Next-Generation Sequencing

Informatics Challenges

Next-Generation Sequencing

- Technology changes have revamped sequencing capabilities
  - Increased throughput
  - Decreased costs per base
- Informatics Challenges remain
  - Assembly, alignment
  - Resolving repetitive sequence

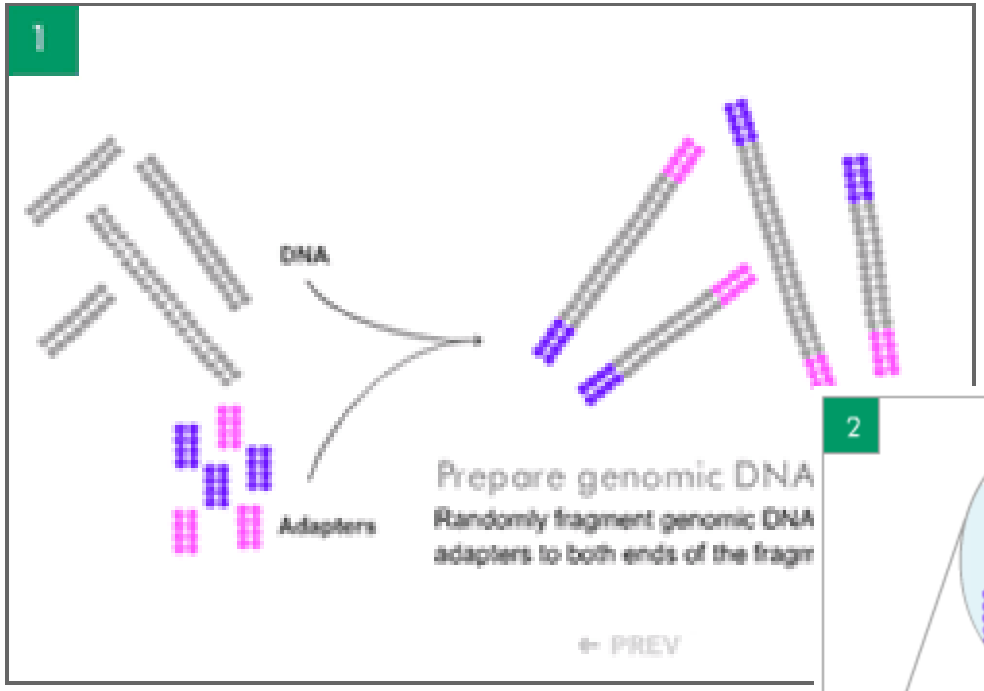
Outline

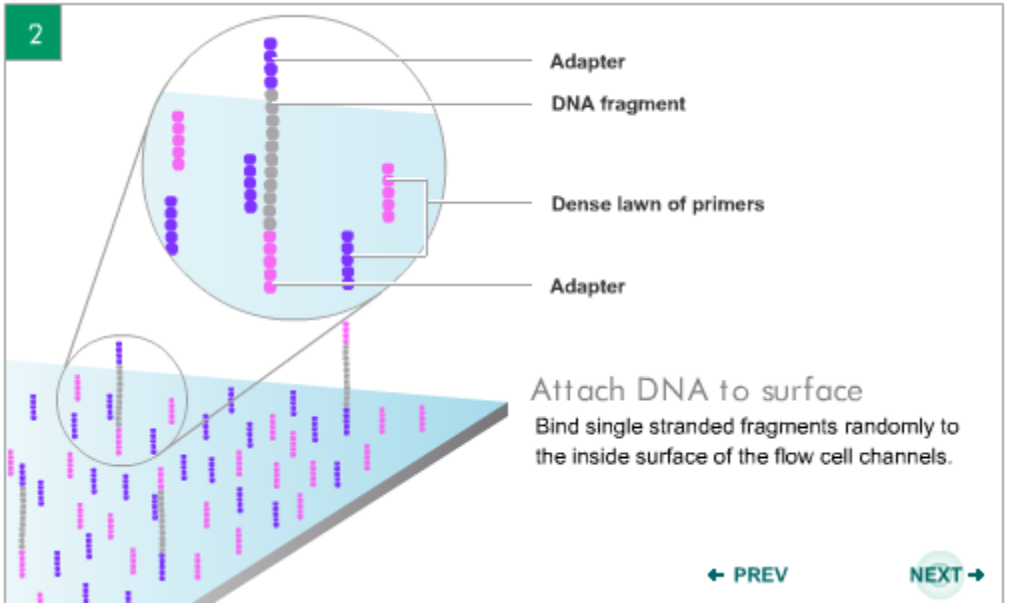
- Technology applied to biology
- Methods
- Informatics applications

Types of Applications to Biology

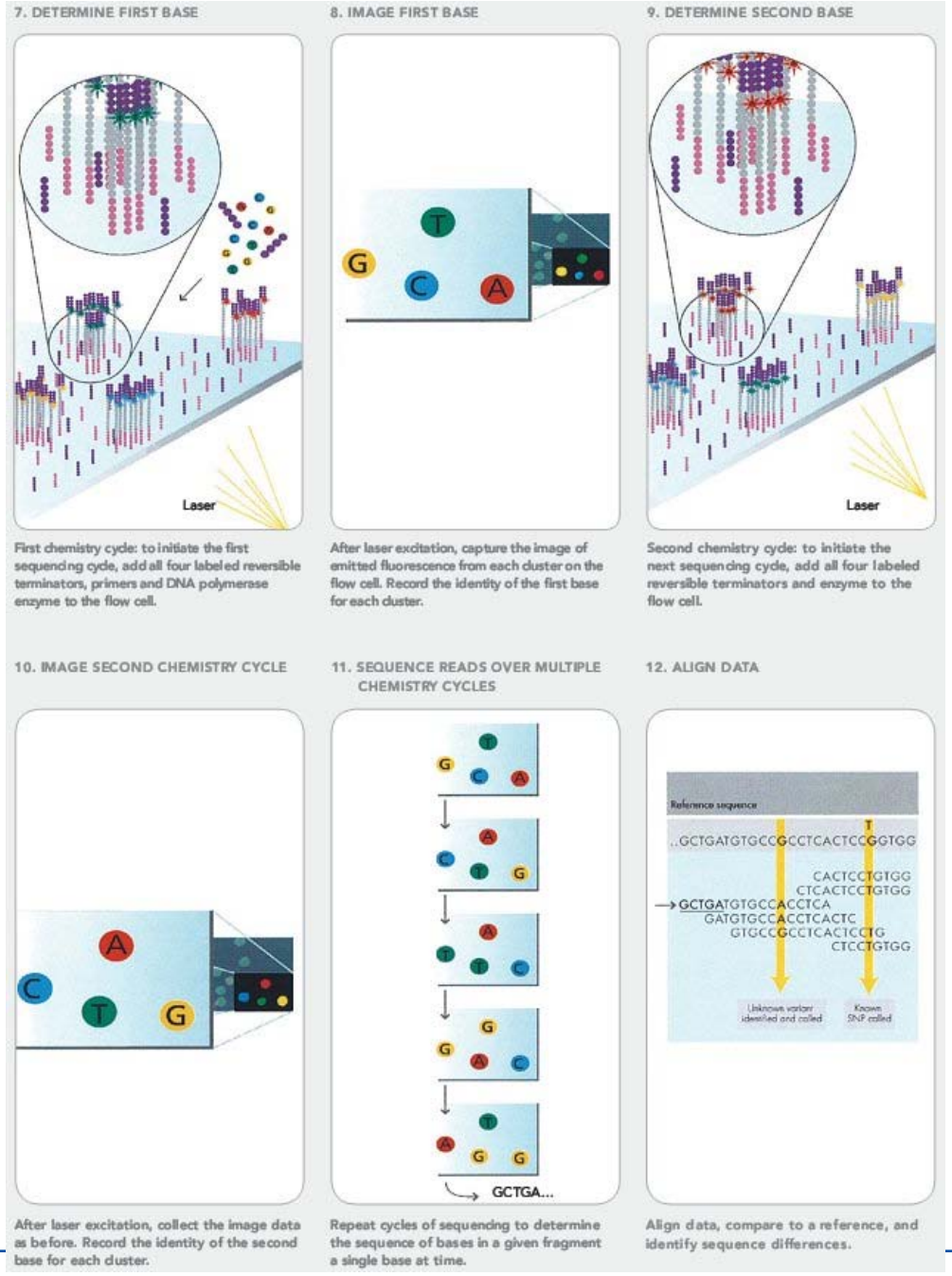
- Genomic Resequencing
  - Sequencing select genome regions, and comparing to a reference genome
- De novo assembly of novel genomes
  - Needs lots of depth of coverage
  - Works best for small (bacterial) genomes
  - Paired ends and different size libraries
    - 454 technology
- RNA Expression
  - Expressed genes and level of expression
- Protein binding to DNA (Promoters)
  - Immunoprecipitation of Protein bound to DNA
- Primer-specific sequencing (16S RNA)
  - Identifies communities of 16S RNA in microbe / samples
- Metagenomic sequencing (microbiome)

Illumina: Solexa

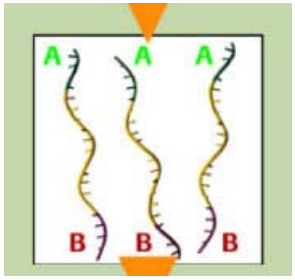




http://www.illumina.com/pages.ilmn?ID=203

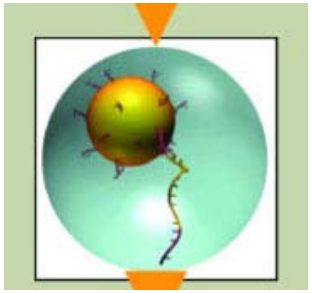

7. DETERMINE FIRST BASE
8. IMAGE FIRST BASE
9. DETERMINE SECOND BASE
Laser
T
G
C
A
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
10. IMAGE SECOND CHEMISTRY CYCLE
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
12. ALIGN DATA
A
C
T
G
GCTGA...
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
Align data, compare to a reference, and identify sequence differences.

454 Process



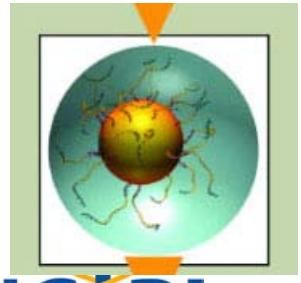
Library Preparation

Using a series of standard molecular biology techniques, short adaptors (A and B) - specific for both the 3' and 5' ends - are added to each fragment. The adaptors are used for purification, amplification, and sequencing steps. Single-stranded fragments with A and B adaptors compose the sample library used for subsequent workflow steps.



One Fragment = One Bead

The single-stranded DNA library is immobilized onto specifically designed DNA Capture Beads. Each bead carries a unique single-stranded DNA library fragment. The bead-bound library is emulsified with amplification reagents in a water-in-oil mixture resulting in microreactors containing just one bead with one unique sample-library fragment.

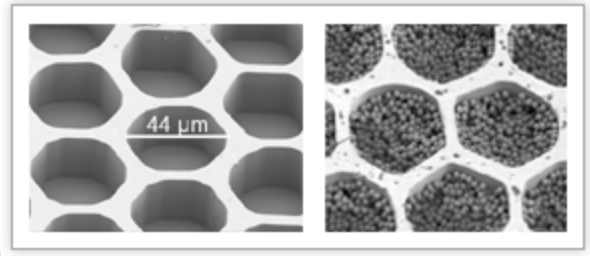


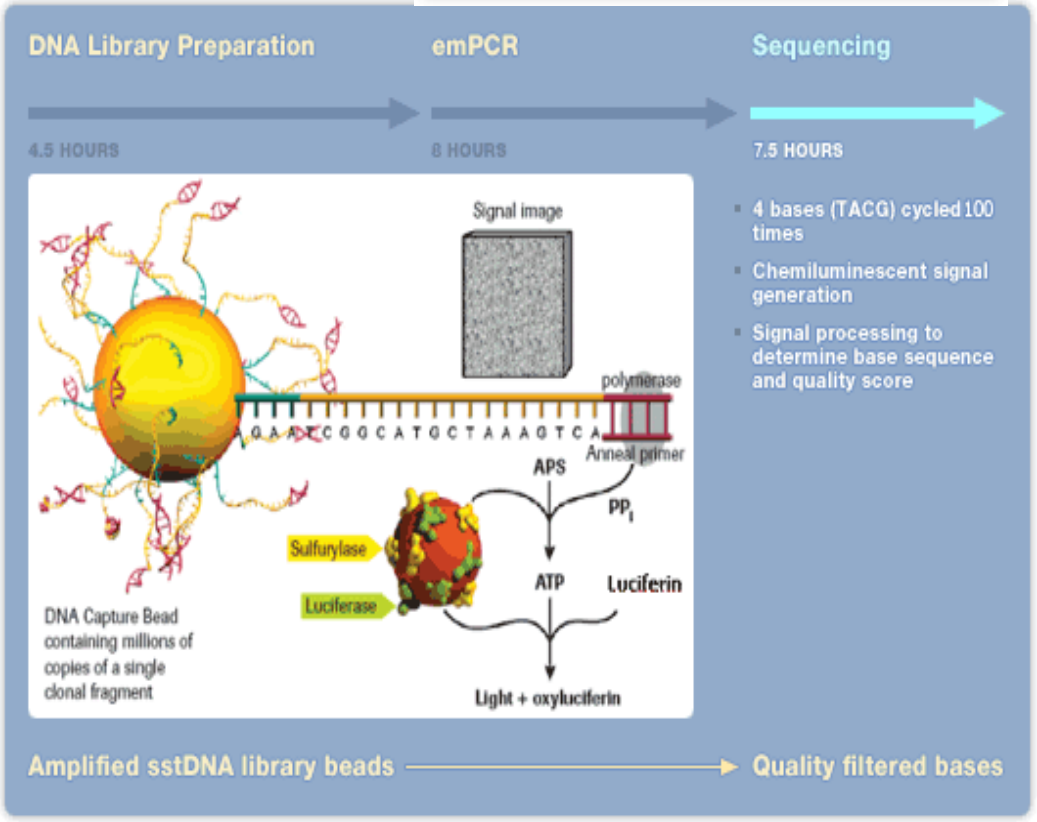
emPCR (Emulsion PCR) Amplification

Each unique sample library fragment is amplified within its own microreactor, excluding competing or contaminating sequences. Amplification of the entire fragment collection is done in parallel; for each fragment, this results in a copy number of several million per bead. Subsequently, the emulsion PCR is broken while the amplified fragments remain bound to their specific beads.

454 (Roche)

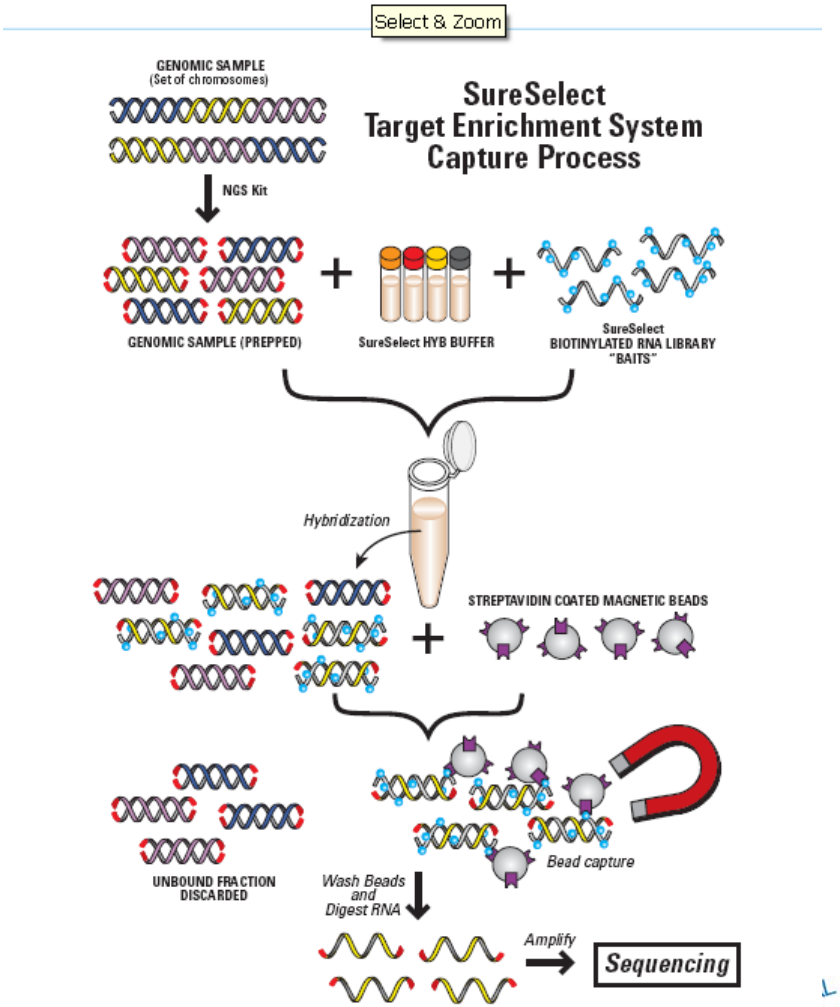
- Beads with millions of copies of DNA are sequenced in parallel.
- Polymerase extends the existing DNA strand by adding nucleotide(s). If a nucleotide complementary to the template strand is flowed into a well,
- The Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal that is recorded by the CCD camera.
- The signal strength is proportional to the number of nucleotides, for example, homopolymer stretches, incorporated in a single nucleotide flow



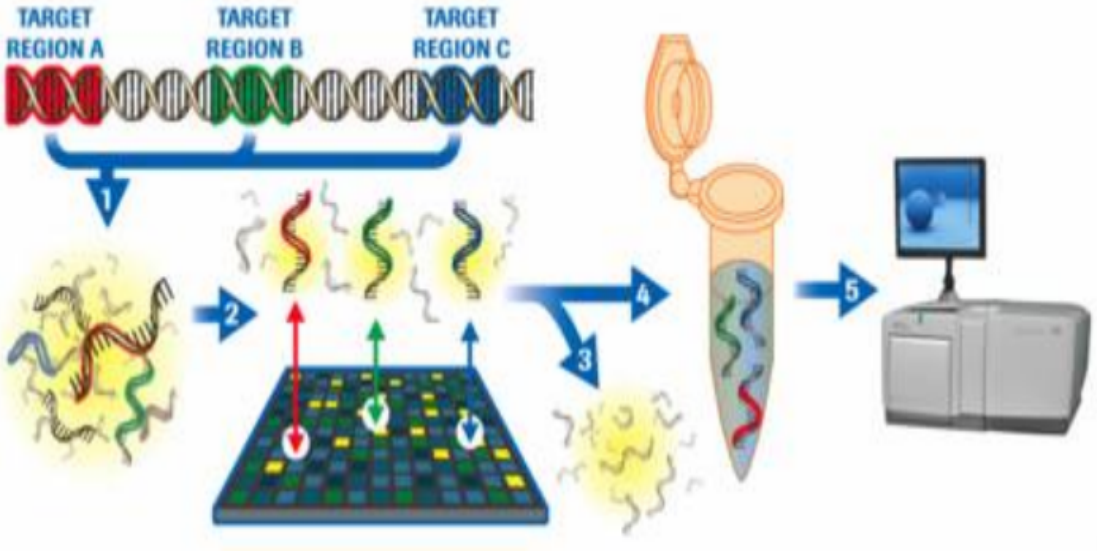


Enriching for portions of the Genome

- Agilent – SureSelect
  - Liquid capture strategy



Nimblegen (chip hybridization)



1. *The genomic DNA sample is fragmented by sonication or nebulization.*
2. *The sample is hybridized to a NimbleGen Sequence Capture array.*
3. *Unbound fragments are washed away.*
4. *The target-enriched pool is eluted and LM-PCR amplified.*
5. *The enriched sample is ready for high-throughput sequencing, such as with a 454 Genome Sequencer FLX instrument.*

Output from GenomeAnalyzer II (Illumina – Solexa)

- Read length 36 or 75 nt (100 nt and more, soon)
  - Paired-end reads as well as Mate-Pair reads
- 8 lanes per flow cell, 12-15 million reads per lane; 96-120 million reads per flow cell. (one lane control)
- ~ 7Gbases per flow cell
- Accuracy is ~99%
  - (34-70 million errors per flow cell)

- For human (diploid) there are ~ 6 Gbases of DNA so you would need 2 full runs (only 1 at 75 nt) per 1X coverage of the genome. To fully resequence you need at least 10-20X coverage (20-30 full chips at 36 nt)

Data output and processing

- Image data output (tiff files)
  - 100 tiles per lane, 8 lanes per flow cell, 36 cycles.
  - 4 images (A,G,C,T) per tile per cycle = 115,200 images
  - Each tiff image is ~ 7 MB = 806,400 MB of data
  - 1.6 TB per 70 nt read,
  - 3.2 TB for 70 nt Paired-end read
- Illumina Pipeline:
  - Firecrest (image analysis)
    - Locates clusters and calculates intensity and noise
  - Bustard (base calling)
    - Deconvolutes signal and corrects for cross-talk, phasing
  - GERALD – generation of recursive analyses linked by dependency
    - ELAND – (Efficient large-scale alignment of nucleotide databases)

Other software applications for assembly and alignment

Align/Assemble to a reference

* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hou workstation with 2 gigabytes of memory. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony Solexa 1G machine.
* EULER - Short read assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research).
* Exonerate - Various forms of alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Bi EMBL. C for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C,
* MOSAIK - Reference guided aligner/assembler. Written by Michael Strömberg at Boston College.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has prelimina handle ABI SOLiD data. Written by Heng Li from the Sanger Centre.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficier library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, A Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. P(required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Available free for evaluatio use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* RMAP - Assembles 20 - 64 bp Solexa reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinforma OS required.
* SeqMap - Works like ELand, can do 3 or more bp mismatches and also INDELs. Written by Hui Jiang from the Wong lab at Stanford. Builds available for n
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Bru Stephen Rumble at the University of Toronto.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a sequence or a set of reference sequences.. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto referenc Author is Ruiqiang Li at the Beijing Genomics Institute. C++ for Unix.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases usi Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq genome builds for a number of organisms. Web Portal. OS independent.

***de novo* Align/Assemble**

* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing te or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecul
* SSAKE - Version 2.0 of SSAKE (23 Oct 2007) can now handle error-rich sequences. Authors are René Warren, Granger Sutton, Steven Jones and Robert Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
* VCAKE - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
* Velvet - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X co paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).

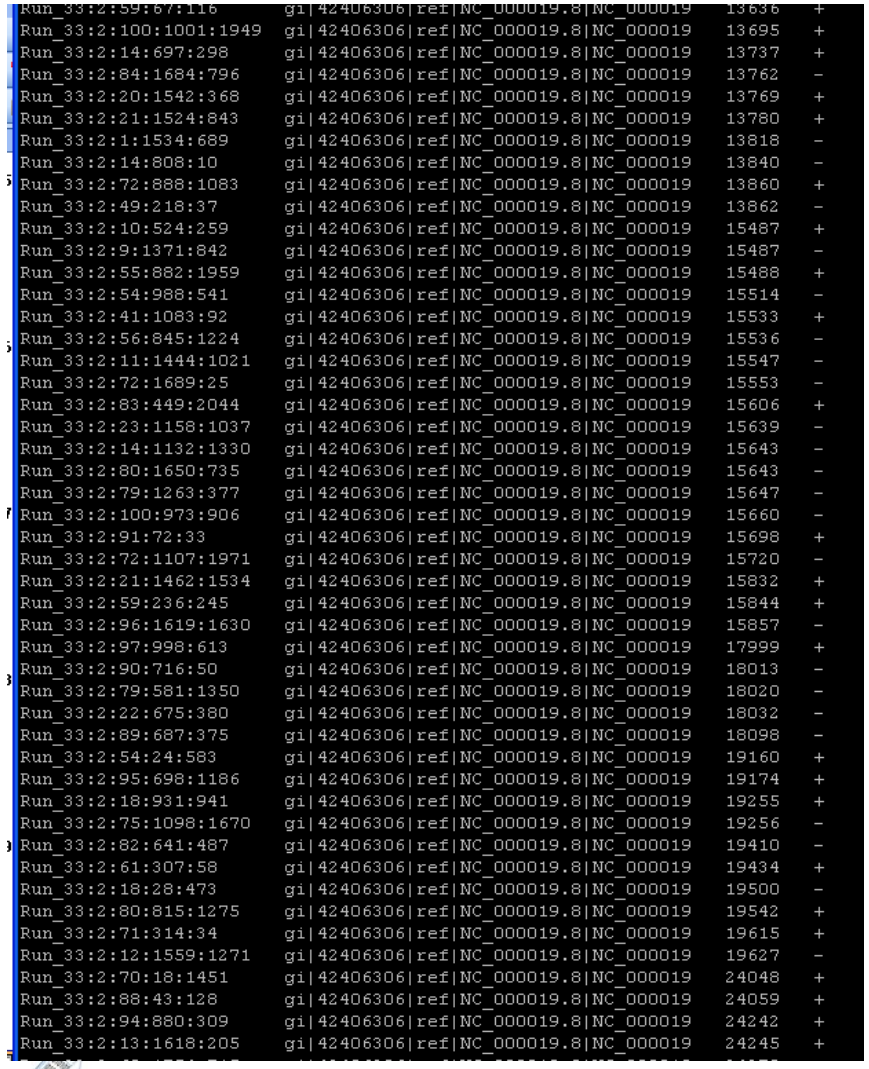
SNP/Indel Discovery

http://seqanswers.com/

Bioinformatics workflow

- Image extraction
- Base Calling, quality scoring
- Align reads to known sequence OR each otherr
- Assemble Reads
- Analysis of genes, regions
- Coverage, quantification
- Annotation

Sequence text output

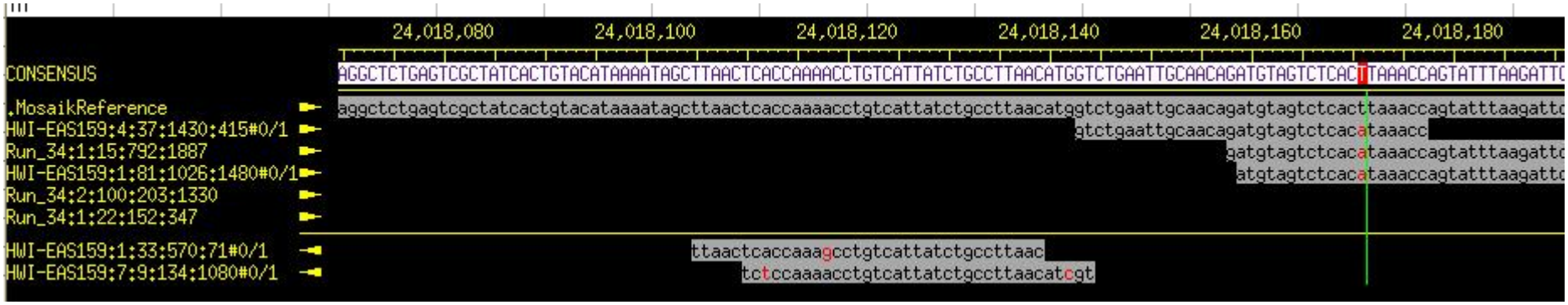


Run_33:2:59:67:116	gi|42406306|ref|NC_000019.8|NC_000019	13636	+
Run_33:2:100:1001:1949	gi|42406306|ref|NC_000019.8|NC_000019	13695	+
Run_33:2:14:697:298	gi|42406306|ref|NC_000019.8|NC_000019	13737	+
Run_33:2:84:1684:796	gi|42406306|ref|NC_000019.8|NC_000019	13762	-
Run_33:2:20:1542:368	gi|42406306|ref|NC_000019.8|NC_000019	13769	+
Run_33:2:21:1524:843	gi|42406306|ref|NC_000019.8|NC_000019	13780	+
Run_33:2:1:1534:689	gi|42406306|ref|NC_000019.8|NC_000019	13818	-
Run_33:2:14:808:10	gi|42406306|ref|NC_000019.8|NC_000019	13840	-
Run_33:2:72:888:1083	gi|42406306|ref|NC_000019.8|NC_000019	13860	+
Run_33:2:49:218:37	gi|42406306|ref|NC_000019.8|NC_000019	13862	-
Run_33:2:10:524:259	gi|42406306|ref|NC_000019.8|NC_000019	15487	+
Run_33:2:9:1371:842	gi|42406306|ref|NC_000019.8|NC_000019	15487	-
Run_33:2:55:882:1959	gi|42406306|ref|NC_000019.8|NC_000019	15488	+
Run_33:2:54:988:541	gi|42406306|ref|NC_000019.8|NC_000019	15514	-
Run_33:2:41:1083:92	gi|42406306|ref|NC_000019.8|NC_000019	15533	+
Run_33:2:56:845:1224	gi|42406306|ref|NC_000019.8|NC_000019	15536	-
Run_33:2:11:1444:1021	gi|42406306|ref|NC_000019.8|NC_000019	15547	-
Run_33:2:72:1689:25	gi|42406306|ref|NC_000019.8|NC_000019	15553	-
Run_33:2:83:449:2044	gi|42406306|ref|NC_000019.8|NC_000019	15606	+
Run_33:2:23:1158:1037	gi|42406306|ref|NC_000019.8|NC_000019	15639	-
Run_33:2:14:1132:1330	gi|42406306|ref|NC_000019.8|NC_000019	15643	-
Run_33:2:80:1650:735	gi|42406306|ref|NC_000019.8|NC_000019	15643	-
Run_33:2:79:1263:377	gi|42406306|ref|NC_000019.8|NC_000019	15647	-
Run_33:2:100:973:906	gi|42406306|ref|NC_000019.8|NC_000019	15660	-
Run_33:2:91:72:33	gi|42406306|ref|NC_000019.8|NC_000019	15698	+
Run_33:2:72:1107:1971	gi|42406306|ref|NC_000019.8|NC_000019	15720	-
Run_33:2:21:1462:1534	gi|42406306|ref|NC_000019.8|NC_000019	15832	+
Run_33:2:59:236:245	gi|42406306|ref|NC_000019.8|NC_000019	15844	+
Run_33:2:96:1619:1630	gi|42406306|ref|NC_000019.8|NC_000019	15857	-
Run_33:2:97:998:613	gi|42406306|ref|NC_000019.8|NC_000019	17999	+
Run_33:2:90:716:50	gi|42406306|ref|NC_000019.8|NC_000019	18013	-
Run_33:2:79:581:1350	gi|42406306|ref|NC_000019.8|NC_000019	18020	-
Run_33:2:22:675:380	gi|42406306|ref|NC_000019.8|NC_000019	18032	-
Run_33:2:89:687:375	gi|42406306|ref|NC_000019.8|NC_000019	18098	-
Run_33:2:54:24:583	gi|42406306|ref|NC_000019.8|NC_000019	19160	+
Run_33:2:95:698:1186	gi|42406306|ref|NC_000019.8|NC_000019	19174	+
Run_33:2:18:931:941	gi|42406306|ref|NC_000019.8|NC_000019	19255	+
Run_33:2:75:1098:1670	gi|42406306|ref|NC_000019.8|NC_000019	19256	-
Run_33:2:82:641:487	gi|42406306|ref|NC_000019.8|NC_000019	19410	-
Run_33:2:61:307:58	gi|42406306|ref|NC_000019.8|NC_000019	19434	+
Run_33:2:18:28:473	gi|42406306|ref|NC_000019.8|NC_000019	19500	-
Run_33:2:80:815:1275	gi|42406306|ref|NC_000019.8|NC_000019	19542	+
Run_33:2:71:314:34	gi|42406306|ref|NC_000019.8|NC_000019	19615	+
Run_33:2:12:1559:1271	gi|42406306|ref|NC_000019.8|NC_000019	19627	-
Run_33:2:70:18:1451	gi|42406306|ref|NC_000019.8|NC_000019	24048	+
Run_33:2:88:43:128	gi|42406306|ref|NC_000019.8|NC_000019	24059	+
Run_33:2:94:880:309	gi|42406306|ref|NC_000019.8|NC_000019	24242	+
Run_33:2:13:1618:205	gi|42406306|ref|NC_000019.8|NC_000019	24245	+

tggtggggagagagtccctccccctgccccctgtctct	(((((((((((((((((((((((((((((((((((((((
aggggaagggttcaaagctggtcacatccccAccaa	((((((((((((((((((((((((((((((((]((($
ccatggacaacgaaaagCCCACtaGcTtGTCCAGTG	(('((((((((((((%(`]VUU("Z![(ZRYYPJOW
cttgtccagtgccacaggaggggcaagtggaggagg	(((((((((((((((((((((((((((((((((((((
agTgccacAggAggGgcAagTggAggAggAgAgGTg	((T((((Q((X((Y(^((V(()[(N((G(Y(^A(
aggggcAAgtggaggAggAgAggtggcggTGCTCCC	((((((OO(((%((U(()(Q(((((("((RXOWIW^
CCCCAcTGCCagtCgTcactggctctcccttccctt	O@VU>(HVMV(((X(K((((((((((("(((((((
ctctcccttcccttcatccTcgttcccctatctgtca	((((((((((((((((((S((((((((((((((((
cgttccctatctgtcaccatttcctgtCGtcGtttc	((((((((((((((((((((((((#DN((R((((
ttccctatctgtcGccatttcctgtcgtcgtttcct	(((((!'((#((((A(((((((((((((((((((((
ggcaaggaaacacaatttctgagggaatggTtttgG	((((((((((((((((((((((((((((((X(((R
ggCaaggaaacacaatttcTgagggaatggttttgg	((Q(((((((((((((Z((((((((((((((((((
gcaaggaaacacaatttctgagggaatggttttggc	((((((((((((((((((((((((((((((((#((
tggttttggcctccattctaagtgctggacatgggg	(((((((((((((((((((((((((((((((((((
aagtgctggacatggggtggccataatctggagctg	(((((((((((((((((((((((((((((((((((
TgCTggacaTggggtggccataatctggagctgatg	R(WP(((((M((((((((((((((((((((((((((
gggtggccataatctggagctgatggctcttaaaga	(((((((((((((((((((((((((((((((((((
ccataatctggagctgatggctcttaaagacctgca	(((((((((((((((((((((((((((((((((((
ccctcgtgcacatttagcacaaagataagcacaaaA	((((((((((((((((((((((((((((((((((N
aaaggTgcatccagcActttgttactattggtggca	(((((Z(((((((((X((((((((((((((((((
gtgcatccagcactTtgttactattggtggcaggtt	(#((((((((((P((((((((((((((((((((((
GTgcaTccagcactttGTtactaTtggtGgcaggtt	YY%(&W(((((((((^`((((&_(((R(((((((
atccagcactttgttactattggtggcaggttcatg	(((((((((((((((((((((((((((((((((((
ttnctaTtggtggcaggttcatgaatggcaaccaaa	((!$((U!(((((((((((((((((((((((((((
cagtgtacgggtcaagattatcgacagggaagagaT	((((((((((((((((((((((((((((((((((_
aacagggaagagatagcatttcctgaaggcttccta	('(((((((((((((((((((((((((((((((((((
attattaccacaacttcacaaatgagaacaccgagg	(((((((((((((((((((((((((((((((((((
acttcacaaatgagaacaccgaggcttagaggggtt	(((((((((((((((((((((((((((((((((((
gaacaccgaggcttagaggggttgggttgcccaagg	(((((((((((((((((((((((((((((((((((
ccacttaacccctgaggAatttgaggcCtGctcct	((((((((((((((((((Y(($(((((Y(O((($((
gaggaatTtgaggccTgcTcctgaaacagactgggc	(((((((X(((((((D((P((((((((((((((((
ttgaggcctgctcctgaaacagactgggcagtggct	(((((((((((((((((((((((((((((((((((
Cctgaaacagactgggcatggctagtgactctagg	V(((((((((((('(((_((((((((((((((((
CTagaGcTTaGGGGgccaagaggaaagaggtgcctg	O`&#(Y&TT#U_[((((((((((((((((((((
tacacctgatgagtggtttactttctgtctgcaaac	(((((((((((((((((((((((((((((((((((
ggtttactttctgtctgcaaacatctactgatcatc	(((((((((((((((((((((((((((((((((((
cactagccagggagagtctcaaaaacaactaaactc	(((((((((((((((((((((((((((((((((((
actagccagggagagtctcaaaaacaactaaactca	(((((((((((((((((((((((((((((((((((
tcggctcacgcctgtaatcccagcactttgggaggc	(((((((((((((((((((((((((((((((((((
actttgggaggcgaaggcagacggatcacctgaggt	(((((((((((((((((((((((((((((((((((
atgaaactCcatctctactaaaaatacaaaattagc	((((((((`(((("(((((((((((((((((((((
tggtggtgcatgcctgtaatccccgctactcgggAg	((((((((((((((((((((((((((((((((((E(
ggTtgcagtgtgccaacatcgcgccattgcactccA	(([((((((((((((((((((((((((((((((((V
cCaagaTTgcgccatGgcacTccagcctaggcaacg	(W(((GK(((((((K(((W((((((((((((((((
tgggcgtggtggctcatgcctgtaatcctagcactt	(((((((((((((((((((((((((((((((((((
gctcatgcctgtaatcctagcactttggtaggctga	((((((((((((((((((((((((((((('((((((
acttgagcttgggagatggaggctgcAgtgagctGT	((((((((((((((((((((((((W(($''!(LR
tgagcttgggagatggaggctgcagTgagctgTgat	((((((((((((((((((((((((((((([(((((D%((
tgagctatgattgcaccactgtactccaggctgggc	(((((((((((((((((((((((((((((((((((
actccagtctggGcaacaGagagagaccctgtctca	$((((((((((^((((L((((((((((((((((((

Looking for mutations

- Consed (Dave Gordon)
- Identifying reads discrepant from reference
- Sorting/prioritizing that list to identify variants for lab followup (perl scripts)



Gordon, D., C. Abajian, and P. Green. 1998. Consed: A Graphical Tool for Sequence Finishing. Genome Research. 8:195-202

Finding Variants (Discrepancies)

A		C		G		T		*		nt	
1	1.9%	42	80.8%r	8	15.4%	0	0.0%	1	1.9%	3712	gi|89106884|ref|AC_000091.1|
0	0.0%	1	1.4%r	0	0.0%	71	98.6%	0	0.0%	10,696	gi|89106884|ref|AC_000091.1|
0	0.0%	0	0.0%	75	83.3%r	15	16.7%	0	0.0%	11,143	gi|89106884|ref|AC_000091.1|
0	0.0%	1	1.3%	64	85.3%r	10	13.3%	0	0.0%	11,378	gi|89106884|ref|AC_000091.1|
0	0.0%	1	1.1%	79	85.9%r	12	13.0%	0	0.0%	11,956	gi|89106884|ref|AC_000091.1|
39	75.0%r	0	0.0%	13	25.0%	0	0.0%	0	0.0%	15,382	gi|89106884|ref|AC_000091.1|
1	2.0%	6	12.0%	39	78.0%r	2	4.0%	2	4.0%	16,839	gi|89106884|ref|AC_000091.1|
0	0.0%	18	23.1%	60	76.9%r	0	0.0%	0	0.0%	17,048	gi|89106884|ref|AC_000091.1|
61	75.3%r	19	23.5%	1	1.2%	0	0.0%	0	0.0%	17,050	gi|89106884|ref|AC_000091.1|
78	87.6%r	2	2.2%	6	6.7%	2	2.2%	1	1.1%	17,698	gi|89106884|ref|AC_000091.1|
2	2.2%	0	0.0%	81	89.0%r	8	8.8%	0	0.0%	18,614	gi|89106884|ref|AC_000091.1|
[illegible]											gi|89106884|ref|AC_000091.1|
0	0.0%	1	2.2%r	1	2.2%	43	95.6%	0	0.0%	23,501	gi|89106884|ref|AC_000091.1|
1	1.8%	7	12.3%	47	82.5%r	2	3.5%	0	0.0%	27,338	gi|89106884|ref|AC_000091.1|
10	8.3%	111	91.7%r	0	0.0%	0	0.0%	0	0.0%	29,486	gi|89106884|ref|AC_000091.1|
0	0.0%	11	8.2%	0	0.0%	123	91.8%r	0	0.0%	29,503	gi|89106884|ref|AC_000091.1|
61	85.9%r	8	11.3%	0	0.0%	1	1.4%	1	1.4%	29,901	gi|89106884|ref|AC_000091.1|
0	0.0%	1	1.3%r	0	0.0%	74	98.7%	0	0.0%	31,683	gi|89106884|ref|AC_000091.1|
63	84.0%r	11	14.7%	0	0.0%	1	1.3%	0	0.0%	35,049	gi|89106884|ref|AC_000091.1|
10	17.2%	0	0.0%	48	82.8%r	0	0.0%	0	0.0%	40,051	gi|89106884|ref|AC_000091.1|
0	0.0%	1	1.8%	10	17.5%	46	80.7%r	0	0.0%	40,719	gi|89106884|ref|AC_000091.1|
3	3.1%	87	89.7%r	7	7.2%	0	0.0%	0	0.0%	41,672	gi|89106884|ref|AC_000091.1|
132	91.0%r	5	3.4%	3	2.1%	5	3.4%	0	0.0%	41,891	gi|89106884|ref|AC_000091.1|
2	2.5%	3	3.8%	5	6.3%	69	87.3%r	0	0.0%	59,287	gi|89106884|ref|AC_000091.1|
4	3.3%	3	2.5%	6	5.0%	107	89.2%r	0	0.0%	59,516	gi|89106884|ref|AC_000091.1|
0	0.0%	1	1.7%r	0	0.0%	57	96.6%	1	1.7%	59,968	gi|89106884|ref|AC_000091.1|
49	83.1%r	9	15.3%	1	1.7%	0	0.0%	0	0.0%	63,942	gi|89106884|ref|AC_000091.1|
10	15.4%	0	0.0%	0	0.0%	55	84.6%r	0	0.0%	65,566	gi|89106884|ref|AC_000091.1|
0	0.0%	0	0.0%	69	87.3%r	9	11.4%	1	1.3%	70,602	gi|89106884|ref|AC_000091.1|
[illegible]											gi|89106884|ref|AC_000091.1|
59	98.3%	0	0.0%	1	1.7%r	0	0.0%	0	0.0%	76,249	gi|89106884|ref|AC_000091.1|
112	91.8%r	6	4.9%	1	0.8%	3	2.5%	0	0.0%	85,535	gi|89106884|ref|AC_000091.1|
[illegible]											gi|89106884|ref|AC_000091.1|

NCBI

NATIONAL INSTITUTES OF HEALTH

Perl Script for parsing and variant detection

PE_Simmulate.perl | discrepantparser_between.perl

```
5
6   $infile = shift @ARGV;
7
8   $outfile = $infile . '.Betweenoutput';
9
10  print STDOUT "$infile \t $outfile";
11
12  open OUTFILE, ">$outfile";
13  open INFILE, "<$infile";
14  while (<INFILE>) {
15      chomp;
16
17  #   print "$_";
18      @discline = split(/\s+/,$_);
19      if ($discline[0] eq '') {
20          $t = shift @discline;
21          }
22
23      $dpos= $discline[10];
24
25  #   positions ACGT*
26      $a_inf = $discline[1];
27      $c_inf = $discline[3];
28      $g_inf = $discline[5];
29      $t_inf = $discline[7];
30      $x_inf = $discline[9];
31      print "D-position= $dpos\n";
32
33      %ref=();
34  # stores the letter of the reference base
35
36      # where is the 'r' - this is the reference.  I want to find where reference is < 5% and some other
37      # base is 90%
38      ($a_pct, $ref{a}) = split(/\%/,$a_inf,2);
39      ($c_pct, $ref{c}) = split(/\%/,$c_inf,2);
40      ($g_pct, $ref{g}) = split(/\%/,$g_inf,2);
41      ($t_pct, $ref{t}) = split(/\%/,$t_inf,2);
42      ($x_pct, $ref{x}) = split(/\%/,$x_inf,2);
43
44  #   print "$a pct\n";
```

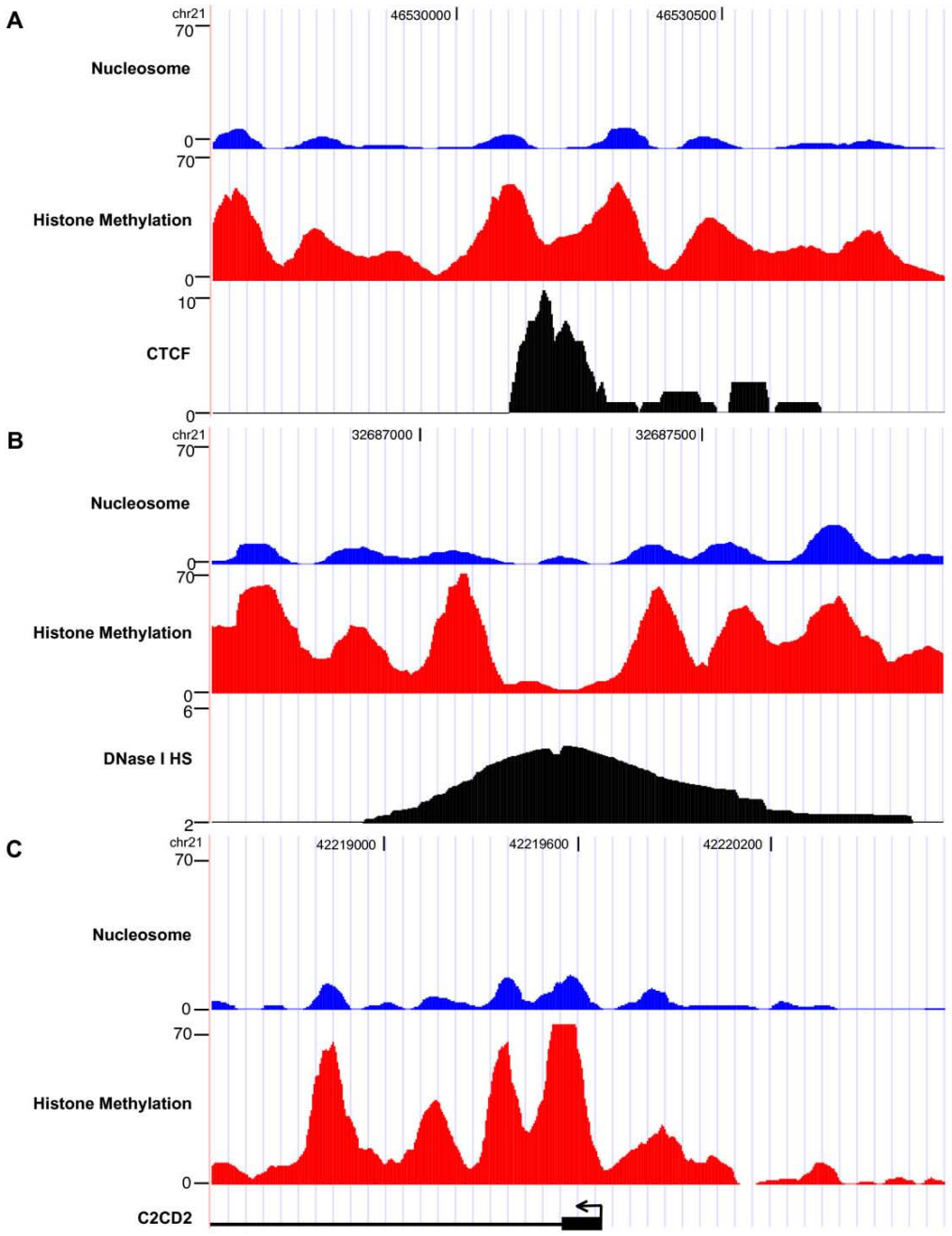
Perl source file | nb char : 1689 | Ln : 8 Col : 31 Sel : 0 | Dos\Windows | ANSI

Tabulating and Prioritizing Variations

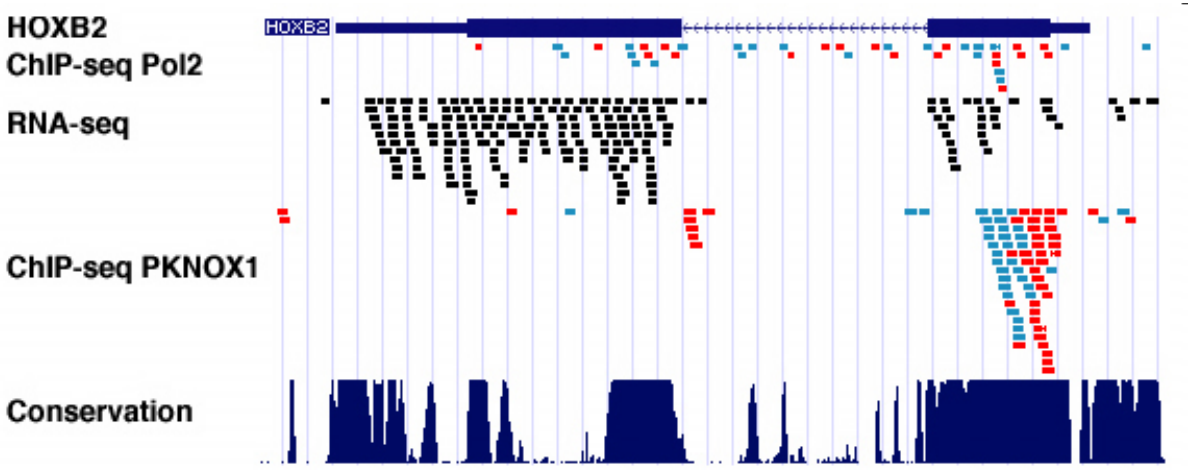
- Identify variants within or near exons
- Identify if these are already known in dbSNP
- Determining whether synonymous or non-synonymous changes at protein level

Risk SNP	consed_pos	genomic _pos	hit_info	SNPbase	Refbase	AA_change
Possible	1457478	51167999	CDSHIT	SNPbase=T	Ref=C	non-synonymous: PRO->LEU (cct->ctt)
No	2988712	52699233	CDSHIT	SNPbase=A	Ref=G	ASP ((gac -> gat) synonymous)
?	5263342	54973863	CDSHIT	SNPbase=T	Ref=C	
	5263910	54974431	CDSHIT	SNPbase=T	Ref=C	
Possible	5263910	54974431	CDSHIT	SNPbase=T	Ref=C	non-synonymous: PRO->LEU (cct->ctt)
?	7326967	57037488	CDSHIT	SNPbase=C	Ref=T	
No	7926212	57636733	CDSHIT	SNPbase=A	Ref=G	LEU ((ctg -> cta) synonymous)
No	15165899	64876420	CDSHIT	SNPbase=G	Ref=C	THR ((acc -> acg) synonymous)
No	15165905	64876426	CDSHIT	SNPbase=G	Ref=C	ARG ((cgc -> cgg) synonymous)
?	15165938	64876459	CDSHIT	SNPbase=C	Ref=T	
?	15165942	64876463	CDSHIT	SNPbase=C	Ref=G	
	24018308	73728829	CDSHIT	SNPbase=C	Ref=G	
Possible	24018308	73728829	CDSHIT	SNPbase=C	Ref=G	non-synonymous: THR->ARG (aca->agg)
	24700674	74411195	rs4892396-G/T	Ref=T	C=0.0 G=66.7 T=33.3 *=0.0	
?	27400682	77111203	CDSHIT	SNPbase=G	Ref=C	

Data from ChIP Seq experiments



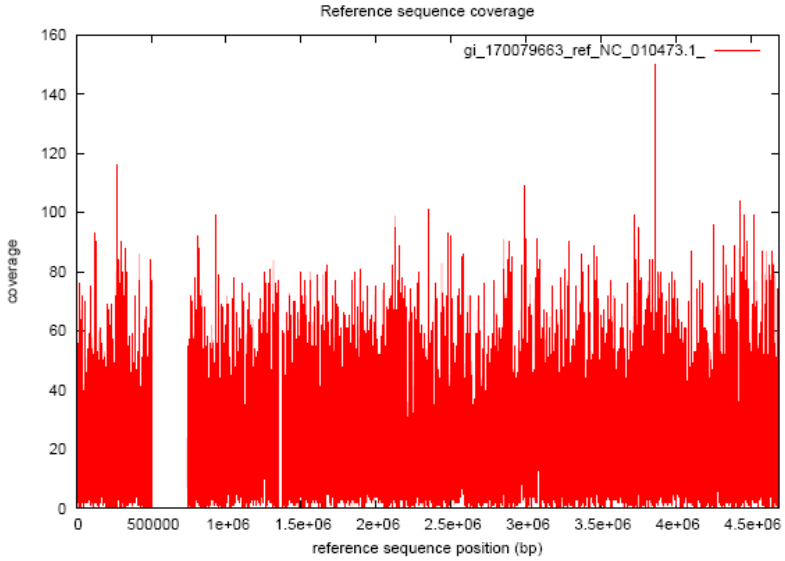
RNA Seq Results



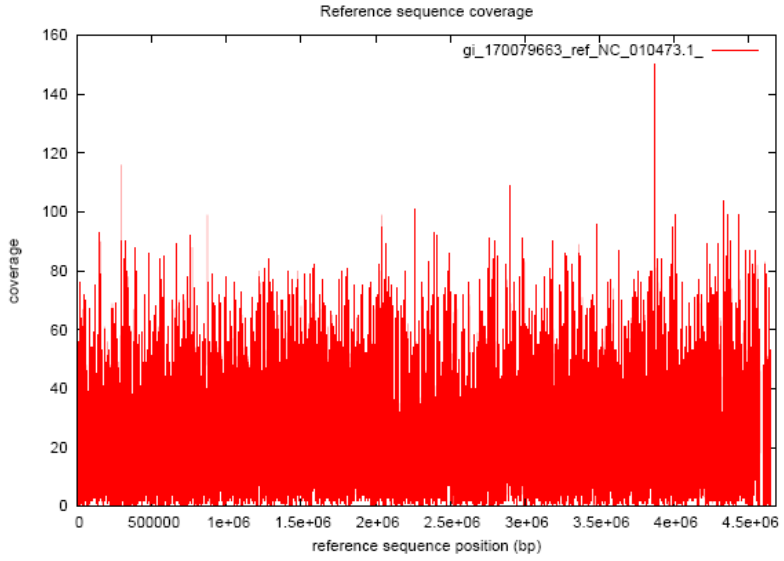
ChIP-seq and RNA-seq data exemplified at the HOXB2 gene

Single lane of Solexa for E. coli Genome (Genome Resequencing)

DH10B



W3110



Conclusion

- Technology is available to rapidly change the application of DNA sequencing to biology questions without need to be a sequencing center
- Bioinformatics challenged to keep up and develop robust methods as the technology is rapidly changing and improving

- *3rd-Gen sequencing is less that 1 year away.*