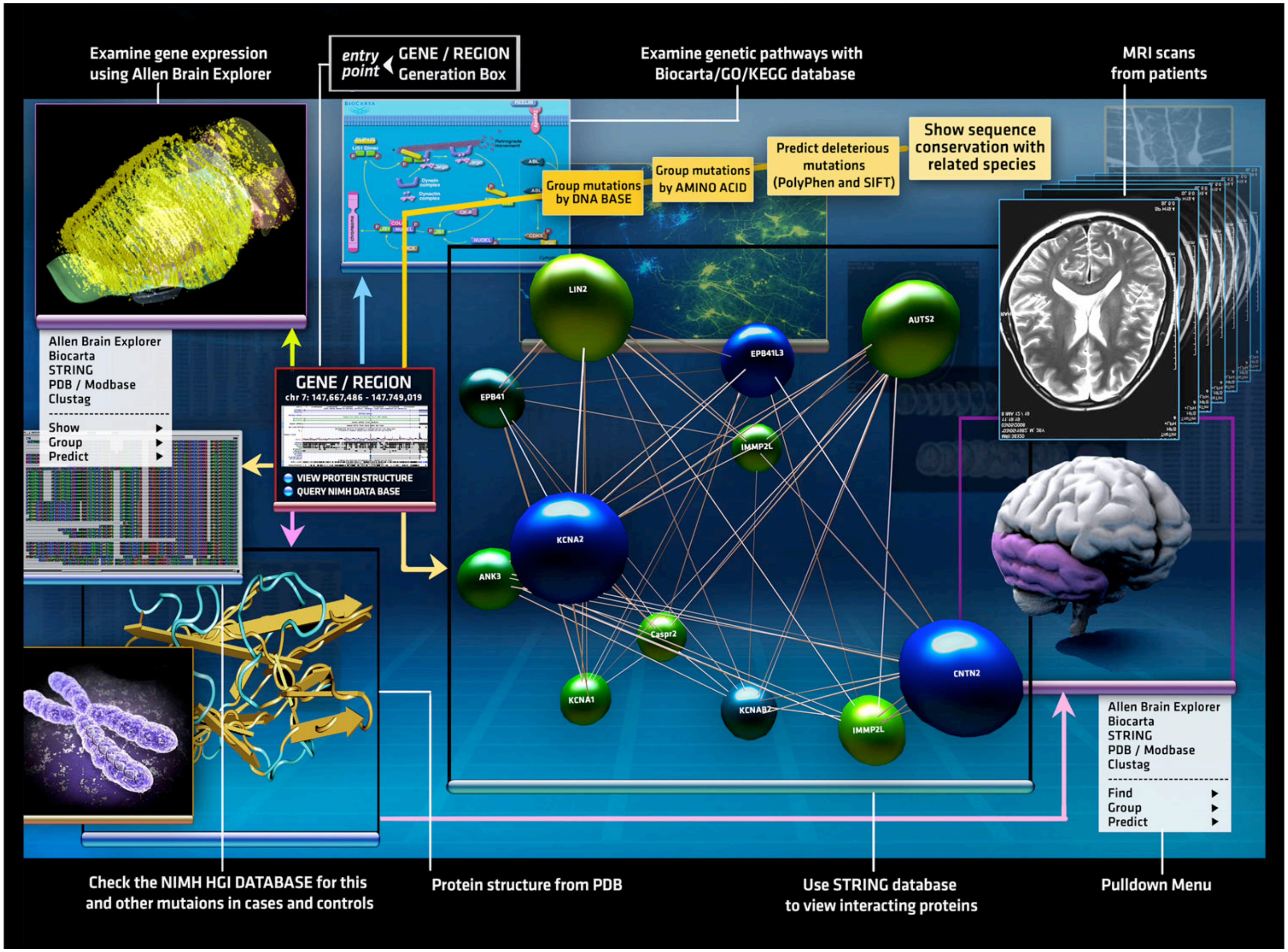
Next-Generation Sequencing (NGS) Technologies and Data Analysis

Christopher E. Mason

TA: Paul Zumbo

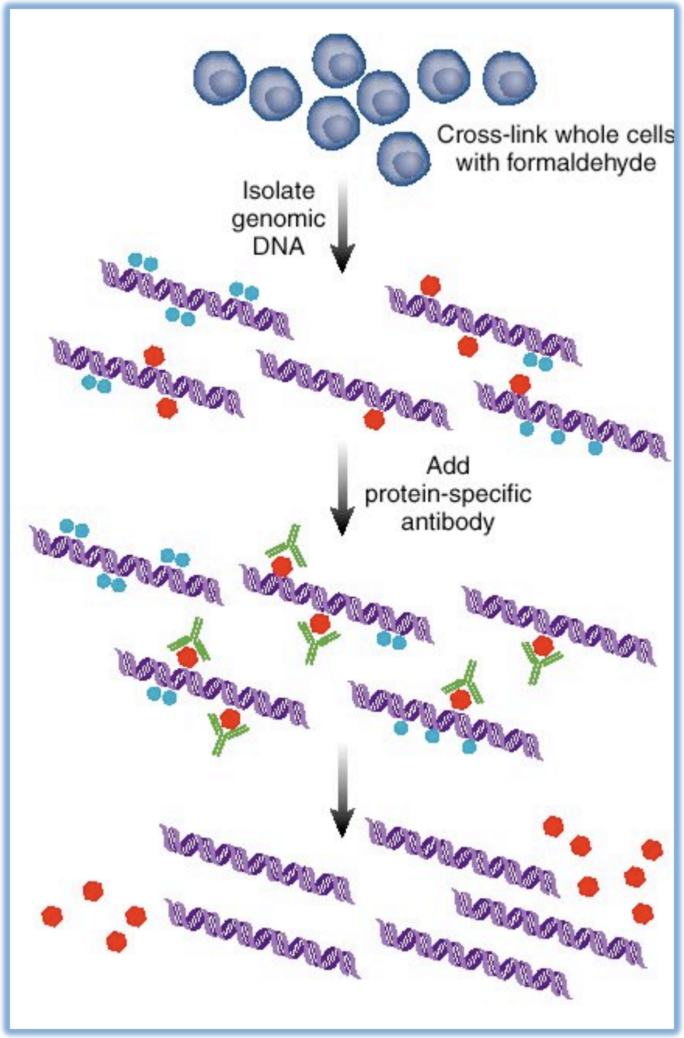
Spring 2010

Class #4: ChIP-Seq, Genome Assembly, and data parallelization, visualization, & integration


Examine gene expression using Allen Brain Explorer
entry point
GENE / REGION Generation Box
Examine genetic pathways with Biocarta/GO/KEGG database
MRI scans from patients
Show sequence conservation with related species
Predict deleterious mutations (PolyPhen and SIFT)
Group mutations by AMINO ACID
Group mutations by DNA BASE
Allen Brain Explorer
Biocarta
STRING
PDB / Modbase
Clustag
Show
Group
Predict
GENE / REGION
chr 7: 147,667,486 - 147,749,019
VIEW PROTEIN STRUCTURE
QUERY NIMH DATA BASE
LIN2
AUTS2
EPB41L3
EPB41
IMMP2L
KCNA2
ANK3
Caspr2
CNTN2
KCNA1
KCNAB2
IMMP2L
Allen Brain Explorer
Biocarta
STRING
PDB / Modbase
Clustag
Find
Group
Predict
Check the NIMH HGI DATABASE for this and other mutaions in cases and controls
Protein structure from PDB
Use STRING database to view interacting proteins
Pulldown Menu

ChIP-Seq

Chromatin Immunoprecipitation (ChIP) Sequencing allows you to assay the amount of binding and location of a protein to DNA, such as a transcription factor bound to the start site of a gene, or a histones of a certain type.


Cross-link whole cells with formaldehyde
Isolate genomic DNA
Add protein-specific antibody

Mardis, Nature Methods, 2008

Many available tools, java or other:
http://havoc.genomecenter.ucdavis.edu/cgi-bin/chipseq.cgi

but we will focus on ChIPseeqer:
http://icb.med.cornell.edu/wiki/index.php/Elementolab/ChIPseeqer_Tutorial

ChIP-Seeqer

From our very own Dr. Oliver Elemento!

Download

```
]http://physiology.med.cornell.edu/faculty/elemento/lab/files/ChIPseeqer-1.0.zip
```

Unzip

```
]unzip ChIPseeqer-1.0.zip
]cd ChIPseeqer-1.0/
```

Compile libmd

```
]cd libmd/
]perl Makefile.PL
]make
```

Compile chip-seeqer

(on mac, type just ‘ cd ../ ; make’)

OR, on Linux, type cd ../ ; make –f Makefile.linux

Set environmental variable (make sure to use back-ticks! cmd substituion)

```
export CHIPSEEQERDIR=`pwd`
```

ChIP-Seeqer

Get raw data from (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13084, Mapping polycomb complexes in human and mouse embryonic stem cells

Make a new directory for data analysis and move into that directory

```
]cd ..
]mkdir CHIP
]cd CHIP
```

Download raw data into the CHIP directory, taken from ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/samples/GSM327nnn/GSM327662/GSM327662_hES.H3K4me3.aligned.txt.gz

Then unzip the chip data

```
]gunzip GSM327662_hES.H3K4me3.aligned.txt.gz
```

Split aligned data

```
]perl ../SCRIPTS/split_bed_or_mit_files.pl GSM327662_hES.H3K4me3.aligned.txt
```

Start chip-seeqer

```
]../ChIPseeqer.bin -chipdir ../CHIP/ -format mit -uniquereads 1
>TF_targets.txt
```

ChIP-Seeqer Targets

Summarize the analysis

```
../ChIPseeqerSummary --targets=TF_targets.txt --lenu=2000 -lend=1000 --
suffix=TF_targets_SUM --db=RefGene
```

```
--targets=FILE file containing genomic regions
--lenu=INT     length upstream of TSS
--lend=INT     length downstream of TSS
--suffix=STR   suffix for output files
--db=STR       can be either RefGene or AceView. Default is RefGene
```

The file that ends with **_ALL.NM** will have: TranscriptID, Chromosome, TSS, TES, #peaks found

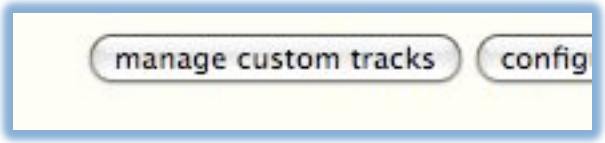
The file that ends with **.NM** will have only the transcripts with detected peaks (from _ALL.NM file).

The file that ends with **.SUM** will have: GeneID, GeneDescription, Chromo, TSS, TES, # peaks found

Visualize the data - Create a wiggle plot

```
../ChIPseeqer2Track --targets=TF_targets.txt --trackname="TF ChIPseeqer peaks“
```

We want to upload data to UCSC genome browser

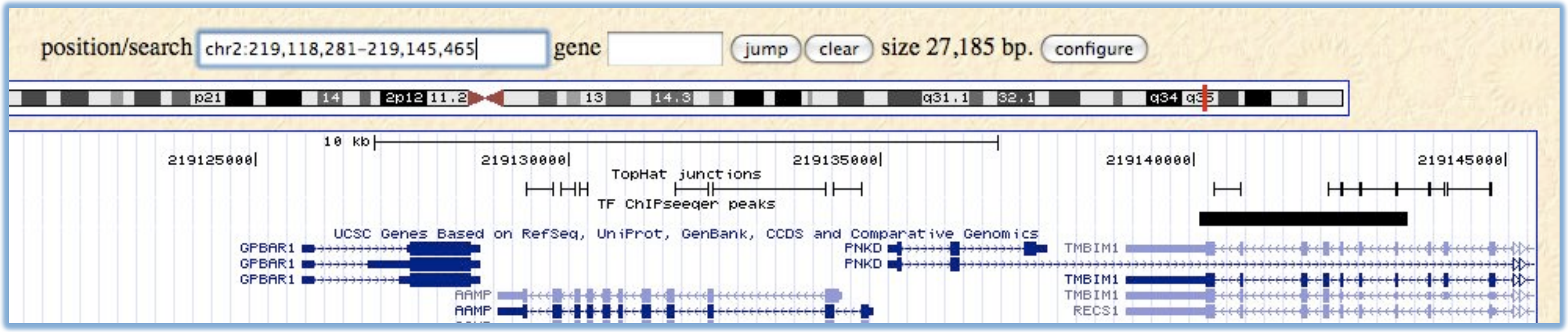


Be careful which genome you use!

Examine the data file

```
]more TF_targets_SUM.RefGene.SUM
```

In the genome browser, go to gene AAMP, at chr2:219,128,853-219,134,893.



position/search chr2:219,118,281-219,145,465 gene jump clear size 27,185 bp. configure

We see a peak more than 5000bp from the TSS of AAMP, yet we reported a peak within 2000bp. What could it be? Wrong genome build!

Discern the size of the file

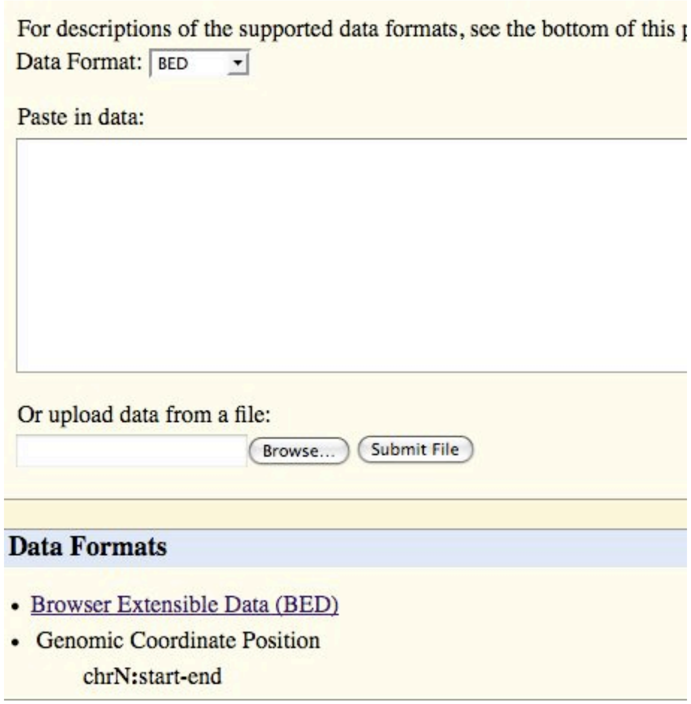
```
]wc TF_targets.txt.wgl
```

Take the tail of the file without the header ((file size in lines) – 1)

```
]tail -n 38002 TF_targets.txt.wgl >hg18targets.txt
```

Upload and convert the hg18target.txt with the LiftOver tool

Liftover Tool



For descriptions of the supported data formats, see the bottom of this p

Data Format: BED

Paste in data:

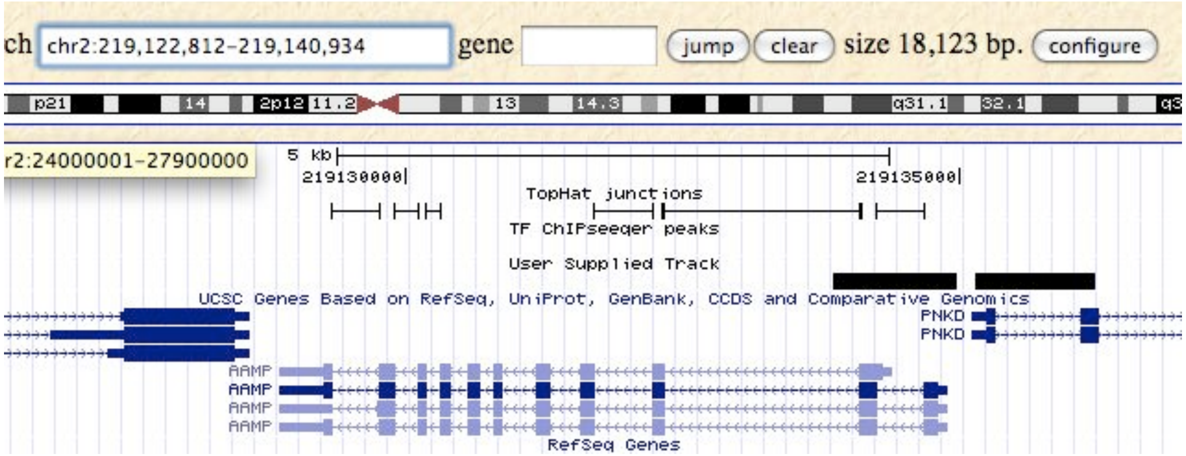
Or upload data from a file:

Data Formats

- Browser Extensible Data (BED)
- Genomic Coordinate Position

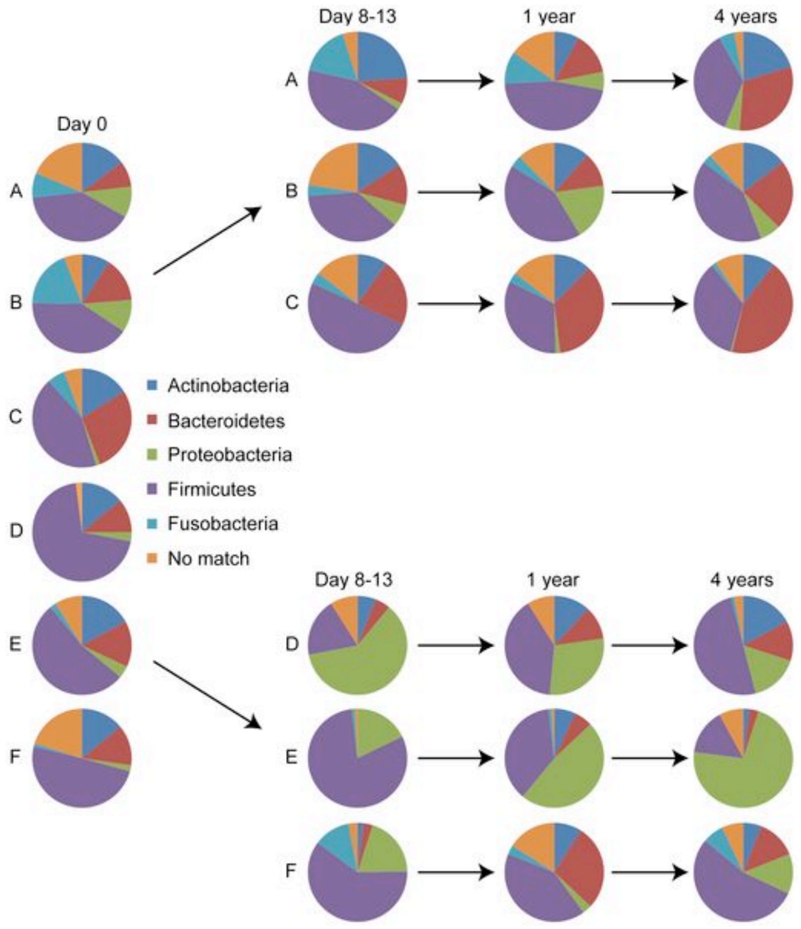
  chrN:start-end

Save as … hg19_chip_peaks.bed and upload



ch chr2:219,122,812-219,140,934 gene size 18,123 bp.

Meta-genomic phenotypes can persist for years, and “passenger genomes” can be a phenotype, as well as their distributions.



Normal throat

Throat + antibiotics

Jakobsson et al, 2010

What do we do with all the un-mapped reads? Answer meta-genomics questions!

First, install Velvet:

http://www.ebi.ac.uk/~zerbino/velvet/velvet_0.7.62.tgz

Then, we can use “grep” and “awk” to filter the non-header reads that have not mapped:

Count the number of mapped vs. unmapped reads in burge_liver_gdna.sam:

```
]wc burge_liver_gdna.sam
]grep –c ‘chr’ burge_liver_gdna.sam
]grep -v 'chr' burge_liver_gdna.sam >burge_liver_nohits.sam
]wc burge_liver_nohits.sam
]awk '{print ">"$1"\n"$10}' burge_liver_nohits.sam >burge_liver_nohits.fa
]cp burge_liver_nohits.fa velvet_07.62/
```

Velvet is good hash-based assembler for small genomes

http://www.ebi.ac.uk/~zerbino/velvet/

Download the file

http://www.ebi.ac.uk/~zerbino/velvet/velvet_0.7.62.tgz

Gunzip and Untar the tarball

```
]gunzip velvet_0.7.62.tgz
]tar –xvf velvet_0.7.62.tar
```

Compile the program

```
]make
```

Let’s see what else was in our liver…

First we create the hash tables and roadmap

```
]./velveth vout/ 31 burge_liver_no-hits.fa
```

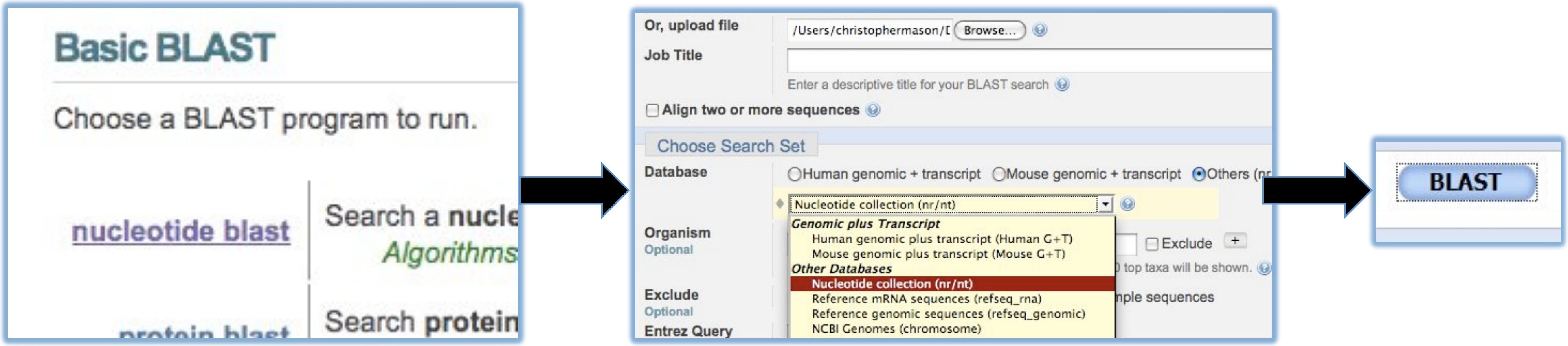
Then, we build de Bruijn graphs and manipulate them

```
] ./velvetg vout/
```

Then, examine what we have built:

```
]cd vout/
]more contigs.fa
```

BLAST file against NT/NR DB: http://blast.ncbi.nlm.nih.gov



Basic BLAST

Choose a BLAST program to run.

nucleotide blast — Search a nucle

Algorithms

protein blast — Search protein

Or, upload file: /Users/christophermason/[Browse...

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr

Nucleotide collection (nr/nt)

Genomic plus Transcript
- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases
- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)

Organism Optional

□ Exclude +

top taxa will be shown.

Exclude Optional

mple sequences

Entrez Query

How do you re-make a fragmented genome?

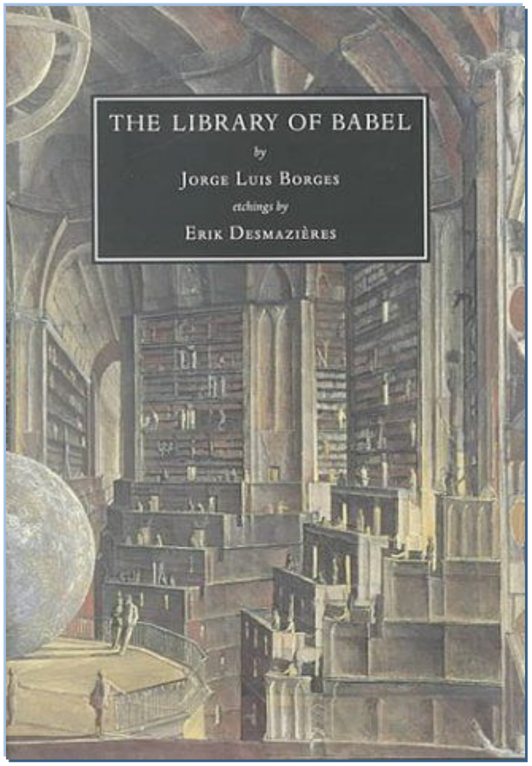
We have to build it from scratch…several options:

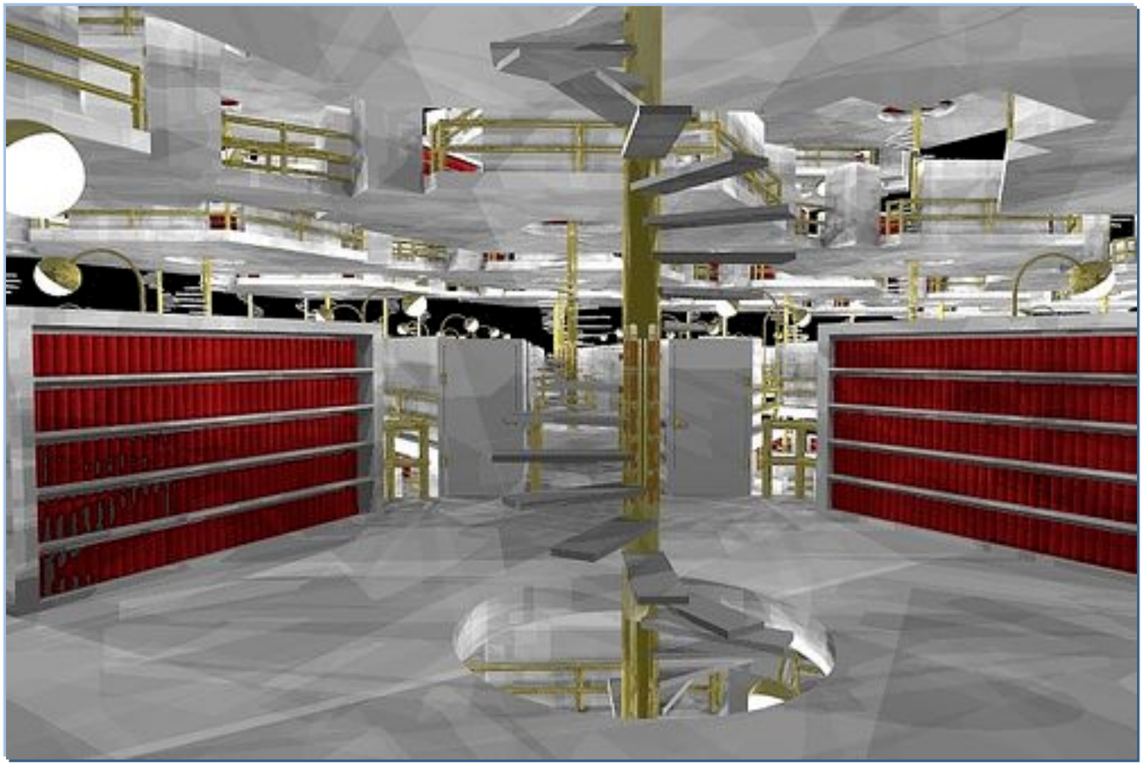
- Overlap/Layout/Consensus (OLC) Graph
- De bruijn graph (DBG)
- Greedy Graph

How do you re-make a genome & avoid “The Library of Babel?”

“His book was known as the Book of Sand, because neither the book nor the sand have any beginning or end.” — Jorge Luis Borges

We must avoid all permutations of a 410-page book.





Overlap/Layout/Consensus (OLC) Graph

(O): All vs. all, pairwise comparison, seed and then extend

Variables: K-mer size, Overlap length, % Identity

(L): Creation of an overlap graph reduces memory footprint and creates a read layout, showing the relationship between the overlaps

(C): Using multiple sequence alignment (MSA), the consensus sequence is generated using pair-wise alignments.

- **Newbler** – ideal for 454 reads, uses two rounds of OLC: first, uncontested unitigs, thencontigs from pair-wise aligned unitigs. Uses coverage to guide layout
- **Celera (CABOG**) – used for original human genome, also makes unitigs first. Uses error-correction from the overlaps, and a "best-overlap" filter.
- **Arachne**
- **CAP (and PCAP)**

De Bruijn Graph Assemblers

- De-Bruijn Graph Approach: k-mer graphs, reduces sequence complexity into hash tables, but memory intensive

- **Euler** – Spectral alignment that removes low frequency k-mers, (4^K must be less than 2xGenome Size). Makes distribution of reads comparing k-mer for one read vs. all (all is usually bi-model distribution of low-frequency k-mers vs. those in repeats). Also uses “mate-threading,” treating paired-end data as long sequences.
- **Velvet –** “Tour bus” algorithm allows each k-mer to start as its own node, builds bubbles and removes low-coverage paths. Also uses a “Rock Band” algorithm to make nodes with two or more long reads that have no contradiction. Uses mate pairs to find the graph that matches best to the insert size (“Breadcrumbs”).
- **ABySS** – Distributed assembler for mammalian sized genomes that iteratively removes spurs, compact version of Velvet. ABySS does not build scaffolds as of March 2010.
- **AllPaths** – Spectral Aligner that does read correcting before assembly, then makes unitigs into a database, then prunes. Creates a global graph for its assembly.
- **SOAPdenovo** – Filters and corrects for pre-set k-mer thresholds. Removes bubbles based on coverage. Processes the edges in order of insert size.

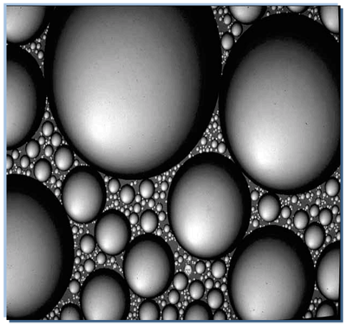
Greedy, Graph-Based Short Read Assemblers

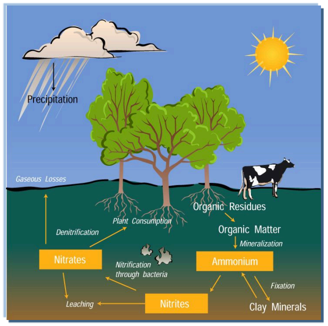
Stores the fragments as a directed graph- very memory intensive

- SSAKE – lookup table by seq-prefixes
- SHARCGS – Iterative extension, three stage pre-filtering for QVs
- VCAKE – Iterative extension, allowing imperfect matches

Common Problems in Assembly

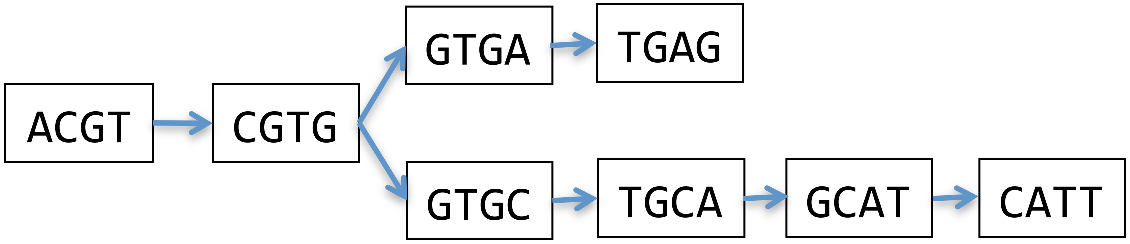
- Spurs – dead-end sequences (errors at end of a read)
- Bubbles – divergent paths that then converge (errors in the middle of a read)
- Frayed Rope- convergent then divergent paths
- Cycles – paths convergent upon themselves (repeats in the target genome)





Assembly Problems

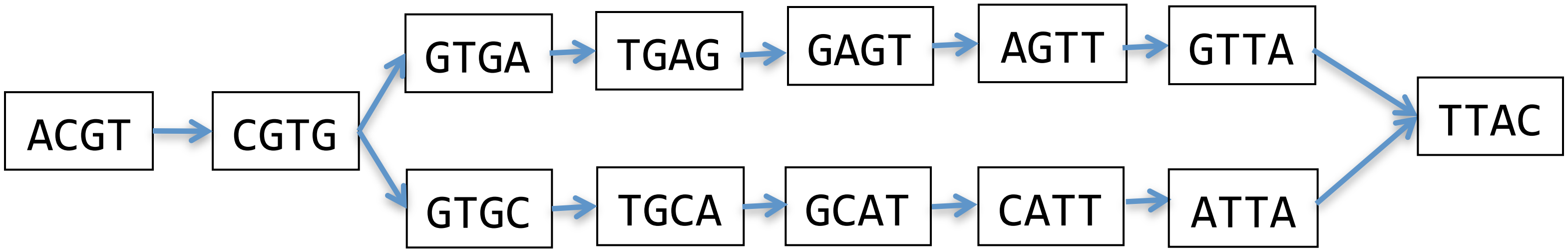
Spurs



ACGT CGTG GTGA TGAG

GTGC TGCA GCAT CATT

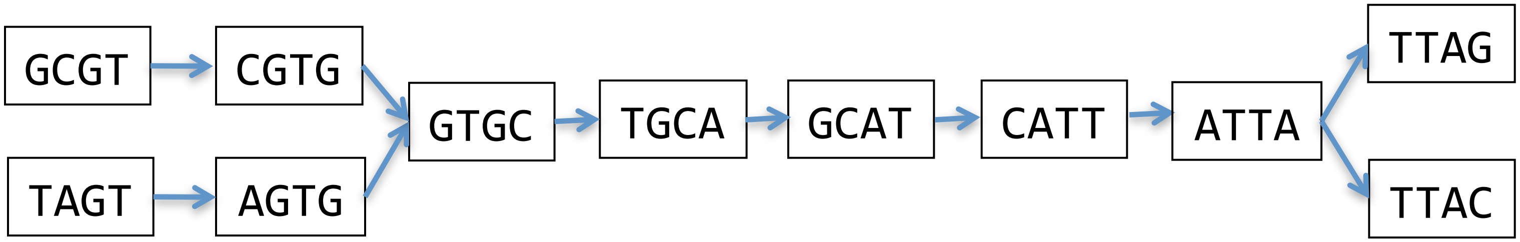
Bubbles



ACGT CGTG GTGA TGAG GAGT AGTT GTTA TTAC

GTGC TGCA GCAT CATT ATTA

Frayed Rope

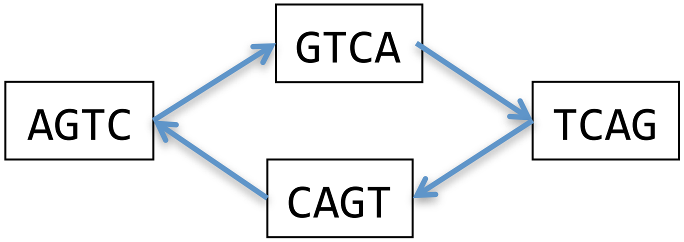


GCGT CGTG TTAG

TAGT AGTG GTGC TGCA GCAT CATT ATTA

TTAC

Cycles



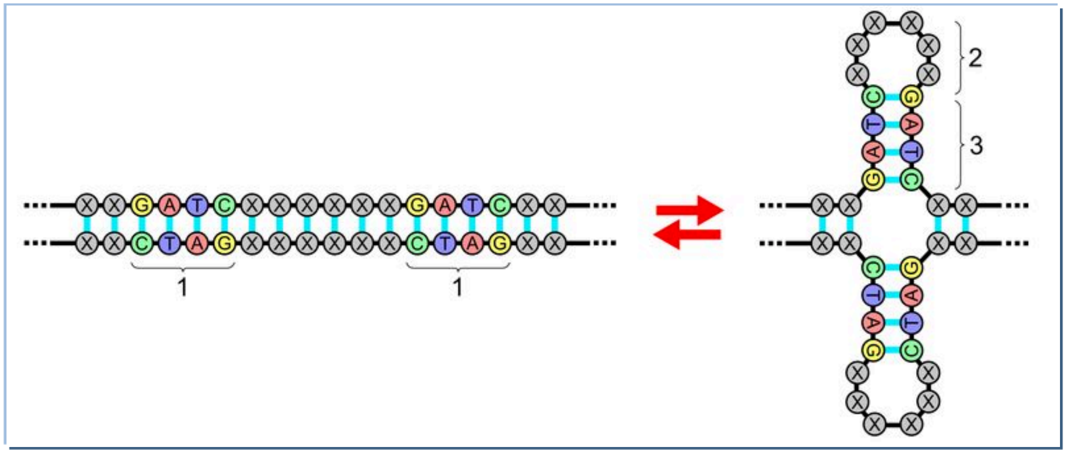
GTCA

AGTC TCAG

CAGT

Other Problems in Assembly

1. DNA is double stranded, so a forward hash may overlap the reverse of another
2. Repeats (tandem, inverted, segdups)
3. Palindromes: ACTAATCA (to avoid this, use an odd-numbered Kmer)



4. Sequencing error and native polymorphisms

Hadoop Map/Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Who uses it?

1. Crossbow (Bowtie)
2. Cloudburst (SNPs)
3. Amazon/Google/AOL
4. Twitter/Facebook

Useful Links for Annotation/Analysis

http://gvs.gs.washington.edu/SeattleSeqAnnotation/

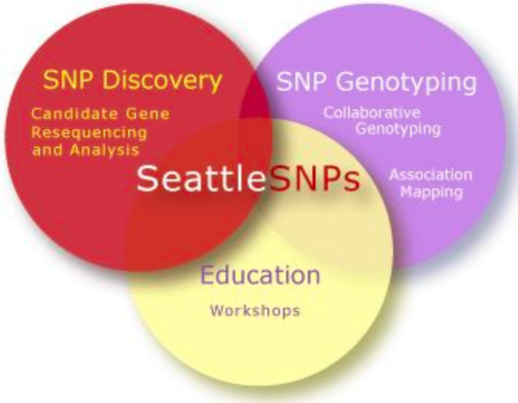
http://genome.ucsc.edu

http://cistrome.dfci.harvard.edu/trac/

http://main.g2.bx.psu.edu/

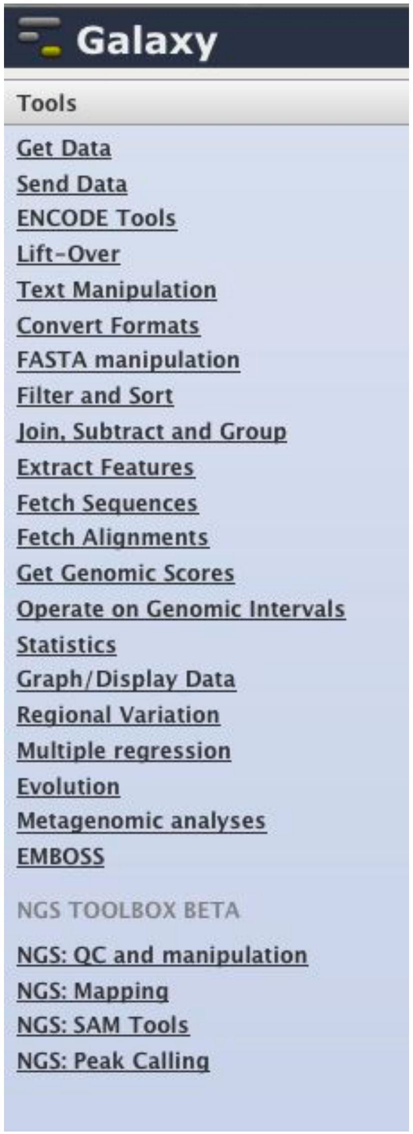
SeattleSNPs
Variation Discovery Resource

lood Institute's (NHLBI) Programs for Genomic Applications (PGA
otide polymorphisms (SNPs) in candidate genes and pathways th



RS&G Resequencing Genotyping NHLBI

Cistrome Project


Galaxy

Tools

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Evolution
Metagenomic analyses
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Peak Calling

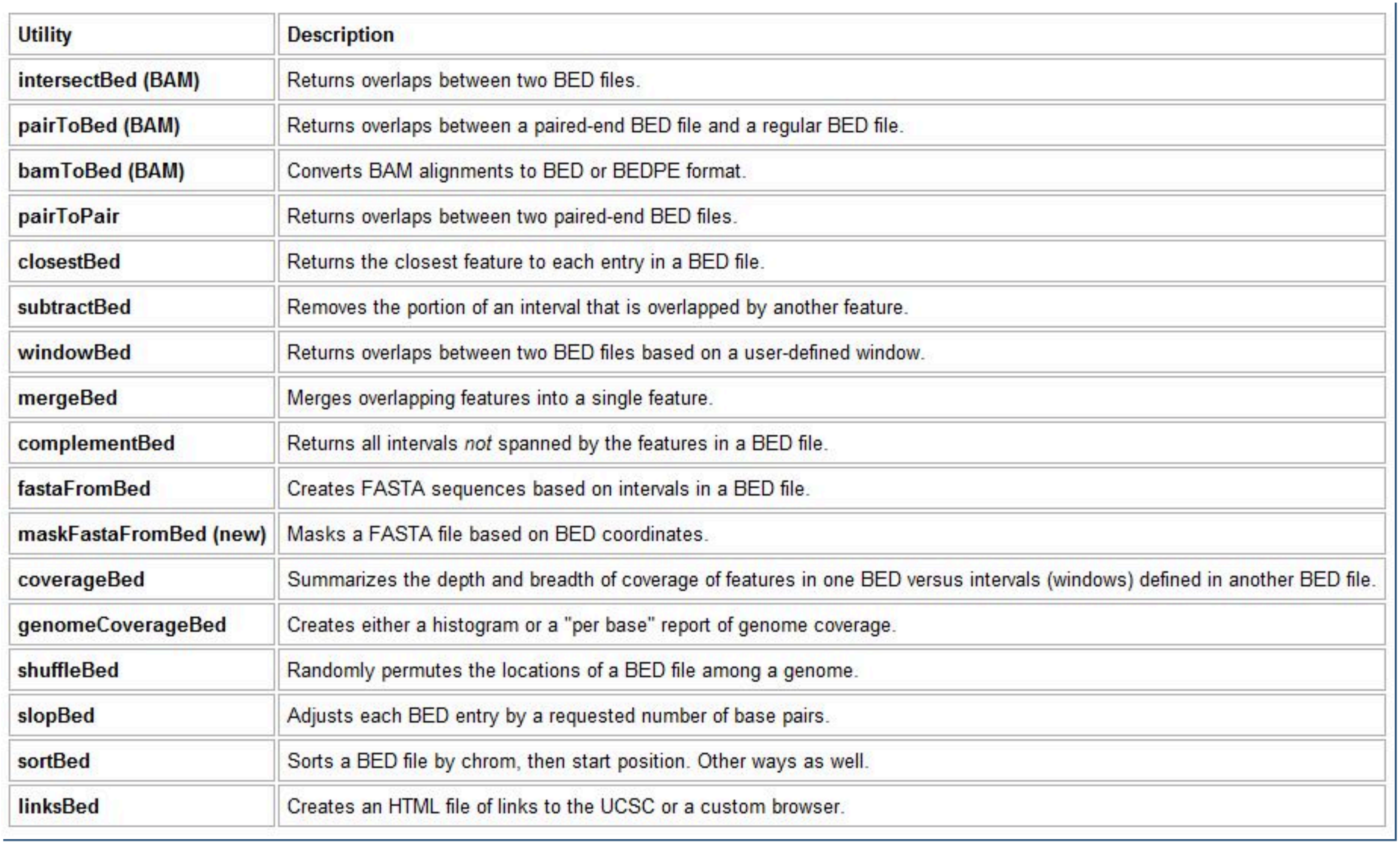
Other Useful Links

Useful Sequencing Sites:
http://getsatisfaction.com/gsa
http://seqanswers.com
http://seqanswers.com/wiki/SEQanswers
http://seqanswers.com/wiki/Software/list
http://code.google.com/p/bedtools/

Fasta files with SNPs-masked:
'ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/snp130Mask/*’

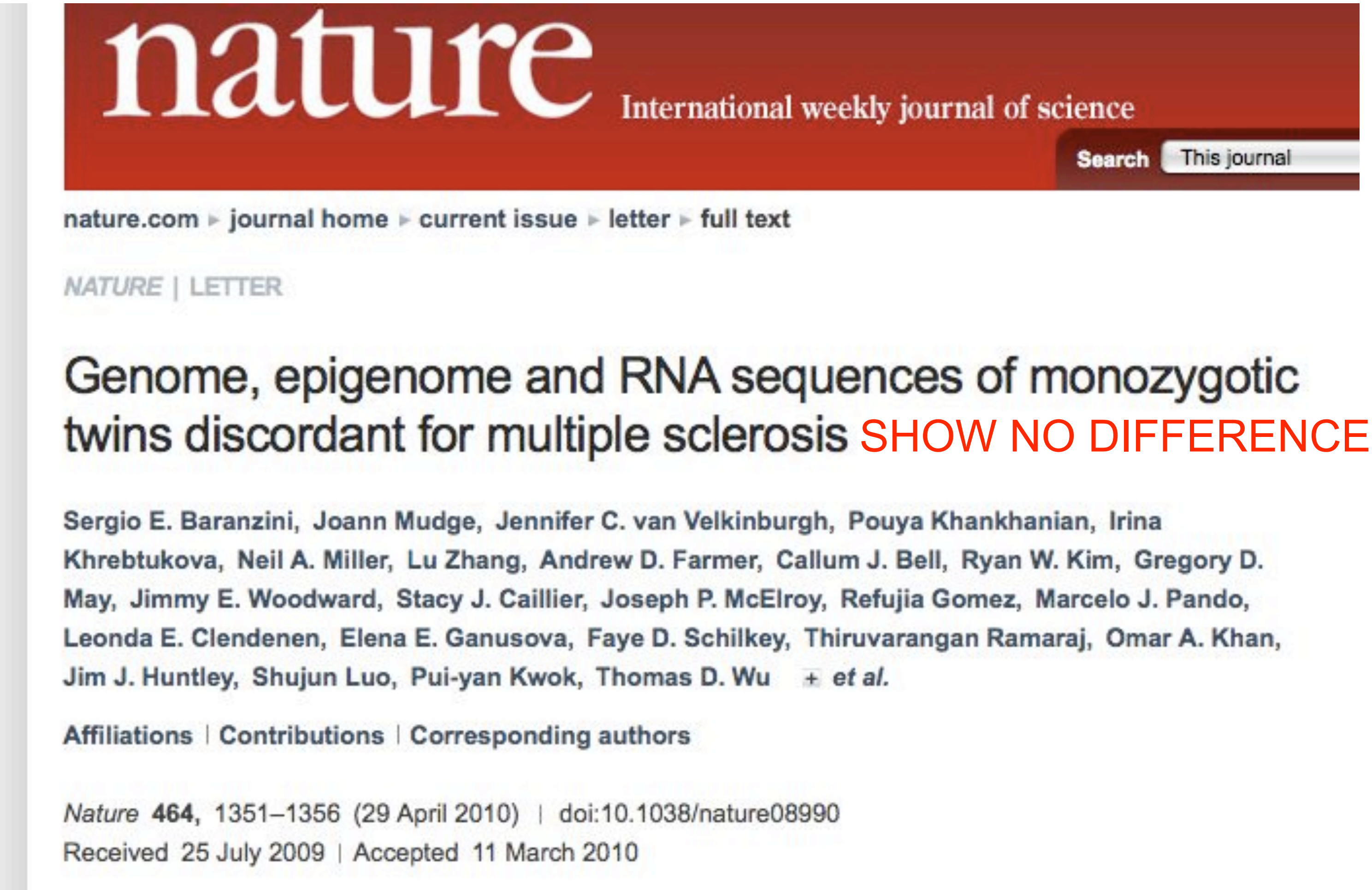
BedTools

http://code.google.com/p/bedtools/



Utility	Description
intersectBed (BAM)	Returns overlaps between two BED files.
pairToBed (BAM)	Returns overlaps between a paired-end BED file and a regular BED file.
bamToBed (BAM)	Converts BAM alignments to BED or BEDPE format.
pairToPair	Returns overlaps between two paired-end BED files.
closestBed	Returns the closest feature to each entry in a BED file.
subtractBed	Removes the portion of an interval that is overlapped by another feature.
windowBed	Returns overlaps between two BED files based on a user-defined window.
mergeBed	Merges overlapping features into a single feature.
complementBed	Returns all intervals *not* spanned by the features in a BED file.
fastaFromBed	Creates FASTA sequences based on intervals in a BED file.
maskFastaFromBed (new)	Masks a FASTA file based on BED coordinates.
coverageBed	Summarizes the depth and breadth of coverage of features in one BED versus intervals (windows) defined in another BED file.
genomeCoverageBed	Creates either a histogram or a "per base" report of genome coverage.
shuffleBed	Randomly permutes the locations of a BED file among a genome.
slopBed	Adjusts each BED entry by a requested number of base pairs.
sortBed	Sorts a BED file by chrom, then start position. Other ways as well.
linksBed	Creates an HTML file of links to the UCSC or a custom browser.

There are other factors than these!



nature International weekly journal of science

Search This journal

nature.com ▸ journal home ▸ current issue ▸ letter ▸ full text

NATURE | LETTER

Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis SHOW NO DIFFERENCE

Sergio E. Baranzini, Joann Mudge, Jennifer C. van Velkinburgh, Pouya Khankhanian, Irina Khrebtukova, Neil A. Miller, Lu Zhang, Andrew D. Farmer, Callum J. Bell, Ryan W. Kim, Gregory D. May, Jimmy E. Woodward, Stacy J. Caillier, Joseph P. McElroy, Refujia Gomez, Marcelo J. Pando, Leonda E. Clendenen, Elena E. Ganusova, Faye D. Schilkey, Thiruvarangan Ramaraj, Omar A. Khan, Jim J. Huntley, Shujun Luo, Pui-yan Kwok, Thomas D. Wu + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **464**, 1351–1356 (29 April 2010) | doi:10.1038/nature08990
Received 25 July 2009 | Accepted 11 March 2010

Systems biology requires spatiotemporal monitoring of the genome, epigenome, transcriptome, proteome, metabolome, and the environment, to see the interactome