Beyond the Genome 2013

Informatics Challenge

One possible solution

Michael Schatz
Sven-Eric Schelhorn, October 2013

There are many possible solutions –
this is one.

Approach

We want to obtain subsequences of several genomes based on fragmented read data.

Since we do not know anything about the insert, we cannot use targeted approaches and have to *assemble* de-novo

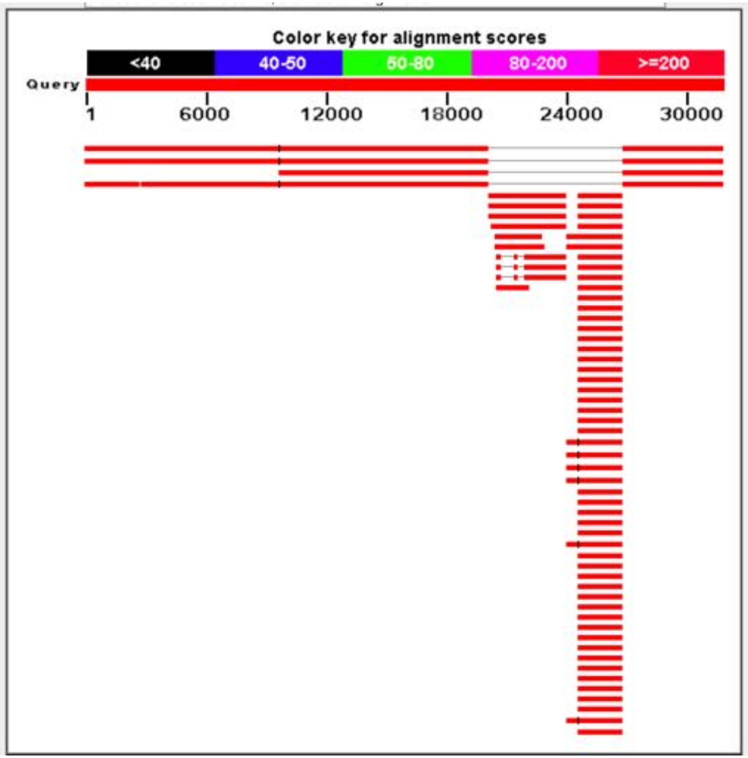
Using Ray

```
$ gunzip  *.gz
$ Ray -p sh_end_1.fastq sh_end_2.fastq -p \
lo_end_1.fastq lo_end_2.fastq -k 21 -o rayout
```

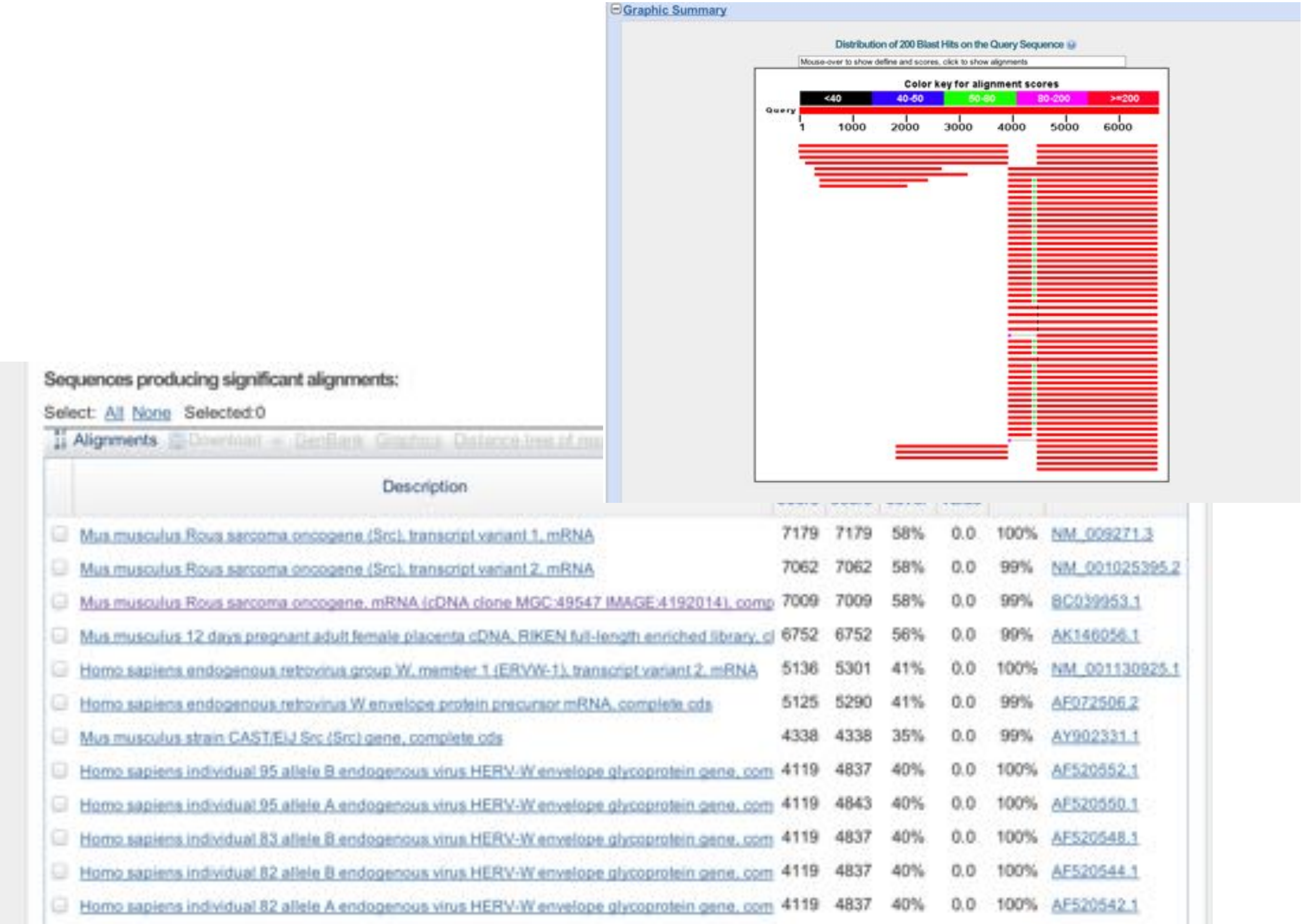
Results are available in rayout/Scaffolds.fa

Identify the insert with BLAST

Blasting this data immediately shows you the species (three giant viruses)
As well as the insert consisting of src *and* syncti*n*



Identify the wildtype insert with BLAST


Graphic Summary
Distribution of 200 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
1000
2000
3000
4000
5000
6000
Alignments
Download
GenBank
Graphics
Distance tree of res

Sequences producing significant alignments:

Select: All None Selected:0

	Description						
□	Mus musculus Rous sarcoma oncogene (Src), transcript variant 1, mRNA	7179	7179	58%	0.0	100%	NM_009271.3
□	Mus musculus Rous sarcoma oncogene (Src), transcript variant 2, mRNA	7062	7062	58%	0.0	99%	NM_001025395.2
□	Mus musculus Rous sarcoma oncogene, mRNA (cDNA clone MGC:49547 IMAGE:4192014), comp	7009	7009	58%	0.0	99%	BC039953.1
□	Mus musculus 12 days pregnant adult female placenta cDNA, RIKEN full-length enriched library, cl	6752	6752	56%	0.0	99%	AK146056.1
□	Homo sapiens endogenous retrovirus group W, member 1 (ERVW-1), transcript variant 2, mRNA	5136	5301	41%	0.0	100%	NM_001130925.1
□	Homo sapiens endogenous retrovirus W envelope protein precursor mRNA, complete cds	5125	5290	41%	0.0	99%	AF072506.2
□	Mus musculus strain CAST/EiJ Src (Src) gene, complete cds	4338	4338	35%	0.0	99%	AY902331.1
□	Homo sapiens individual 95 allele B endogenous virus HERV-W envelope glycoprotein gene, com	4119	4837	40%	0.0	100%	AF520552.1
□	Homo sapiens individual 95 allele A endogenous virus HERV-W envelope glycoprotein gene, com	4119	4843	40%	0.0	100%	AF520550.1
□	Homo sapiens individual 83 allele B endogenous virus HERV-W envelope glycoprotein gene, com	4119	4837	40%	0.0	100%	AF520548.1
□	Homo sapiens individual 82 allele B endogenous virus HERV-W envelope glycoprotein gene, com	4119	4837	40%	0.0	100%	AF520544.1
□	Homo sapiens individual 82 allele A endogenous virus HERV-W envelope glycoprotein gene, com	4119	4837	40%	0.0	100%	AF520542.1

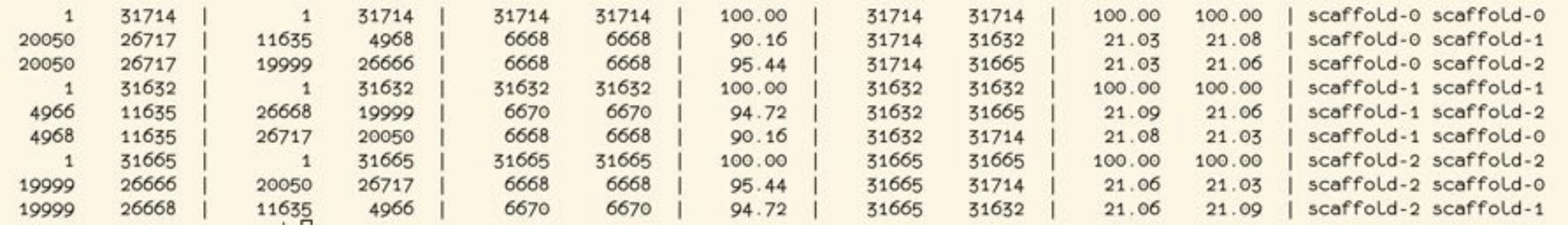
Identify the insert with Nucmer

*The one thing we know about the insert is that it is present in **all** the three viruses.*

*So let's find common regions within the contigs using **mummer/nucmer***

Identify the insert with Nucmer

```
$ nucmer -maxmatch rayout/Scaffolds.fasta
rayout/Scaffolds.fasta -p nucout
$ show-coords -Hrcl nucout.delta
```



```
    1    31714  |      1    31714  |  31714    31714  |  100.00  |  31714    31714  |  100.00   100.00  | scaffold-0 scaffold-0
20050    26717  |  11635     4968  |   6668     6668  |   90.16  |  31714    31632  |   21.03    21.08  | scaffold-0 scaffold-1
20050    26717  |  19999    26666  |   6668     6668  |   95.44  |  31714    31665  |   21.03    21.06  | scaffold-0 scaffold-2
    1    31632  |      1    31632  |  31632    31632  |  100.00  |  31632    31632  |  100.00   100.00  | scaffold-1 scaffold-1
 4966    11635  |  26668    19999  |   6670     6670  |   94.72  |  31632    31665  |   21.09    21.06  | scaffold-1 scaffold-2
 4968    11635  |  26717    20050  |   6668     6668  |   90.16  |  31632    31714  |   21.08    21.03  | scaffold-1 scaffold-0
    1    31665  |      1    31665  |  31665    31665  |  100.00  |  31665    31665  |  100.00   100.00  | scaffold-2 scaffold-2
19999    26666  |  20050    26717  |   6668     6668  |   95.44  |  31665    31714  |   21.06    21.03  | scaffold-2 scaffold-0
19999    26668  |  11635     4966  |   6670     6670  |   94.72  |  31665    31632  |   21.06    21.09  | scaffold-2 scaffold-1
```

Based on pairwise similarity, scaffold 2 seems to be the wildtype
This information can also be obtained by BLAST.

Let’s extract the insert regions

```
$ samtools faidx rayout/Scaffolds.fasta

$ samtools faidx rayout/Scaffolds.fasta
scaffold-2:19999-26668 > wildtype.fasta

$ samtools faidx rayout/Scaffolds.fasta
scaffold-1:4966-11635 > variant1.fasta

$ samtools faidx rayout/Scaffolds.fasta
scaffold-0:20050-26717 > variant2.fasta
```

Align the regions and call the variants

```
$ nucmer -maxmatch wildtype.fasta variant1.fasta -p var1 >& /dev/null
$ show-snps -H var1.delta | awk '{print $3}' | ./dna-encode.pl –d
$ furywithwhichhecreatedthiscomplexpieceofcode

$ nucmer -maxmatch syntenic_region.fasta variant2.fasta -p var1 >& /dev/null
$ show-snps -H var1.delta | awk '{print $3}' | ./dna-encode.pl –d
$ Hehadtoicehiswristsatnightbecauseofthe
```

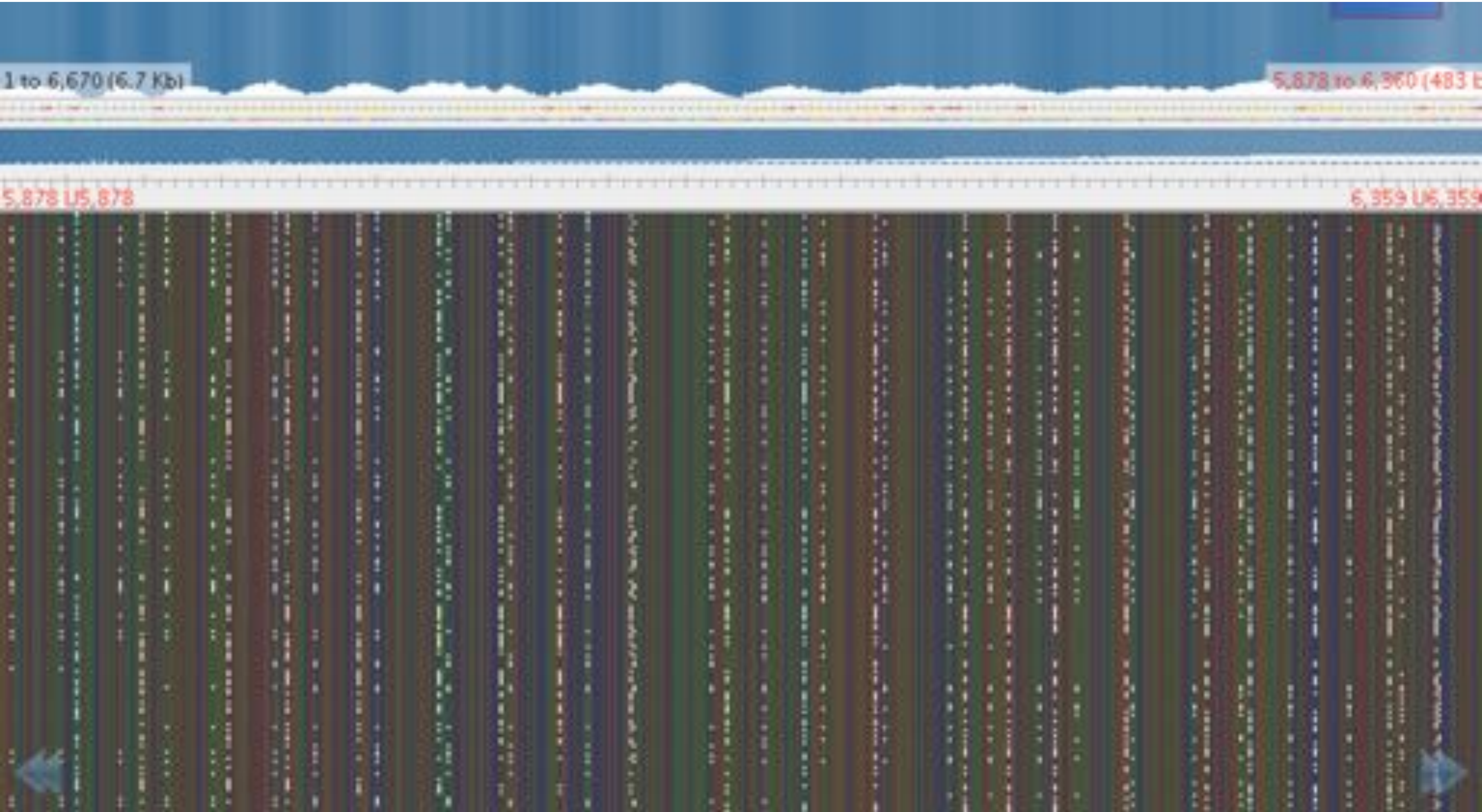
Solution

"He had to ice his wrists at night because of the fury with which he created this [extraordinarily] complex piece of code."

*--David Haussler** (*on* Jim Kent, the grad student who created GigAssembler and his crucial role in assembling the first ~70%of the human reference

** See keynote*

One alternative



Call variants from pileup

```
$ bwa index wildtype.fasta
$ bwa mem wildtype.fasta sh_end_1.fastq sh_end_2.fastq | samtools
view -Su - > sh.bam
$ bwa mem wildtype.fasta lo_end_1.fastq lo_end_2.fastq | samtools
view -Su - > lo.bam
$ samtools merge -f merged.bam sh.bam lo.bam
$ samtools sort merged.bam merged.sorted
$ samtools index merged.sorted.bam
$ samtools faidx wildtype.fasta
$ samtools mpileup -f wildtype.fasta merged.sorted.bam >
syntenic_region.pileup
```

<parse pileup, separate variants by frequency and feed them into dna-encode.pl; cone be done in ~10 lines of code >

The winners

Congratulation to *Jacob Kitzman*
for being the first conference attendee to send in the solution

and *Shaun Jackman (author of Abyss)*
for being the first to solve the problem on twitter

and all other participants for taking part!