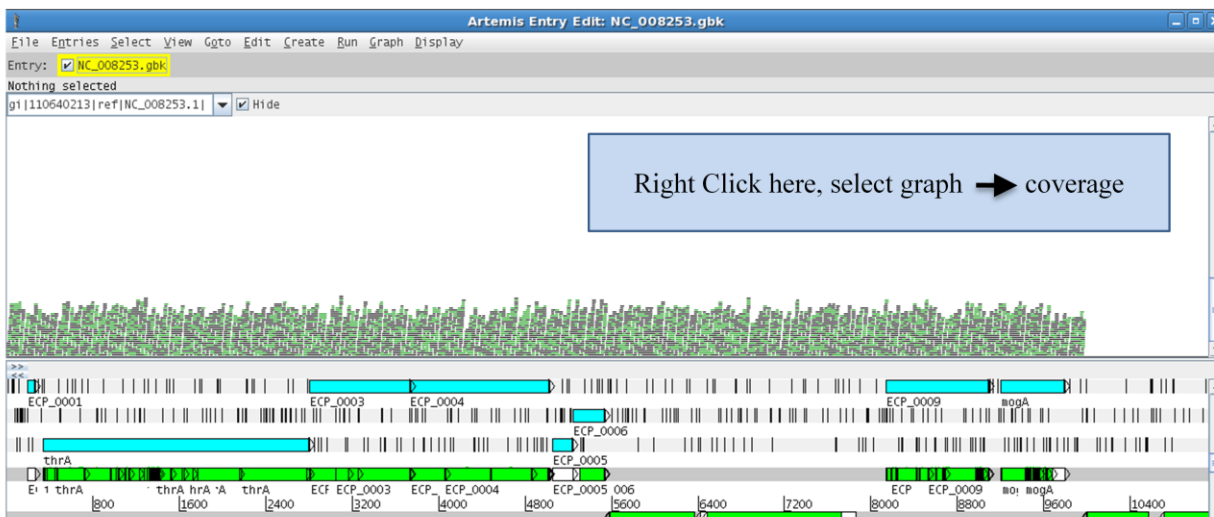


You should see the BAM window appear as in the screenshot below.



E. Visualizing the mapped reads in Artemis...continued...

Now we want to view the coverage of the reads mapped to the reference genome.



You should see the coverage plot in the BAM window appear, as in the screenshot below.



E. Visualizing the mapped reads in Artemis...continued...

Zoom in and right click on a read. Select 'show details of ...'

The screenshot shows the Artemis genome browser interface. The main window displays a genomic track with a read (r1309) mapped to a specific location. The read is highlighted in red. A right-click context menu is open, showing the option 'show details of ...'. The details panel for read r1309 is displayed, showing the following information:

Field	Value
Read Name	r1309
Coordinates	3727..3761
Length	35
Reference Name	gl 110640213 ref NC_008253.1
Inferred Size	0
Mapping Quality	255
Strand (read)	+
Cigar String	35M
Flags:	
Duplicate Read	no
Read Paired	no
Read Fails Vendor	no
Quality Check	no
Read Unmapped	no
Read Bases:	AACTAAATGAAACTCTACAATCTGAAAGATCACAA

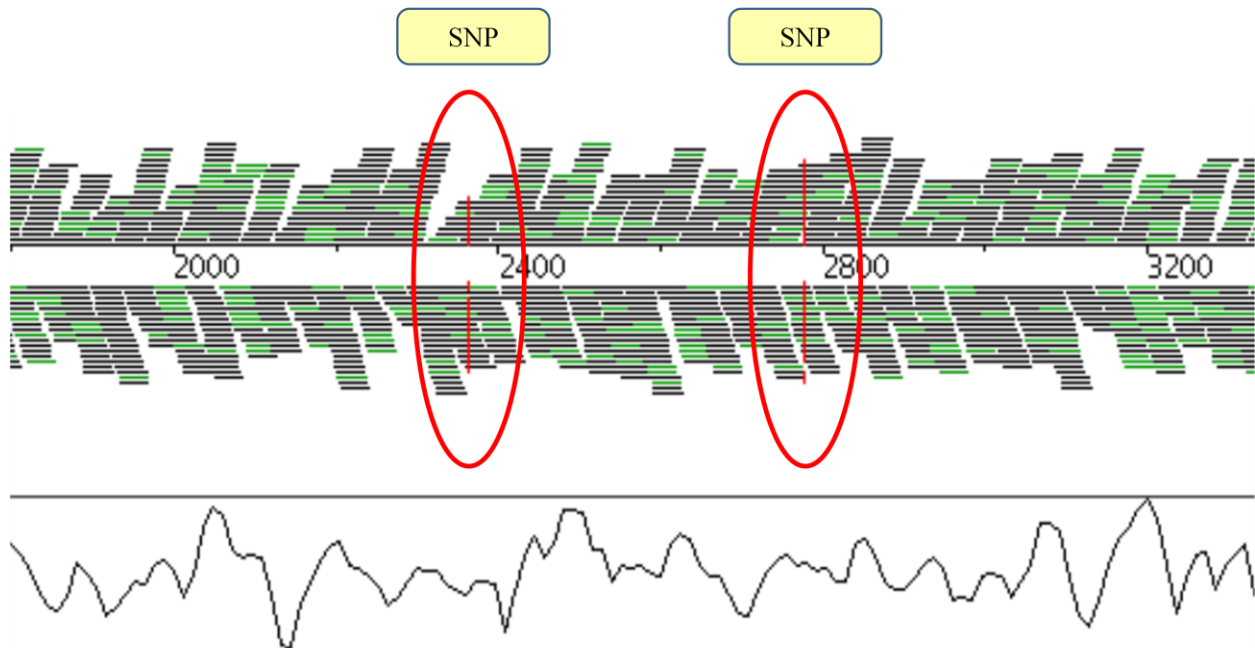
Notice the “mapping quality”. The maximum value for this is 255. the mapping quality depends on the accuracy of the sequence read and the number of mismatches with the reference. A value of zero means that the read mapped equally well to atleast one other location and therefore is not reliably mapped. The flags described the reads mate pair mapping.

F. Identifying Single Nucleotide Polymorphism

Some differences between the reference and the mapped reads are due to sequencing errors. On average, 1 in every 1000 bases in the reads is expected to be incorrect. However if reads mapping to the same location consistently have a base which is different from the reference, it is likely that this base is mutated in the other genome.

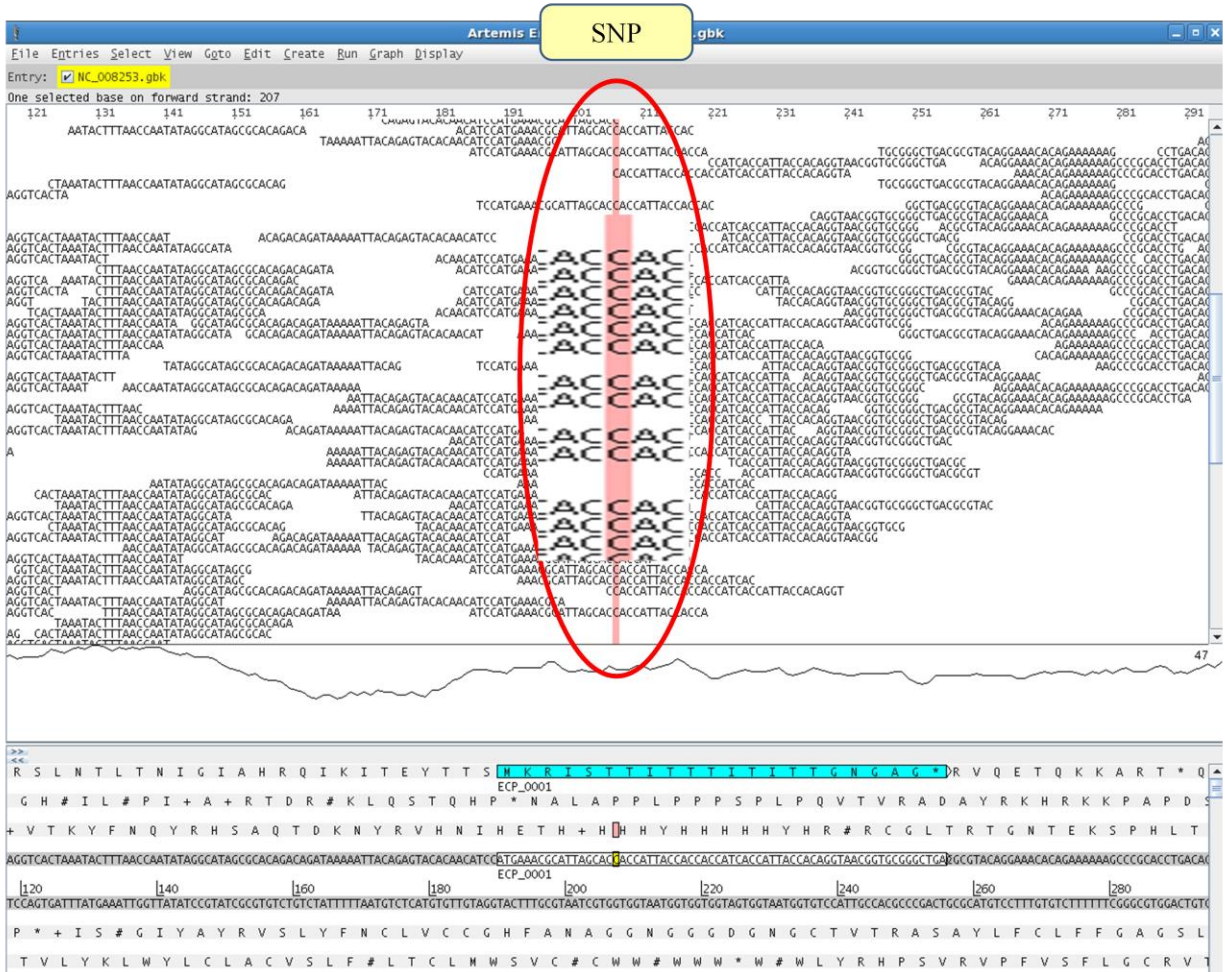


Red marks appear on the stacked reads highlighting every base in a read which does not match to the reference. If you zoom in you can distinguish SNPs as vertical red lines. Red lines, while the random sequencing errors or mismatches would be more disperse.



F. Identifying Single Nucleotide Polymorphism..continued...

Zoom in as far as you can go to the one of the vertical red line of the mismatch. You can see that while the sequence of the reads (in black) is generally the same as the reference sequence. The vertical red lines identify the consistent difference as in the case below.



What could be the consequence of the SNPs you have identified? Many SNPs will have no effect – Why is this?

The SNP examples are quite clear with this dataset, however this is not always the case. What if the read depth is very low? If there are only two reads mapping, the reference is T and both reads are C. is this enough evidence to say that the genomes are different?

Module 3 ACT

(using prokaryotic example)

Introduction

The Artemis Comparison Tool (ACT), also written by Kim Rutherford, was designed to extract the additional information that can only be gained by comparing the growing number of sequences from closely related organisms (Carver *et al.* 2005). ACT is based on Artemis, and so you will already be familiar with many of its core functions, and is essentially composed of three layers or windows. The top and bottom layers are mini Artemis windows (with their inherited functionality), showing the linear representations of the DNA sequences with their associated features. The middle window shows red and blue blocks, which span this middle layer and link conserved regions within the two sequences, in the forward and reverse orientation respectively. Consequently, if you were comparing two identical sequences in the same orientation you would see a solid red block extending over the length of the two sequences in this middle layer. If one of the sequences was reversed, and therefore present in the opposite orientation, there would be a blue 'hour glass' shape linking the two sequences. Unique regions in either of the sequences, such as insertions or deletions, would show up as breaks (white spaces) between the solid red or blue blocks.

In order to use ACT to investigate your own sequences of interest you will have to generate your own pairwise comparison files. Data used to draw the red or blue blocks that link conserved regions is generated by running pairwise BLASTN or TBLASTX comparisons of the sequences. ACT is written so that it will read the output of several different comparison file formats; these are outlined in Appendix II. Two of the formats can be generated using BLAST software freely downloadable from the NCBI, which can be loaded and run a PC or Mac. Whilst having a local copy BLAST to generate ACT comparison files can be very useful, it means that you are tied to a particular computer. Another way of generating comparison files for ACT is to use either of the WebACT web resource (Appendix V). Both of these sites allow you to cut and paste or upload your own sequences, and generate ACT readable BLASTN or TBLASTX comparison files.

Aims

The aim of this Module is for you to become familiar with the basic functions of ACT by using a series of worked examples. Examples will touch on exercises that were used in the previous Artemis Module, this is intentional. Hopefully, as well as introducing you to the basics of ACT these examples will also show you how ACT can be used for not only looking at genome evolution. You will also be shown or use a web resource, WebACT to generate your own comparison files and view them in ACT, depending on the time available.

1. Starting up the ACT software

Make sure you're in the **Module_3_ACT** directory.

Then type

act & [return]

A small start up window will appear.

Now let's load up a *S. Typhi* versus *Escherichia coli* comparison.

The files you will need for this exercise are: *S_typhi.dna*

S_typhi.dna_vs_EcK12.dna.crunch

EcK12.dna

1 Click 'File' then 'Open'

2 Use the File manager to drag and drop files or see 4

3

4, 5 & 6 Click and select appropriate files

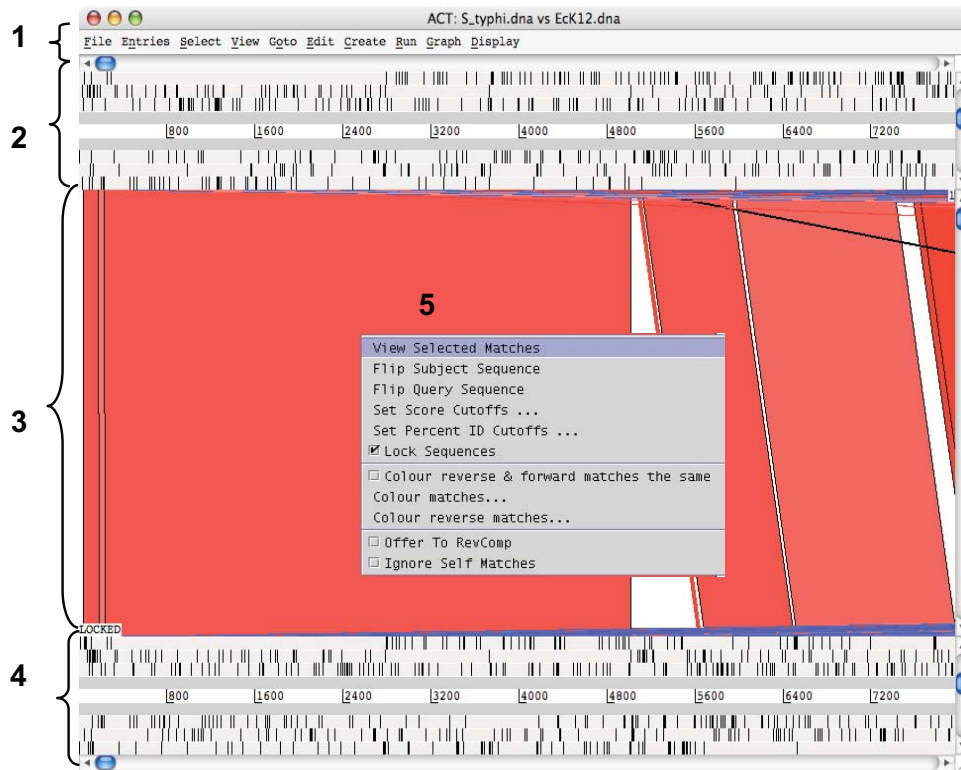
6 Click 'Apply' and wait.....

For comparing more than two genomes!

Comparison files end with '.crunch'.

2. The basics of ACT

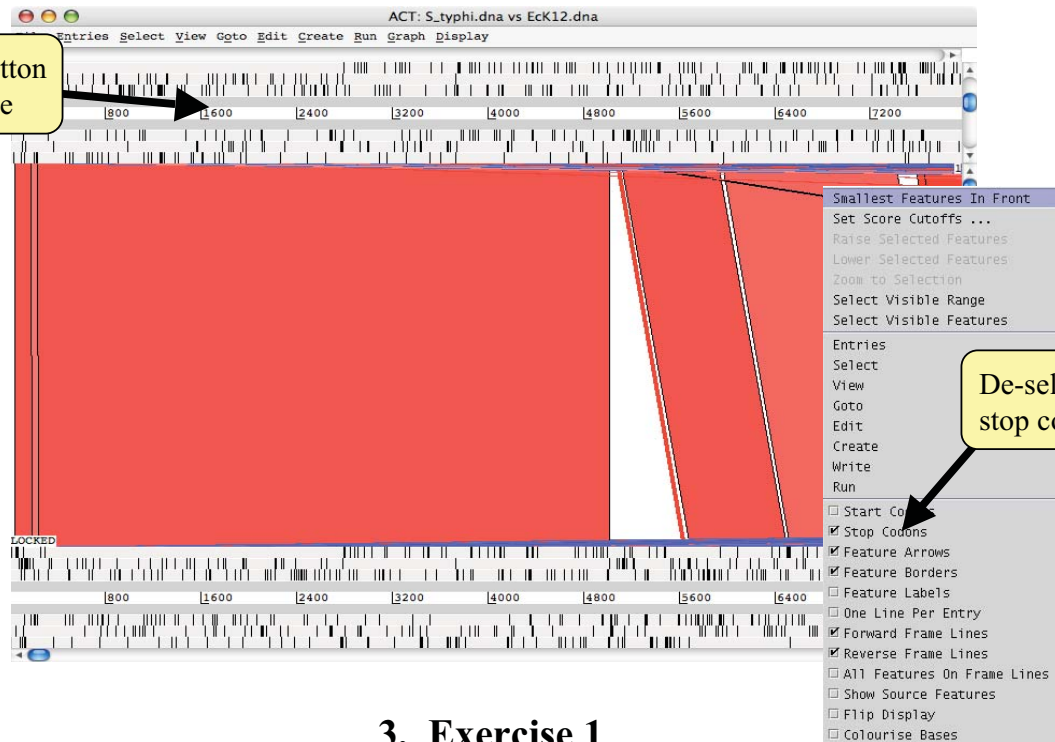
You should now have a window like this so let's see what's there.



1. Drop-down menus. These are mostly the same as in Artemis. The major difference you'll find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you selected earlier. It's a slightly compressed version of the Artemis main view panel. The panel retains the sliders for scrolling along the genome and for zooming in and out.
3. The Comparison View. This panel displays the regions of similarity between two sequences. Red blocks link similar regions of DNA with the intensity of red colour directly proportional to the level of similarity. Double clicking on a red block will centralise it. Blue blocks link regions that are inverted with respect to each other.
4. Artemis-style Sequence View panel for 'Sequence file 2' (Query Sequence).
5. Right button click in the Comparison View panel brings up this important ACT-specific menu which we will use later.

1

Right button
click here



2

De-select
stop codons

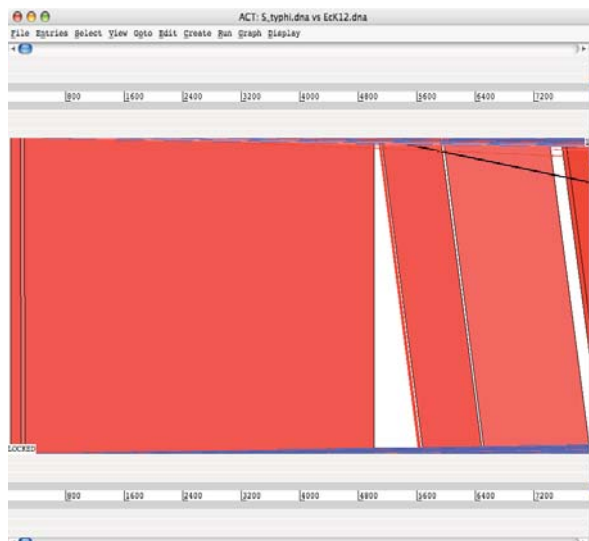
3. Exercise 1

Introduction & Aims

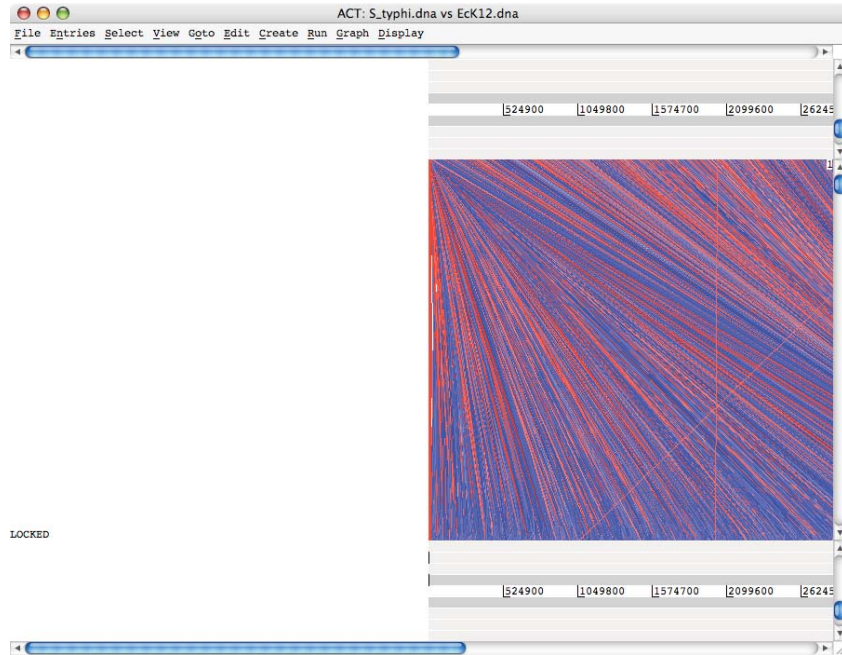
In this first exercise we are going to explore the basic features of ACT. Using the ACT session you have just opened we firstly are going to zoom outwards until we can see the entire *S. Typhi* genome compared against the entire *E. coli* K12 genome. As for the Artemis exercises we should turn off the stop codons to clear the view and speed up the process of zooming out.

The only difference between ACT and Artemis when applying changes to the sequence views is that in ACT you must click the right mouse button over the specific sequence that you wish to change, as shown above.

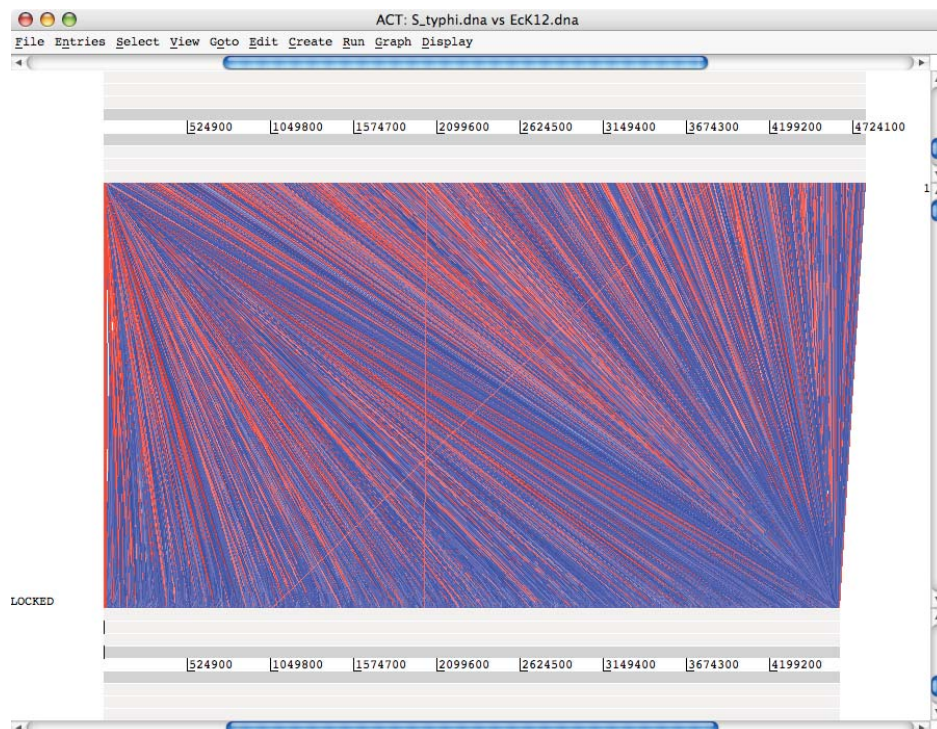
Now turn the stop codons off in the other sequence too. Your ACT window should look something like the one below:



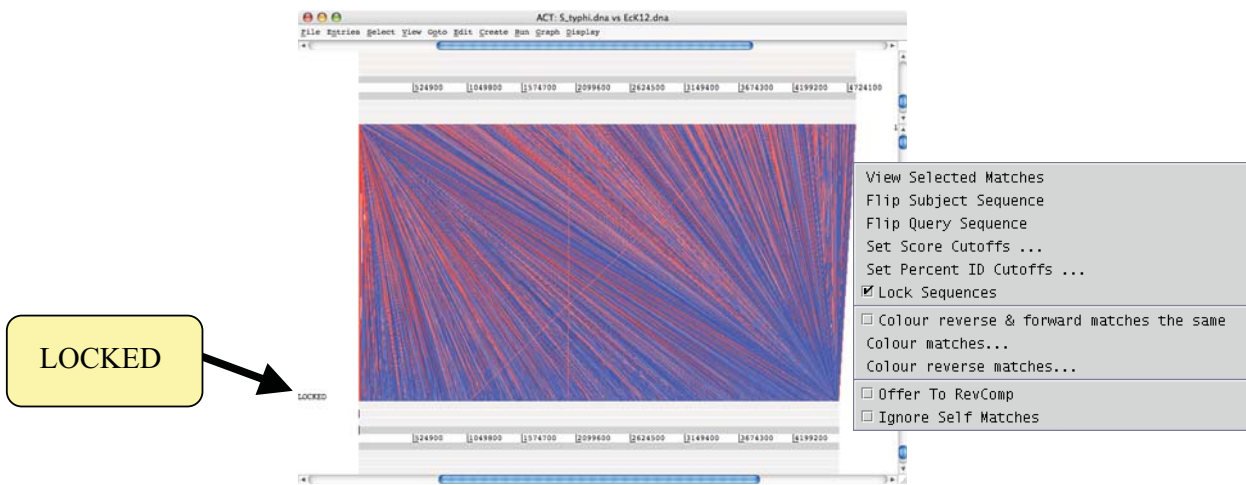
Use the vertical sliders to
zoom out. Drag or click
the slider downwards
from one of the genomes.
The other genome will
stay in synch.



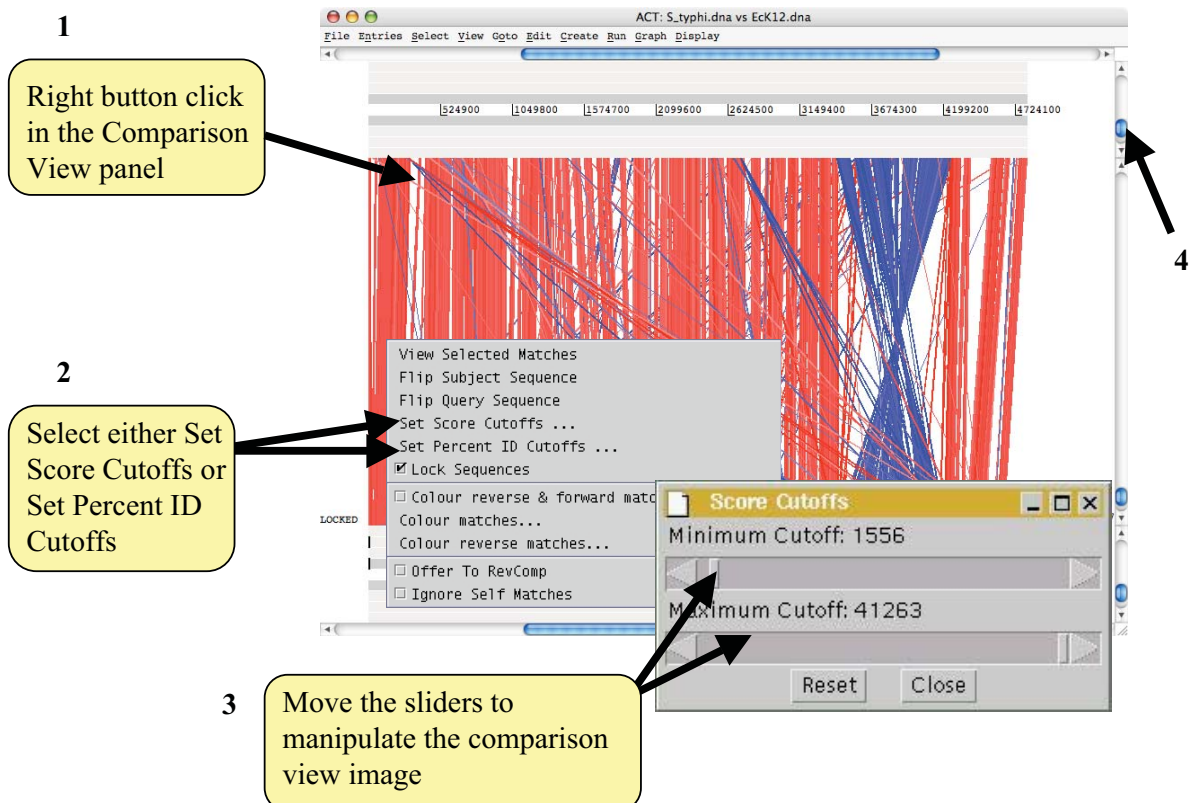
Once zoomed out your ACT window should look similar to the one shown above. If the genomes fall out of view to the right of the screen, use the horizontal sliders to scroll the image and bring the whole sequence into view, as shown below. You may have to play around with the level of zoom to get the whole genomes shown in the same screen as shown below.

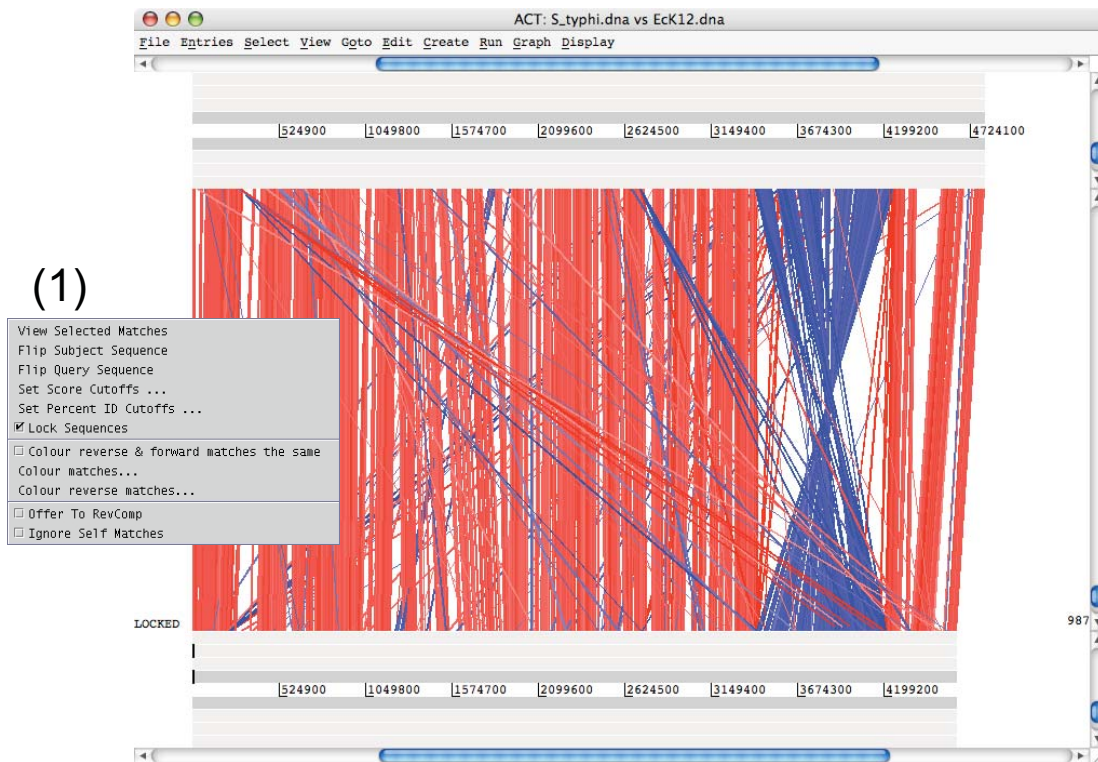


Notice that when you scroll along with either slide both genomes move together. This is because they are 'locked' together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that 'LOCKED' has disappeared from the comparison view panel and the genomes will now move independently



You can optimise your image by either removing 'low scoring' (or percentage ID) hits from view, as shown below 1-3 or by using the slider on the the comparison view panel (4). The slider allows you to filter the regions of similarity based on the length of sequence over which the similarity occurs, sometimes described as the "footprint".





4. SPI-2 in ACT

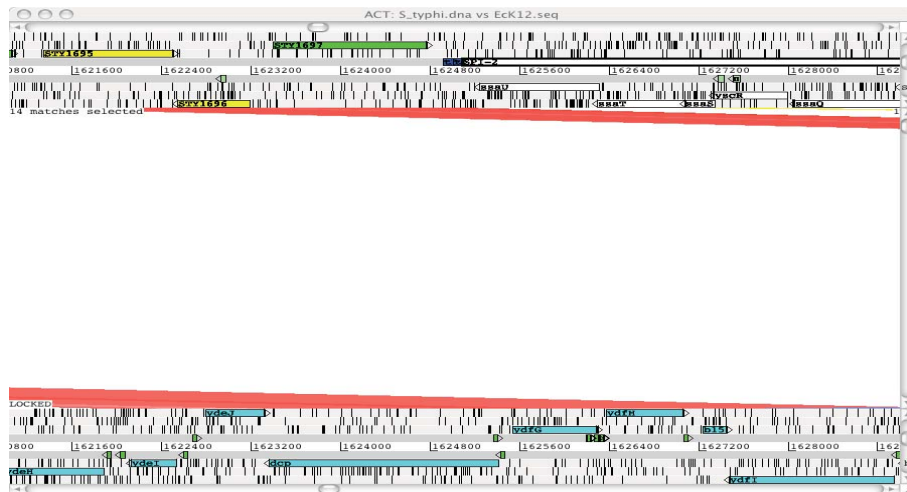
You should now be looking at an image of both the *E. coli* and *S. Typhi* genomes similar to that shown above. It is apparent that there is a backbone sequence shared with *E. coli* K12. Into this various chunks of DNA, specific the *S. Typhi* (with respect to *E. coli* K12) have been inserted.

•Key functions you should now try out in ACT

1. Double click (left mouse button) on the red boxes to centralise them.
2. Zoom right in to view the base pairs and amino acids of each sequence.
3. Also try using some of the other Artemis features e.g. graphs etc.
4. Find an inversion in one genome relative to the other then flip one of the sequences. To do this use the middle window menu shown above (1).

Load into the top sequence (*S. Typhi*) the annotation file '*S_typhi.tab*'. You will need to use the 'File' menu and then select the correct genome sequence ('*S_typhi.dna*') before you can read in the appropriate annotation file or entry. The *E. coli* K12 annotation file (*EcK12.tab*) is also in the directory for this module so you can load this in too.

Now we are going to go to the region of the *S. Typhi* genome that we looked at earlier, SPI-2. Either by using the sliders or the 'navigator' find the SPI-2 region of the *S. Typhi* genome. If you are unsure of where this region is or how to get to it refer back to the earlier Artemis Module as these functions are the same in ACT.

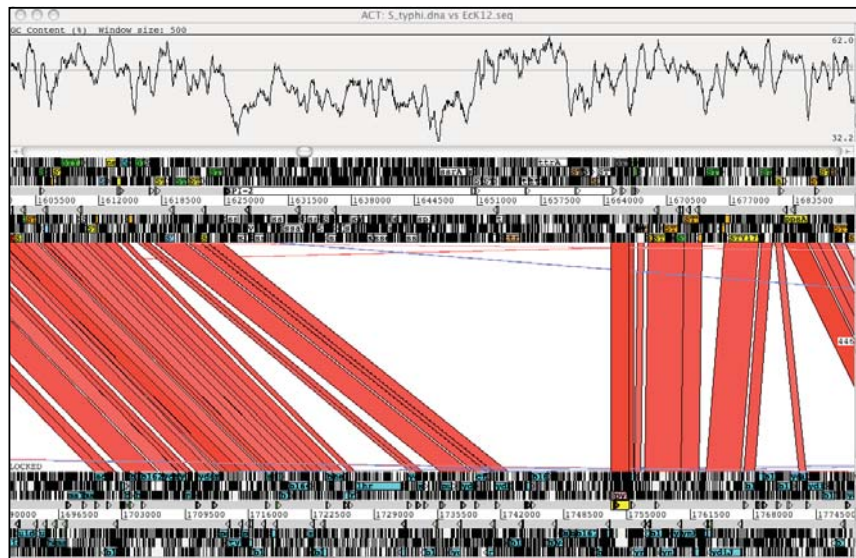


Once you have found SPI-2 in the *S. Typhi* genome double click (left mouse button) on the red boxes and use the sliders to centralise the sequences. Once you have done this it should look similar to the view below. This region is a clear insertion in the *S. Typhi* CT18 genome (see below) and has many of the characteristics of a classical pathogenicity island (PAI):

Jorg Hackers' Definition of a PAI

- Carry mobility functions e.g. integrases
- Inserted next to tRNA
- Anomalous G+C (add G+C plots see below)
- Carry virulence genes
- High number of pseudogenes
- In pathogens absent from non-pathogens

Take this opportunity to explore this region more fully and look for some of these features.



References

Langemead *et al.* (2009) *Genome Biology* 10:R25

Ultrafast and memory efficient alignment of short DNA sequences to the human genome.

Li *et al.* (2009). *Bioinformatics*, 25:1754-60

Fast and accurate short read alignment with Burrows-Wheeler Transform.

Carver T.J. *et al.* (2010). *Bioinformatics*, (doi:10.1093/bioinformatics/btq010)

Bam View: viewing mapped read alignment data in context of the reference sequence.

Carver T.J. *et al.* (2005) *Bioinformatics*, 21:3422-3

ACT: the Artemis Comparison Tool.

Berriman, M., and K. Rutherford (2003) *Brief Bioinform*, 4 (2) 124-132

Viewing and annotating sequence data with Artemis.

Rutherford *et al.* (2000) *Bioinformatics* 16 (10) 944-945

Artemis: sequence visualization and annotation.

Teaching manual of open door workshop (2010) *Welcome Trust Sanger Institute*

Working with Pathogen Genomes

Abbot, J. C. *et al.* (2005) *Bioinformatics* 21(18)3665-3666

WebACT – an online companion for the Artemis Comparison Tool.

Appendices

PROCESSES

^c <ctrl>-c kills (definitely stops) current job
^z <ctrl>-z suspends the current job. This can either be moved to the background or resumed in the foreground by using **bg** or **fg**

bg moves the current process to the background
fg moves a process to the foreground. (If there is more than one suspended job, use **jobs** to decide which you want to **fg**)

fg 2 moves process number 2, as listed by **jobs**, to the foreground

jobs lists background and suspended processes (created with **bg** or **^z**)
jobs -l ("el" not one) includes the pid (process id number)

ps lists all your processes

kill stops a process (use **ps** or **jobs** to find your processes)
kill 2986
kills off the process with pid 2986

MISCELLANEOUS

finger tells you who is logged on (see also **w**)

w shows information about logged in users

who produces similar result (see **finger**)

tar create (or extract) a tarball from (to) a list of files
tar -cvf tarball.tar subdir/*
tar -xvf tarball.tar
the option **-z** compacts the files by **gzip**

wc word count
wc long.file
prints the number of lines, words and characters in *long.file*. Options include **-l** to count lines only, and **-c** to count characters only

ln create a link or an alias for a file
ln -s subdir/orig.file alias.file

history displays last several commands used
!! re-executes the last command
!51 executes command 51 in the history list use also **<up>** - and **<down>** - arrows to navigate in the history

date displays current date and time

passwd invokes a password changing program

exit leaves the current shell (same as **^d** or **<ctrl>-d**) usually = logout

GRAPHIC DISPLAY

To display graphics, most Unix require the configuration of the X-Window server.

Commands on your local computer:

xhost set the list of allowed X-Window clients
xhost +
The "+" allows any remote computer to display on your local display

ifconfig gives information about the network configuration (e.g., the current IP_address, usually similar to 123.145.167.189)

Commands on the remote computer:

setenv set up an environment variable (tc-shell)
setenv DISPLAY IP_address:0.0
required to tell the remote computer where it should display its graphics

xclock starts a graphic clock (e.g., used to test the X-Window server or to get the current time... ;-)

This document was originally written and designed by Aoife McLysaght and Andrew Lloyd© from the Irish EMBnet node, and modified by Laurent Falquet from the Swiss EMBnet node and distributed by the Publications Committee of EMBnet.

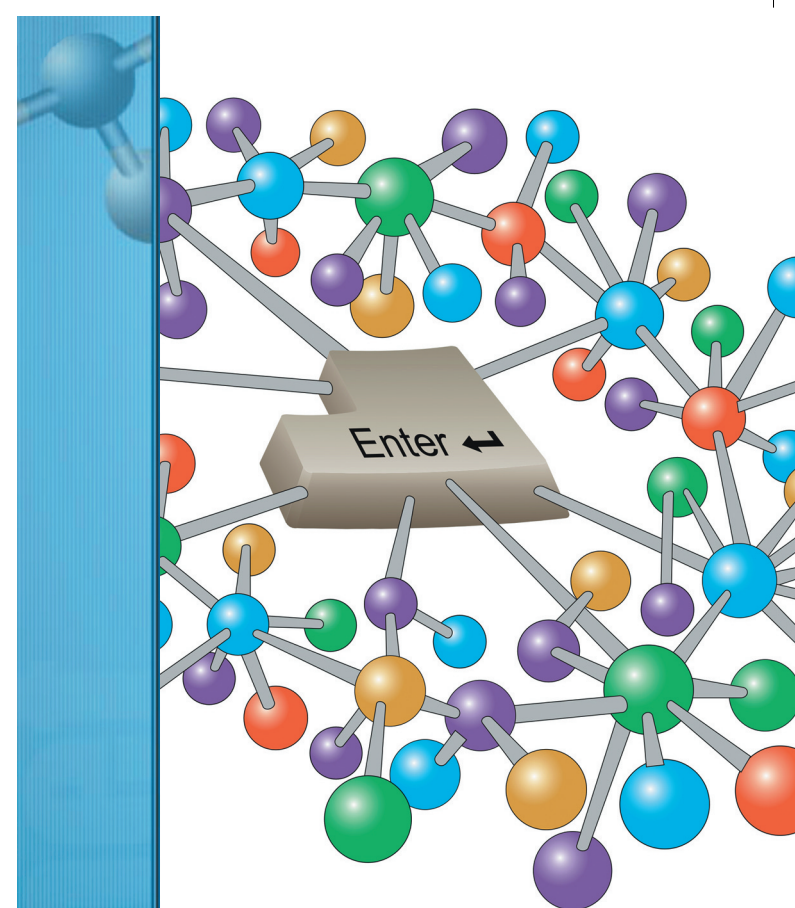
EMBnet - European Molecular Biology network - is a network of bioinformatics support centres situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.

Further information about UNIX is available from your national node. You can find contact information about your national node from the EMBnet web site:

<http://www.embnet.org/>

If you have found this publication useful, please let us know.
If you have ideas for similar documents we'd like to hear from you: emb-pr@embnet.org

A Quick Guide To UNIX
Revised edition 2003



A Quick Guide UNIX

EMBnet

A Quick Guide To UNIX

This is an introduction to the UNIX operating system. Unix may seem idiosyncratic, even impenetrable, to begin with but it has the virtue of minimising the number of keystrokes and so speeding up your access to the computer.

The commands listed here are common to different operating systems and shells. They include some of the most useful and frequently used commands in UNIX. The power and utility of most UNIX commands can be enhanced with switches or options preceded by a “-” sign.

More information on the options, the effects and how to use the commands is available by using the **man** command:

man gives manual information on a topic
man grep
displays the manual page about grep
apropos lists all the man(ual) entries relating to a topic
(same as **man -k**)
apropos print

Another useful source of information is the on-line EMBnet tutorial which includes a page on UNIX

<http://www.dk.embnet.org/Embnetut/Univsl/unixcmds.html>
or equally

<http://www.uk.embnet.org/Embnetut/Univsl/unixcmds.html>

The general format of this document is that anything in **bold** is a command you can enter. Anything in *italic* is a fake file or directory name you must change according to yours. Anything preceded by a hyphen “-” is an option which will modify the effects of a command. A general description of each command is followed by one or several examples of its use.

FILES

ls lists files in a directory
ls -alF
lists **-a** all files in **-l** long format **-F** identifies directories **/**, executable files ***** and symbolic links **@**, in the current directory
cat concatenates and displays files
cat my.file
displays *my.file* on the screen

chmod modifies the read (**r**), write and delete (**w**), and execute (**x**) permissions of specified files and the search permissions of specified directories. The permission can be set for user (**u**), group (**g**) or other (**o**)
chmod go-w my.file
stops (**-**) anyone else (**go**) changing or deleting (**w**) *my.file*
chmod g+rx my.file
allows (**+**) anyone of my group (**g**) reading, changing, deleting or executing (**rx**) *my.file*
cp copies files
cp orig.file copy.file
cp orig.file subdir/new.file
copies *orig.file* to *new.file* in *subdir* directory
cp subdir/orig.file .
copies *orig.file* from *subdir* to the current directory (**.**) without changing its name
mv moves/renames a file (or directory)
mv oldname newname
mv my.file subdir/my.file
a move (**mv**) is equivalent to a copy (**cp**) followed by a remove (**rm**)
rm removes/deletes a file.
rm oldfile
rm -i *.file
option **-i** (interactive) advised if wildcards (*****) in use
diff compares two files and prints how they differ
diff file1 file2
prints differences to screen options include **-b** to ignore differences in blank space, and **-i** to ignore case
find searches the directory tree for a file
find . -name lostfile -print
will search your current directory (**.**) (and any subdirectories) for *lostfile*
grep searches a file for a string
grep word my.file
grep "two words" my.file
options include **-i** to ignore case and **-n** to print line numbers
vi simple screen oriented text editor

pico simple display oriented text editor
pico myfile.txt
head prints the first few (default = 10) lines of a file
head oddfile
head -20 oddfile
displays first twenty lines of *oddfile*
tail displays last few lines of a file (see head)
more displays a file one screenful at a time
more longfile
hit **<spacebar>** to see the next screen
Note: some people prefer **less**

OUTPUT REDIRECTION

> redirects output of a command to a file
diff file1 file2 > new.file
puts differences into *new.file*
cat one.file two.file > both.file
writes the output of the cat command into *both.file* (overwrites *both.file*)
>> appends a file to the bottom of another
cat three.file >> both.file
appends *three.file* to the bottom of *both.file*
| “pipe” - uses the output of the first command as the input of the second
grep string my.file | wc -l
finds how many lines on which “*string*” occurs (see **grep** and **wc**)

DIRECTORIES

cd changes current directory
cd /etc
go to */etc* directory
cd ..
go up one level in directory tree
cd ../subdir2
go “sideways” to *subdir2*
mkdir creates a new subdirectory
mkdir subdir
rmdir removes a directory - you must delete all the files in it first
rmdir subdir
pwd print working directory, tells your current location (path)

EMBOSS

A Quick Guide

European Molecular Biology Open Software Suite

History

Since 1988, the sequence analysis package EGCG has provided extensions to the market leading commercial sequence analysis package GCG. EGCG development was a collaboration of groups within EMBnet and elsewhere.

That project has reached the limits of what we can achieve using the GCG package. Specifically, it is no longer possible to distribute academic software source code which uses the GCG libraries and has become difficult even to distribute binaries.

As a result, the former EGCG developers have been designing a totally new generation of academic sequence analysis software. This has resulted in the present EMBOSS project.

EMBOSS is a new suite of freely available programs and libraries for sequence analysis. It incorporates and integrates a range of currently available public packages and tools into a general, publicly available, suite specially developed for the needs of the Sanger Centre and the EMBnet user community.

Licensing

The EMBOSS core application suite is licensed under the General Public License (GPL) allowing free copying, modification and distribution of the package.

The EMBOSS Libraries are licensed under the the Library General Public License.

Associated packages may be licensed under different terms, all of which permit free redistribution of the software.

Obtaining EMBOSS

EMBOSS and the associated packages can be obtained via FTP from the Sanger Centre, UK at <ftp.sanger.ac.uk/pub/EMBOSS>

EMBOSS home page

<http://www.sanger.ac.uk/Software/EMBOSS>

Running EMBOSS

All EMBOSS programs are designed to be run from the command line. Each program has a specific description file (ACD file) that describes the input and output parameters. All the parameters can be specified on the command line, allowing modular integration into graphical interfaces.

To run an EMBOSS program, just type its name. Your system administrator should ensure that the programs are available in your \$PATH.

The Uniform Sequence Address (USA)

The USA is a method of specifying the location of a sequence and its format. The general form is:

Format::database:sequencename

eg. **embl::em:scact**

EMBOSS is normally very good at identifying sequence *formats* automatically but occasionally needs a hint. *Database* will be one of the databases already set up at your site. The command **% showdb**

lists the databases available on your system.

The *sequencename* can be either its name, accession number, the filename in which the sequence is found, or the sequence itself if **asis::** format is specified. If you are taking one sequence from a multiple sequence file, put the sequence number in braces after the filename, eg:

allmyseqs.fasta{32}

EMBOSS programs

You can obtain a list of EMBOSS programs with the command **wossname**. Useful qualifiers for **wossname** are :

-alphabet	List all programs in alphabetical order
-auto	List all programs without asking for a keyword.

% wossname -alphabet -auto

will list all the available emboss programs with a short description of the function of each program

EMBOSS will by default only prompt you for the minimal input it needs to run the program. The default behaviour can be changed using command line qualifiers.

Important qualifiers

The behaviour of EMBOSS programs can be modified by using a large number of qualifiers. This is a list of the more useful ones.

-help	Prints a summary of the options the program can take. With -verbose it gives a more detailed list.
-options	Prompt the user for the optional parameters
-auto	Accept all the default settings and run without prompting the user.
-sask	Ask for the start, end and reverse of the sequence input
-stdout	Print output to stdout (the screen) instead of to a file.
-filter	Take input from stdin (keyboard) and output to stdout

What -help tells you

The **-help** option lists the inputs to the program along with the input type (sequence, integer etc). There are additional qualifiers associated with many types. **-verbose** will list all the additional qualifiers related to the input types for the program.

The qualifiers are listed in three sections:

Mandatory Qualifiers

These are the minimum inputs the program needs to run. Some of these have default values which can be selected using **-auto**

Optional Qualifiers

These are qualifiers for which you will be prompted if you use the **-option** qualifier. All these qualifiers have default values.

Advanced Qualifiers

You will never be prompted for these. If you wish to use them you must specify them on the command line.

EMBOSS parameter types

Type	Allowed values
bool	yes: -param no: -noparam
integer	Whole numbers -param=5
float	decimal numbers -param=23.9
range	sequence ranges. eg. -param=1-12,35-99
regex	a regular expression pattern
string	ordinary text. -param='text with *'
infile	path of a file
matrix	integer scoring matrix for alignments
matrixf	floating point scoring matrix
codon	codon usage table
sequence	Uniform sequence address (USA) for the sequence or set of sequences.
segset	
segall	
features	Feature table
list	list of options
selection	selection list of options
outfile	path to a file for nonsequence output
segout	output sequence USA
segoutset	multiple sequence file for output
segoutall	multiple or single sequence output files
featout	output feature table
graph	output device for graphics images
xygraph	output device for XY graphs

See the descriptions below for many of these.

Associated qualifiers: sequence, seqset, seqall

-sbegin	integer	first base used [start]
-send	integer	last base used [end]
-sreverse	bool	reverse sequence [N]
-sask	bool	prompt for begin/end/reverse [N]
-snucleotide	bool	Sequence is nucleotide [N]
-sprotein	bool	Sequence is protein [N]
-slower	bool	Make sequence lowercase[N]
-supper	bool	Make sequence uppercase[N]
-sformat	string	input sequence format
-sopenfile	string	input filename
-sdbname	string	database name
-sentry	string	entry name/accession number
-ufo	string	Feature table (UFO)
-fformat	string	features format

Associated qualifiers: seqout, seqoutset, seqoutall

-osformat	string	output sequence format
-osextension	string	filename extension
-osname	string	base filename
-osdbname	string	database name to add
-ossingle	bool	seperate file for each entry[N]
-oufo	string	features UFO
-offormat	string	features format
-ofname	string	features filename

Associated qualifiers: features

-fformat	string	features format
-fopenfile	string	features filename
-fask	bool	prompt for fbegin , fend , and freverse
-fbegin	integer	features starting position
-fend	integer	features end position
-freverse	boon	features on the reverse strand [N]

Associated qualiifers: featout

-offormat	string	feature format
-ofopenfile	string	output filename
-ofextension	string	filename extension
-ofname	string	filename
-ofsingl	bool	write one feature per file

Associated qualifiers: graph, xygraph

-gprompt	bool	graph prompting
-gtitle	string	graph title
-gsubtitle	string	graph subtitle
-gxtitle	string	x axis title
-gytitle	string	y axis title
-grtitle	string	right axis (y2) title
-gpages	integer	number of pages
-goutfile	string	output filename

EMBOSS and Graphics

EMBOSS can support a number of different graphics output types depending on the features available on your system. It will prompt for a graphics device:

Graphics device [x11]:

Typing rubbish here then pressing return will give a lengthy list of devices, many of which are equivalent.

The main graphics options are:

[X]	x11	Output to an X-window
	postscript	Output to a postscript file (good for printing on a laser printer)
	cps	Output to a colour postscript file
	text	Output to a text file
	data	Output XY data points to a file. (good for importing into a graphing package)
[P]	png	Output to a PNG image file (good for web pages)
[X]	Tek	Output to tektronics terminal
[X]	xterm	Output to an Xterm window
[X]- requires X-windows [P] – requires PNG support		

The default filename is *prog.format* eg. **octanol.ps**

Some useful programs

General	
wosname	lists all EMBOSS programs
showdb	Shows the available databases

Sequence retrieval

segret	retrieves and/or changes format of a sequence
segretset	retrieve and or change formats of a number of sequences at once
transeq	translate a DNA sequence to protein
backtranseq	translate a protein sequence to DNA
extractseq	extract regions from a sequence
cutseq	remove a region from a sequence
pasteseq	inserts a sequence into another sequence
infoseq	display information about a sequence
splitter	split a sequence into smaller sequences

Sequence comparison

needle	Needleman-Wunsch sequence alignment
water	Smith-Waterman sequence alignment
stretcher	Myers and Miller global alignment
matcher	Huang and Miller local alignment
dotdup	dotplot comparisons of two sequences.
dotmatcher	
prettyplot	plots multiple sequence alignments
polydot	dotplot comparisons of multiple sequences.
supermatcher	

Sequence parameters

cusp	generates a codon usage table
syco	synonymous codon usage plot
dan	calculates DNA/RNA melting temperature
compseq	sequence composition tables

DNA Sequence features

remap	restriction map of the sequence
cpgplot	CpG island detection
cpgreport	
etandem	finds tandem and inverted repeats
einverted	
plotorf	plots potential ORFs
showorf	pretty display of potential ORFs
fuzznuc	DNA pattern search
tfscan	scans sequence for TF binding sites

Protein Sequence features

ief	Isoelectric point calculation
antigenic	Finds potential antigenic sites
digest	protein digestion map
findkm	Vmax and Km calculations
fuzzpro	protein pattern search
garnier	protein 2D structure prediction
helixturnhelix	finds nucleic acid binding motifs
octanol	displays protein hydropathy
pepwindow	
patmatdb	searching with motifs vs protein sequences
patmatmotifs	
pepcoil	predicts coiled coil regions
pepinfo	Protein information
pepstats	
pepwheel	shows protein sequences as a helix.

File formats supported by EMBOSS

IntelliGenetics, Genbank, NBRF, EMBL, GCG, DNASTrider, Fitch, FASTA, Phylip, PIR, MSF, ASN.1, PAUP, ClustalW

This Quick Guide was written by and is copyright Dr David Martin at the Norwegian EMBnet node.

Comments and suggestions for improving this guide should be addressed to him at david.martin@biotek.uio.no

EMBnet is a network of academic and commercial bioinformatics institutes, supporting bioinformatics research and collaboration in more than countries worldwide.

More information about EMBnet and details of your local node can be found at <http://www.embnet.org>

An unlimited noncommercial right to redistribute the unamended document in printed or electronic form is granted without restriction.