The Software and IT Needed to Deliver NGS Diagnostics

Graham Taylor

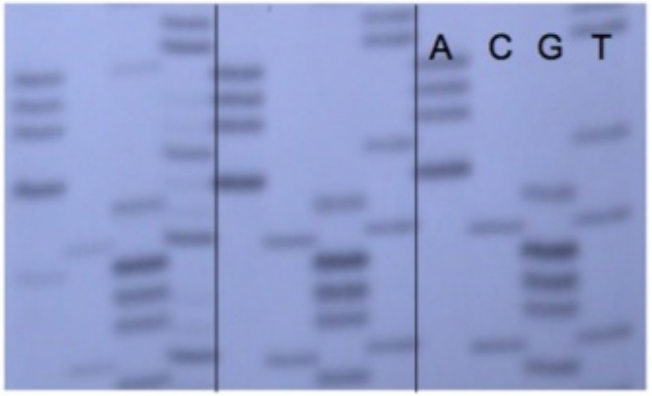
Leeds

Why is it an issue?

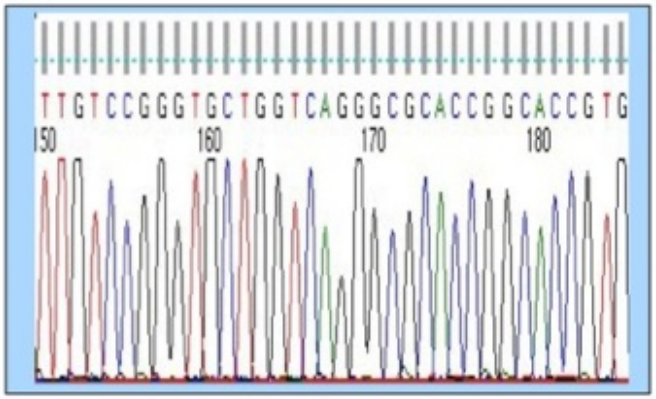
- Data volume changes with NGS
- Infrastructure
  - Hardware
  - Network
  - Software
- Solutions in operation in the Leeds medical/ academic centre
- Other models/solutions

Data volume changes with NGS

Last Gen



Current Gen



10^1 sequences per run

10^3 bases per run

10^2 sequences per run

10^4 bases per run

Next Gen

```
HWI-EAS318     11   2     17    453  1023 0     2
ATGAAGATGATGCTGGAGAGCCGGTTCCGCGTGCCCATCCAGCTGCTGTCC
a1VZ1V1ZVVFMT_BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
hs_ref_chr7.fa          288283    F
16T1C15T5T1A2G3T1 0     56              22    R
```

X 10^8

10^9 -10^{10} bases per run

Million-fold increase in data output

What does this mean for clinical molecular genetic diagnostics?

The currency of genetic analysis will become DNA sequence

- Sequence count
- Sequence variation
- Sequence arrangement

The currency is dropping to commodity prices

Skill set required is less lab, more informatics biased

Space and Computation

- One sequencing run with Illumina produces about 0.5 TB of data.
- Most of this information is “technical” and is usually not required for analysis (images…).
- The most “upstream” data are sequences + read quality. Approx 1 GB of data per lane (8 lanes per run) when compressed.
- From there we can always re-compute derived data (alignments, counts…)

Space and Computation

- If we perform one run every 4 days (maximum capacity of one machine) we will produce about 720 GB of “upstream” data per year. This data need to be **backed up and kept safe** (off site, duplicated) and accessible from different users.
- **FILE SERVER**
- Extra space is required (2 to 10 times the raw data?) for intermediate data, databases, reference genomes, results. No need to thorough backup if easily reproducible. Backup the scripts and programs ;)

Space and Computation

- At the moment alignment runs on an 8 core server & I can analyze data on my workstation.
- Sooner or later, a dedicated computer/cluster will be required (time consuming analysis and/or lot of input data)
- HPC system in Uni of Leeds (Arc1) can provide 84 servers with 12 GB(!) of RAM. Total of 672 processors(!). Plus 4 servers with 128 GB of RAM
- When analyzing big files, the bottleneck is accessing those files.

Accessing data

Me

10 MB/s

File Server

Prototyping and testing server (?)

You

About 3 days to copy 100 GB of data!!!

10 MB/s

HPC Leeds Uni (Arc1)

Anybody

Infrastructure: Hardware

- Servers: each run on an Illumina GAIIx generates
  - 16 Gbytes compressed qseq or FASTQ files
  - 0.5 Gbytes compressed SAM files
- Image files are larger but no longer need to be stored
- Servers use Centos or RedHat Unix with Perl, Python and C programs
- Typically 8 core processor, 32 Gbyes RAM and 7 Tb drive is sufficient for one or two machines
- Not your typical hospital server
- Leeds solution: 3 servers: one live for sequencer, one for research use, one to hold data prior to export

- Recommendations
- Linux/Unix server with hardware RAID5
- Mirror server identical to above for backup
- Long term data storage to tape backup – IT facilities ?
- External RAID enclosures for extending data capacity when needed
- Possible SSD for fast access to heavily used data
- SATA disks as SAS disks too expensive and performance increase small

• Costs

- Linux Server with 16Tb Hardware RAID 5 - £3400
    - 2 x Quad Core Xeon 3u rack mounted ~£2000
    - 1 x Hardware RAID card ~£300
    - 8 x 2Tb SATA disks ~£1100
- External SATA RAID enclosure 16Tb - ~£3000
- NAS system 16Tb - £4900
- *(All prices from "trusted" suppliers, ie Rackservers.com & CCL)*

Infrastructure: Network

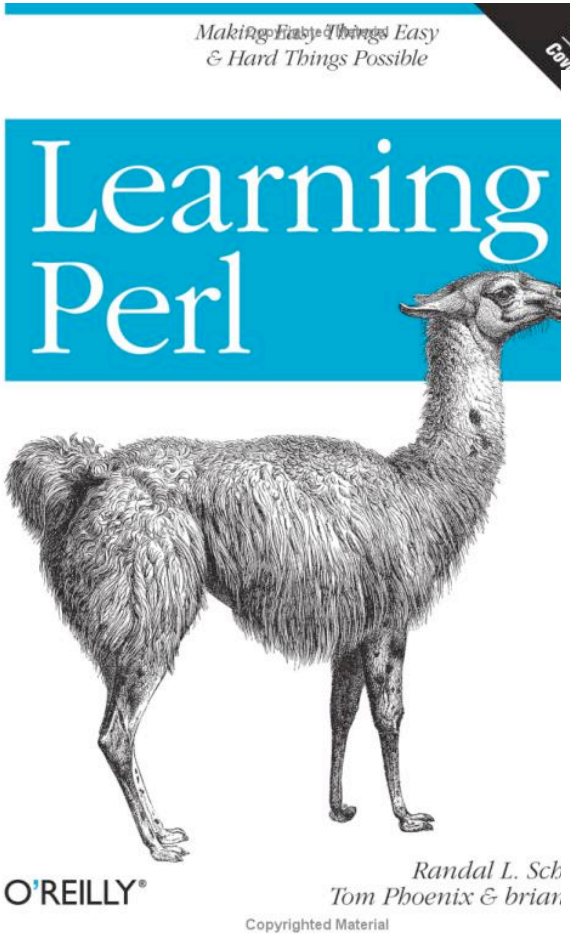
- Gbit ethernet and switching from machine to server
- Offload data to users to avoid data storage problems
- Need sftp or ssh plus firewall permissions between systems (e.g. University/Hospital)
- Often easier to transfer by removable disk!

Infrastructure: Software

- Mix of operating systems, hand written scripts and commercial software
- Regular updates on all requires on site skills to install and run

Handling the data

```perl
#!/usr/bin/perl -w
use strict;
use warnings;

my $regex;
my $count_regex;
my $errors;
my $dna_filename;
my $DNA;
my $location;
my $revcomp;
my $rcregex; # revcomp of search sequ
my $count_rcregex ;

# Get the DNA sequence data
#print "Please type the filename of t

$dna_filename = "BRCA1_BRCA2.txt";

chomp $dna_filename;

# Does the file exist?
unless ( -e $dna_filename) {

    print "File \"$dna_filename\" doe
    exit;
}

# Can we open the file?
unless ( open(DNAFILE, $dna_filename)

    print "Cannot open file \"$dna_fi
    exit;
}

$DNA = <DNAFILE>;

close DNAFILE;

# Remove whitespace
$DNA =~ s/\s//g;

#print "please enter regular expressi

$regex= <STDIN>;

chomp $regex ;

$rcregex = $regex ;

$rcregex = reverse $rcregex ;

$rcregex =~ tr/ACGT/TGCA/ ; # $rcregex is now rev comp of $regex

```




```perl
use strict;
use warnings;

# BRCA1_intro
my $filename

my $seq;

my $match

my @matches

my $firstline
my $line;

my $FNAFILE;

open ($FNAFIL
      or die

# read and ig
if ( !($first
{
    print "Co
    exit();
}

$seq = "";

# WHILE not E
while ( $line
{
        chomp
"chop" (remov

#       $line

        $line

#       $line

        $seq

} # end WHILE

close($FNAFIL

#       chomp

#chomp $first
```

```
AACCATGC        70113074
AACCATGC        70120370
AACCATGC        70260735
AACCATGC        70309147
AACCATGC        70537135
AACCATGC        70548630
AACCATGC        70657740
```

BRCA2_primers

```
errors  0
CAAGGGACAGAAATATCAGGTCTAA       18997952
total   0
revcomp 1
errors  0
ATGTTGGATGTGCCTCTAATCC  19003061
total   1
revcomp 0
errors  0
GTAAGCCTCTTTGAACCTCTGATT        19006948
total   0
revcomp 1
errors  0
GATTACTGGGTGCAAAGGTAAATG        19021022
total   1
revcomp 0
errors  0
GCATGGAGAAGGTCTTAAATCAGA        19026832
total   0
revcomp 1
errors  0
CTCCTGAGTAACTGGCACCAC   18962679
total   1
revcomp 0
errors  0
GCATGGAGAAGGTCTTAAATCAGA        19026832
total   0
revcomp 1
errors  0
```

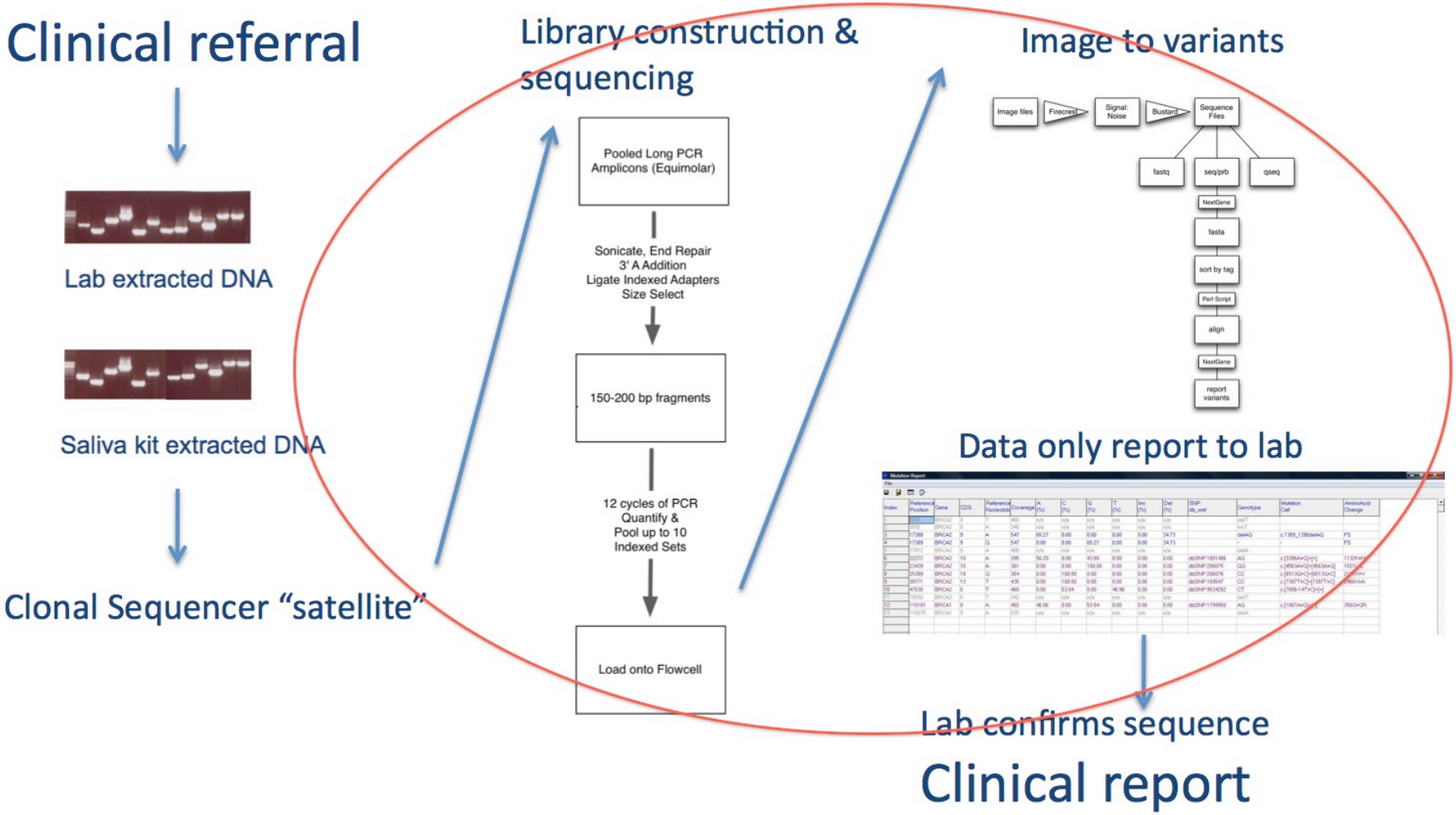
```
AACCATGC        72520283
AACCATGC        72591617
AACCATGC        72610335
total   878
revcomp 875
errors  0
limm-mac1141:Toolkit grahamtaylor$ perl find_motif_BRCA1_BRCA2.pl
please enter regular expression to search for:GGGGGGGAATTTTT
GGGGGGGAATTTTT  47411788
total   1
revcomp 0
errors  0
```

Operations in Leeds

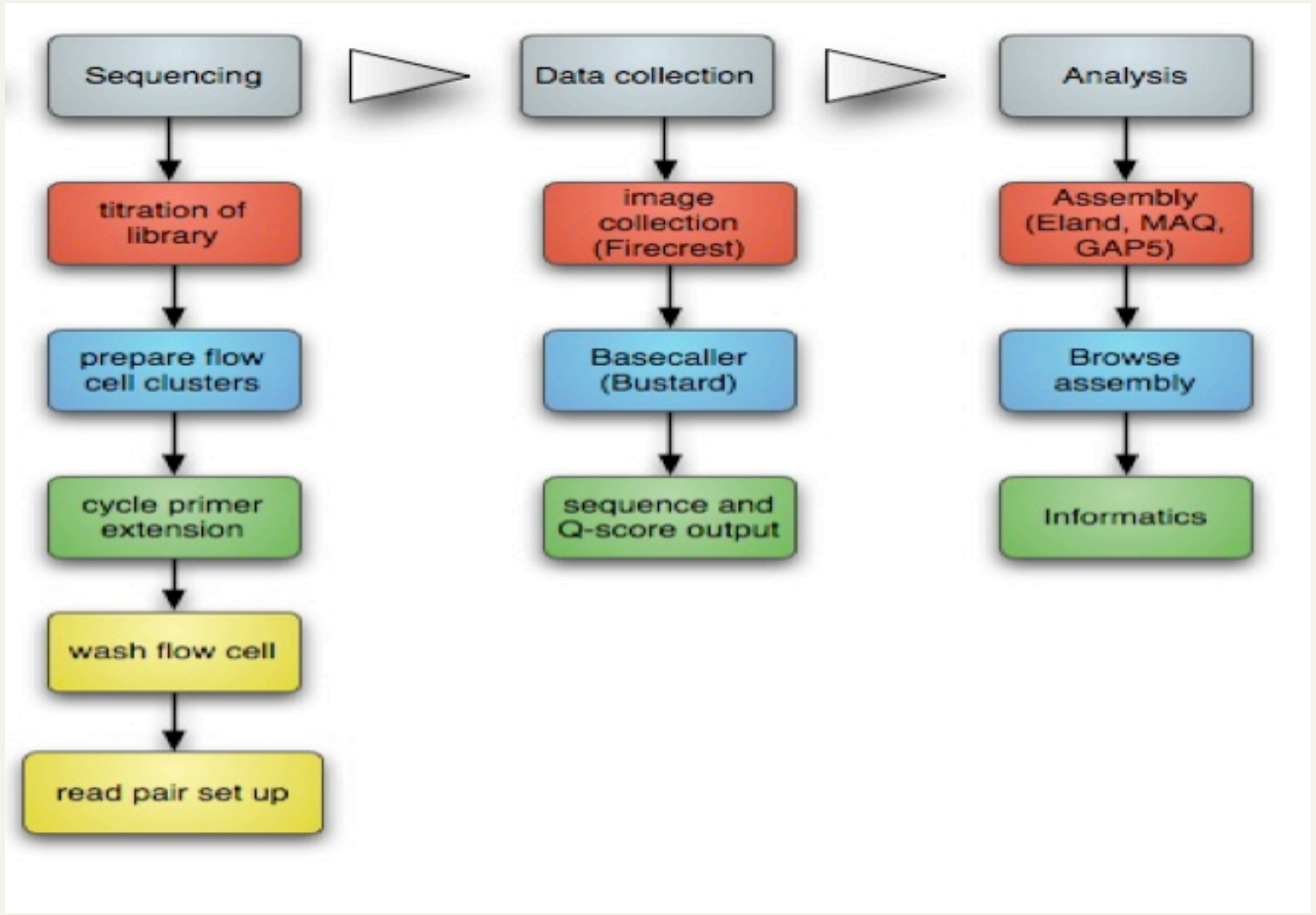
- Management
  - Sequencing “hotel” based around University IT infrastructure
- Confidentiality
  - Encoding of samples (as lab numbers)
  - All patient details remain inside NHS network
  - Build a tag index that converts tag to lab number
- Audit
  - Evaluating changes in process e.g. 10-20 sample batches
- Troubleshooting & assay development

BRCA1 & *BRCA2* Testing

CPA-approved & in service: contact Ruth Charlton for details)

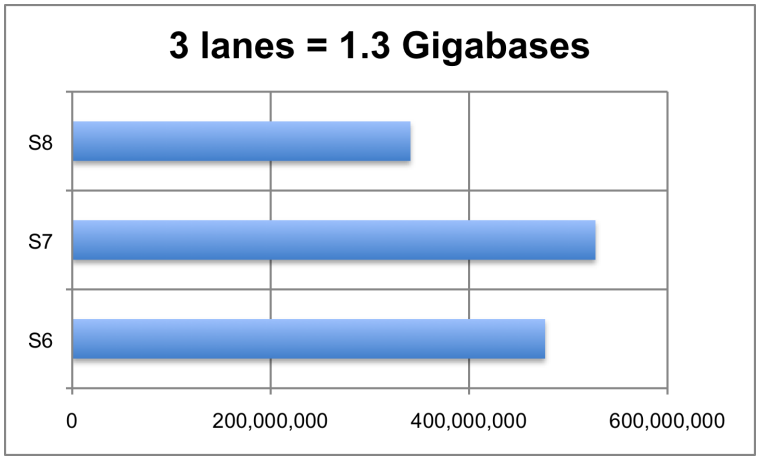


Workflow


Sequencing
titration of library
prepare flow cell clusters
cycle primer extension
wash flow cell
read pair set up
Data collection
image collection (Firecrest)
Basecaller (Bustard)
sequence and Q-score output
Analysis
Assembly (Eland, MAQ, GAP5)
Browse assembly
Informatics

6 base tags 51 base reads

(Spring 2009 stats: recent changes increase sequencing yield 5-fold)



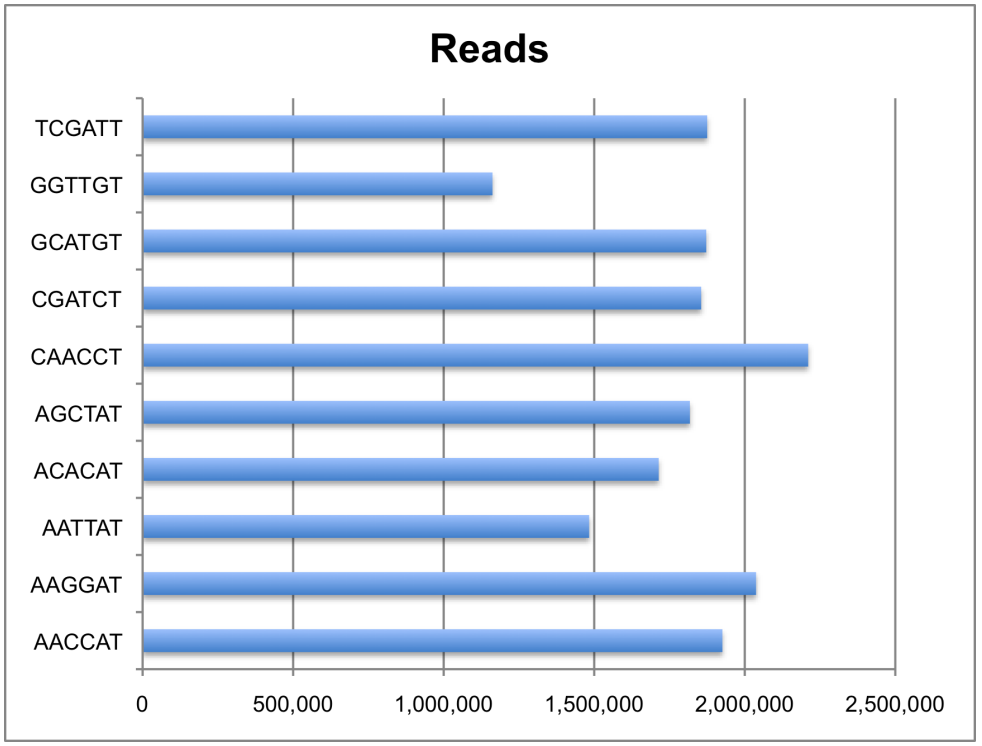


Image to Variants

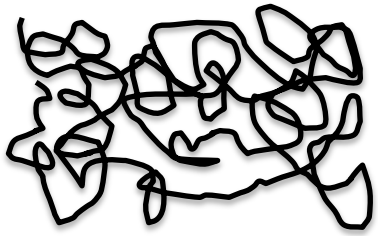
- Image to intensity/ position
- Intensity to base call with q scores (qseq)
- De-multiplex
  - grep, Perl, Python, NextGene, Novalign, Eland
- Qseq to aligner
  - BWA, Bowtie, Novalign (fastq)
  - NextGene, Illuminator (q-filtered fasta)
- Coverage: custom scripts onto eland “sorted” files, SAM files or NextGene coverage files to produce tables (Perl, Python & R) or bedgraph files.
- SAM files can be browsed directly in IGV
- Variant tables to database and reports

BRCA sequencing with 20 tags per lane

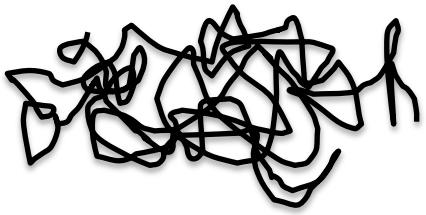
	Minimum Coverage			Position of Minimum Coverage		
	Forward	Reverse	Combined	Forward	Reverse	Combined
2010_0960	167	53	433	145763	79761	79761
2010_0976	140	54	335	99053	79756	79642,3 + 79646
2010_0983	173	70	476	106800	79754	79643-4
2010_0985	156	55	405	99048	79761	79761
2010_0989	87	35	251	123625	79761	79761
2010_0990	140	40	365	106779	79754-5	79672
2010_0991	177	60	580	106789	79755	79665, 79671
2010_1016	92	28	233	106778-80, 123595	79754	79630
2010_1051	173	43	416	106787-8	79761	79761
2010_0699	86	***21***	155	145706	79754	106891
2010_0707*	349	79	831	145706	79754	79752
2010_0798	108	***18***	167	106892	79754	106892
2010_0858	82	***16***	154	21350	79754	106892
2010_0860	78	***10***	142	106892	79753-5	106892
2010_0861	92	24	201	106892	79753-4	106892
2010_0862	50	***11***	137	106892	79754	106892
2010_0863	26	39	***69***	106890-2	79770-1	106892
2010_0900	30	45	***75***	106892	106892	106892
2010_0965	86	23	163	106892	79753,5	106892

Copy number by sequence read depth

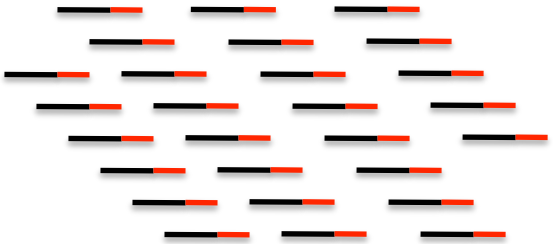
or – aCGH by next generation sequencing



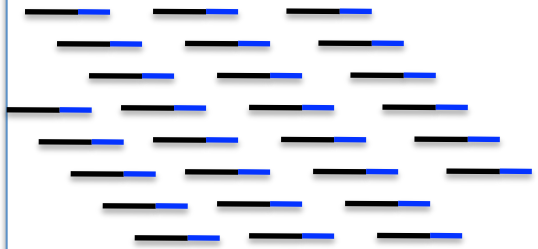
1 µg test genome



1 µg reference genome

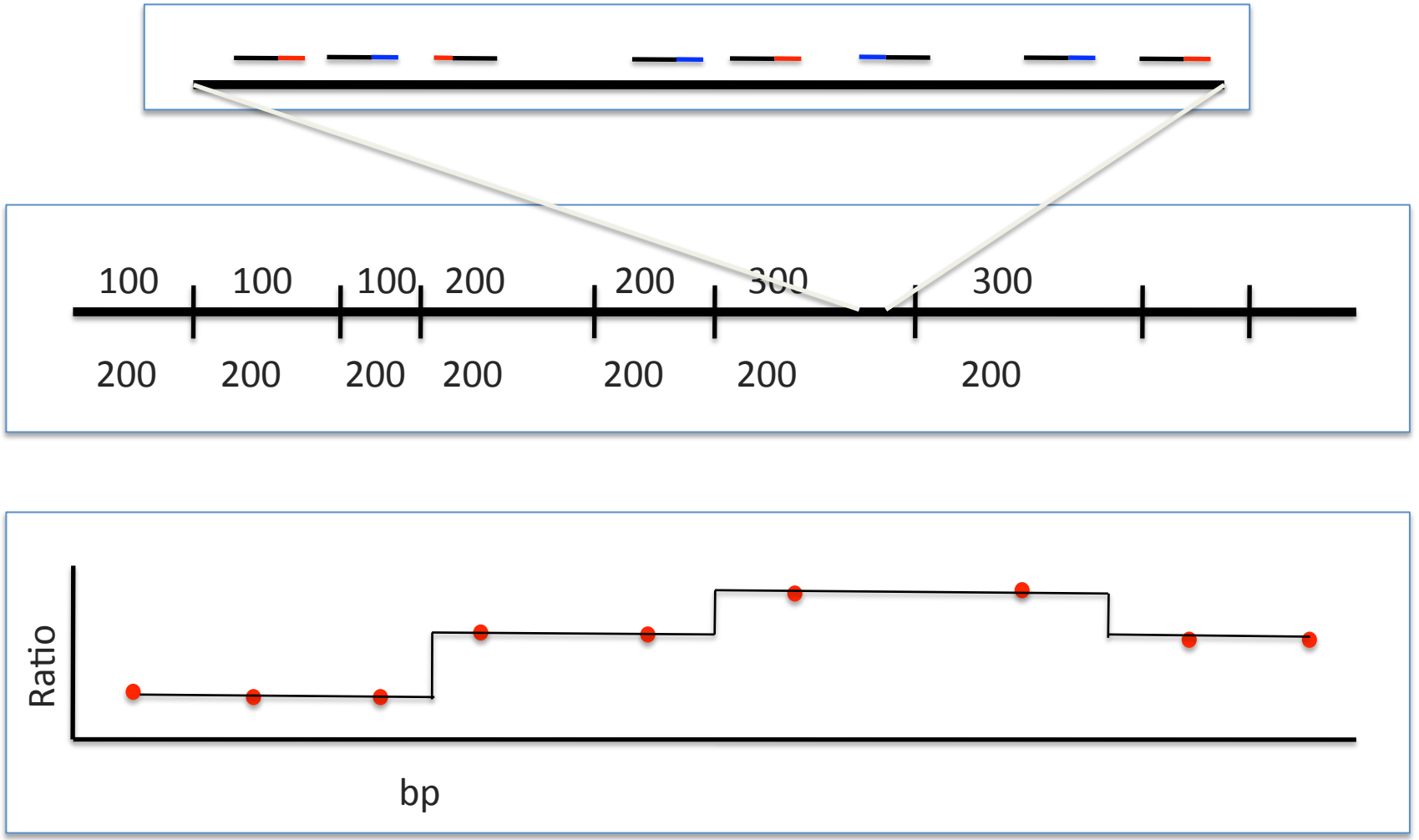


Library prep and tagging



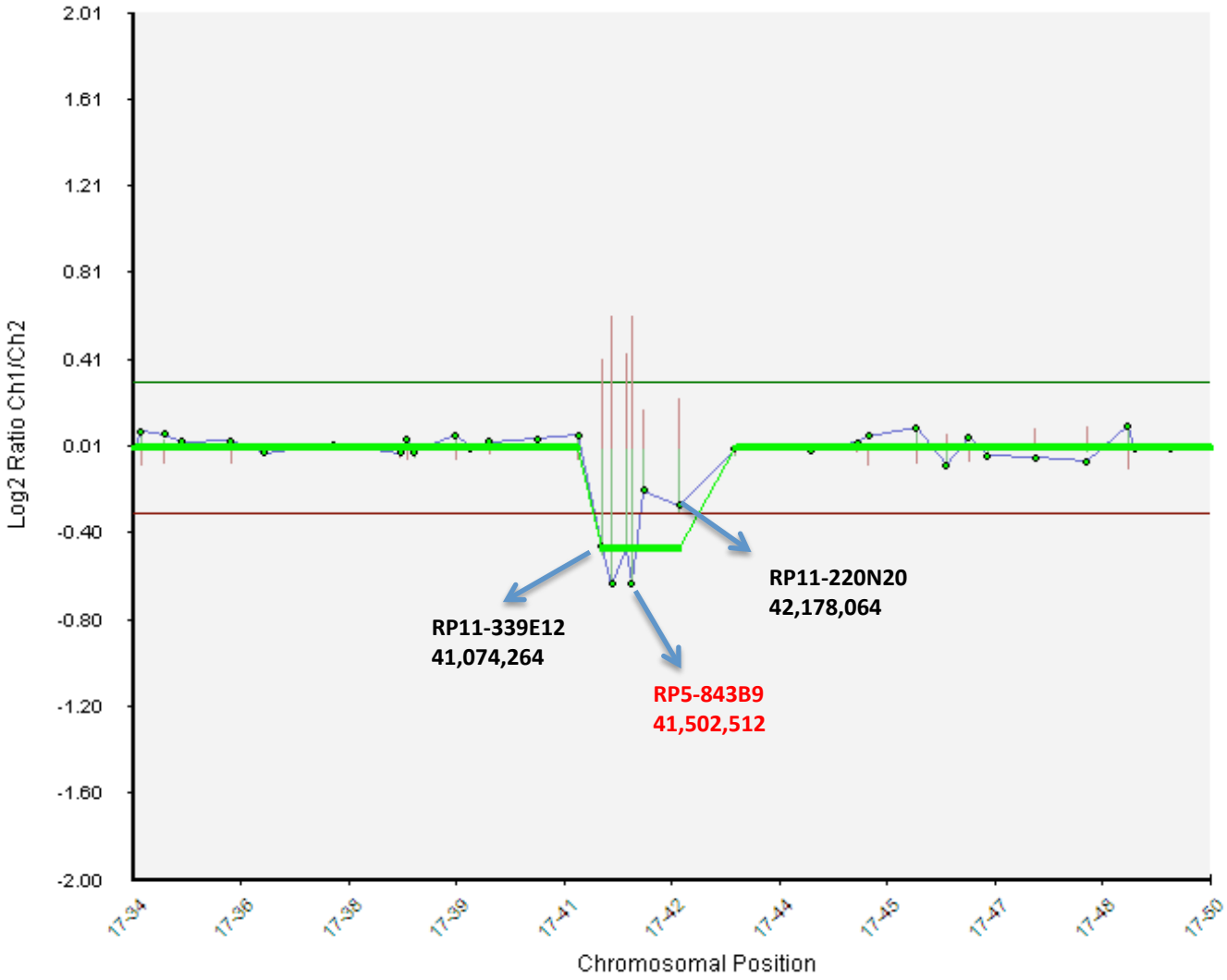
Pool, sequence and uniquely map to human genome

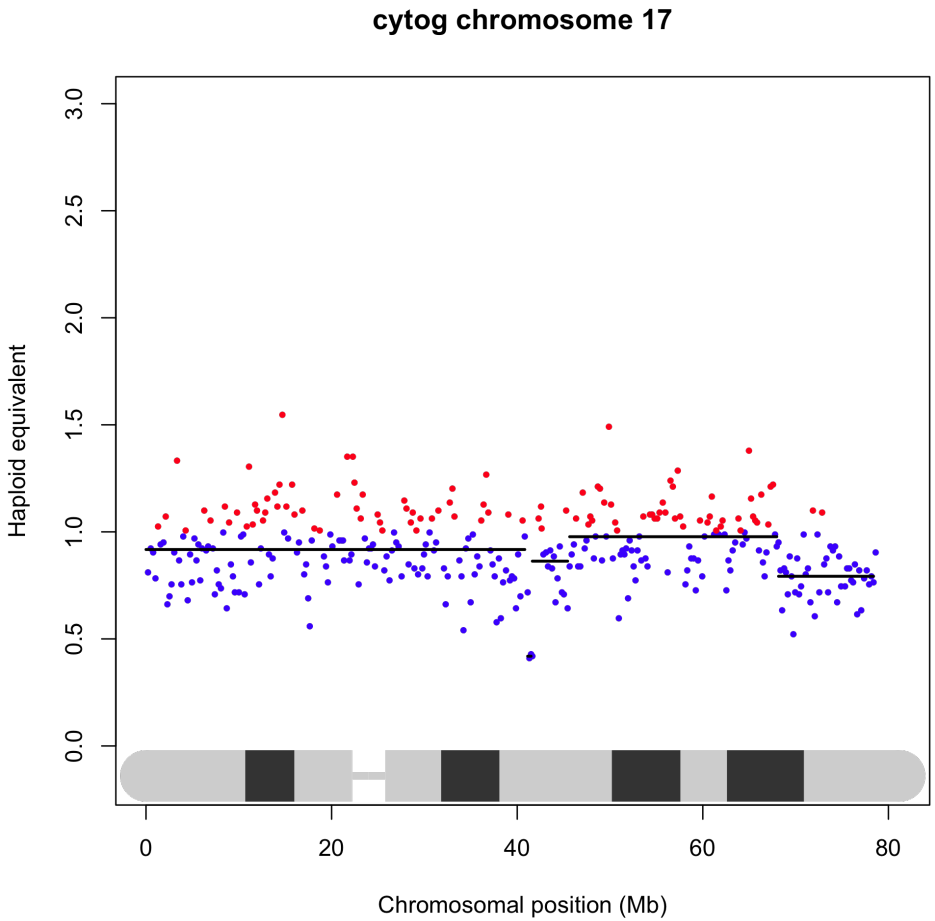
Work done using standard Illumina pipeline



Non – standard post sequencing analysis: splitting the genome into windows, counting reads in each window, calling breakpoints. Custom written scripts and aCGH based packages.

Abnormal case 2 – Del(17)(q21.31q21.32)





Sequencing – sample2 vs LS043-LTN (50 read window)

Results with 3 different controls (50 read window)

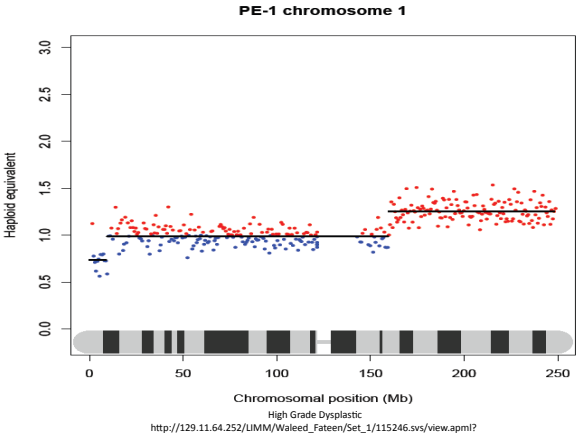
Control	Breakpoints
aglcl	41,142,573 – 41,758,121
LS043-LTN	41,128,557 – 41,631,854
LS035-BC	41,012,833 – 41,976,225

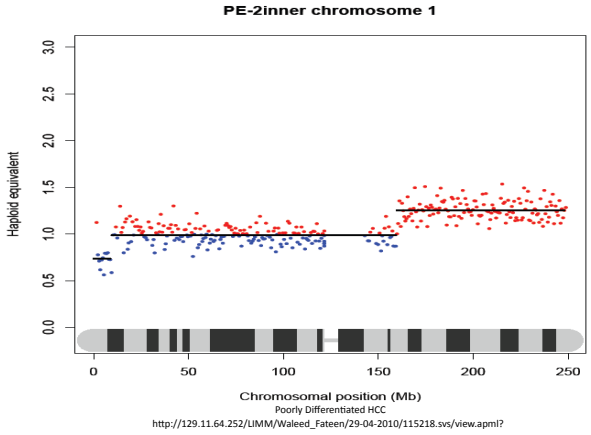
Summary of results (sample 2)		
	Breakpoints	Size of abnormality
BACS	41Mb – 42.1Mb	~1,1Mb
Sequencer	41.1Mb – 41.8Mb	<1Mb

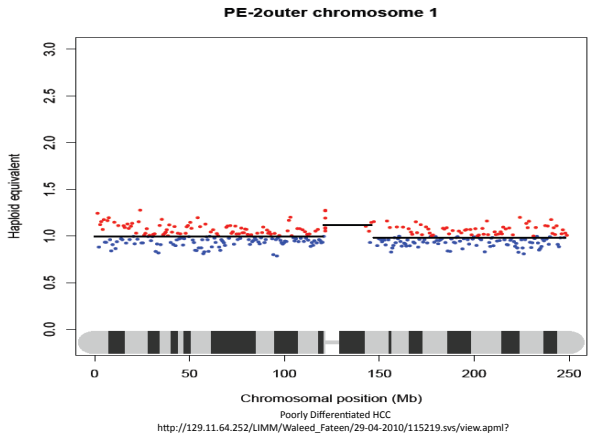
Multiplexing and effect on resolution

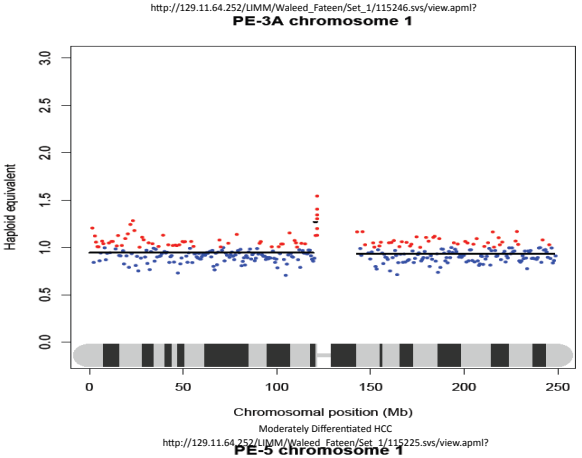
Samples per lane (cost)	Number of reads	Mean distance between reads	Mean size of 200 read window
2 (£500)	5 million	620bp	123Kb
5 (£200)	2 million	1165bp	233Kb
10 (£100)	1 million	3062bp	612Kb
80 (£12)	125,000	27Kb	5.5Mb

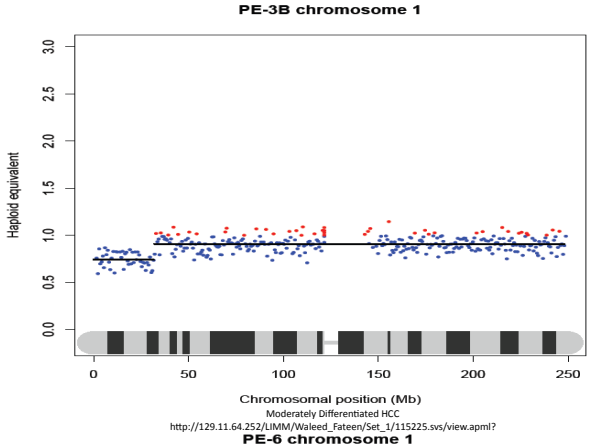
Chromosome 1

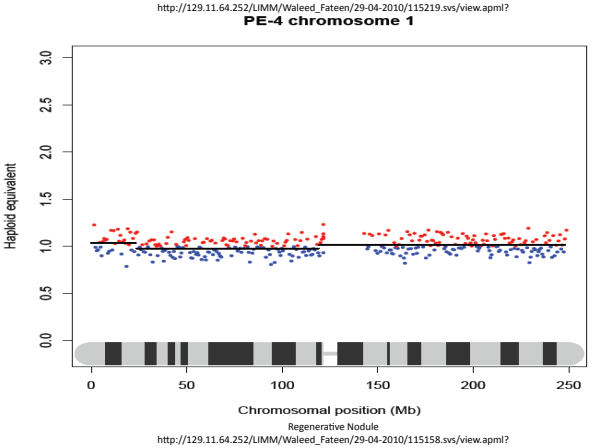


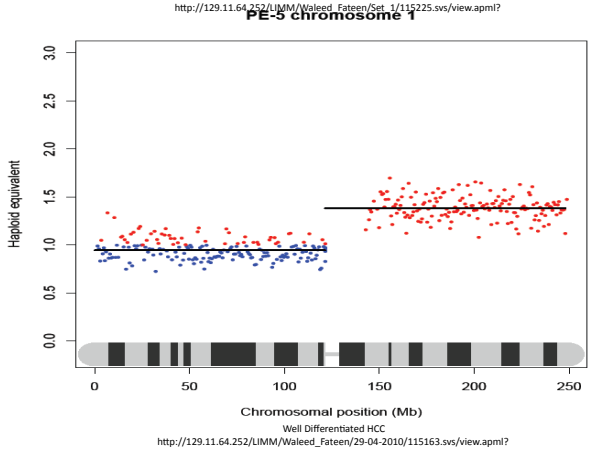


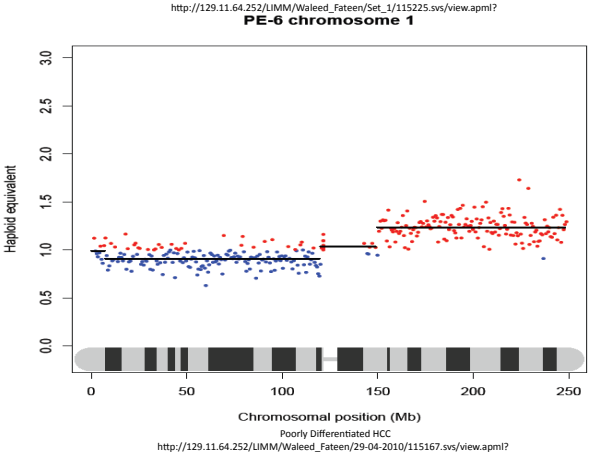


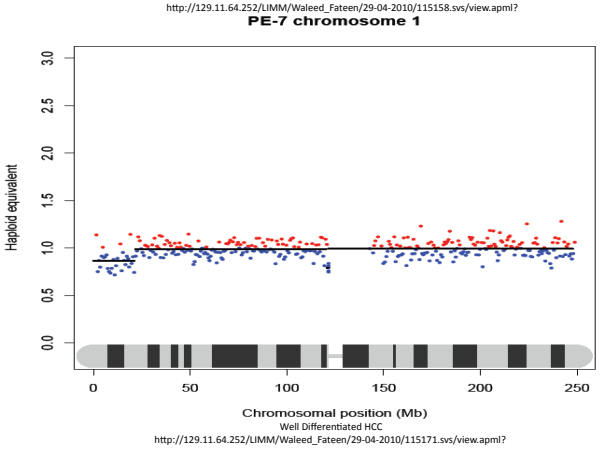




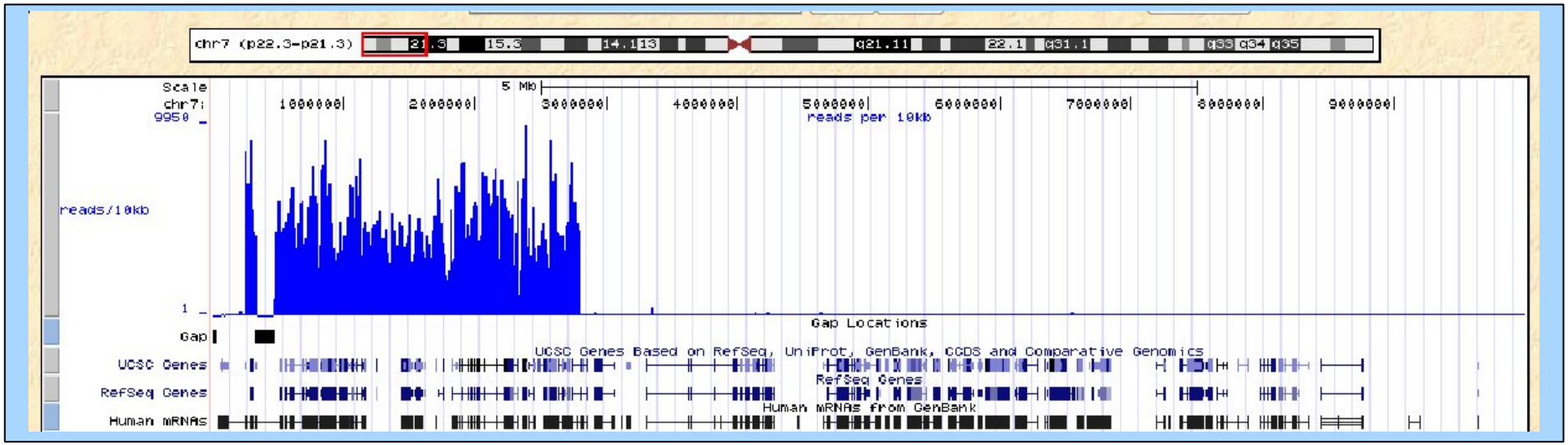








Chr7p target region enrichment

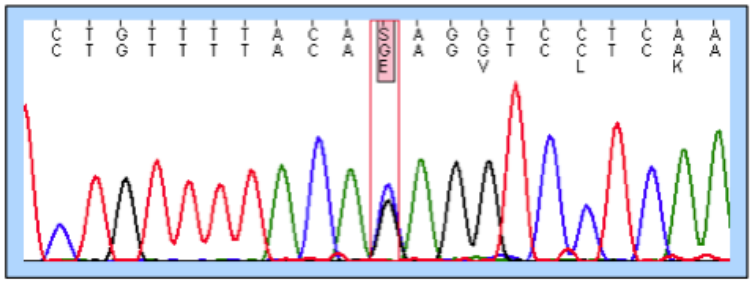


17M reads aligned to genome, 2.2M aligned to target region

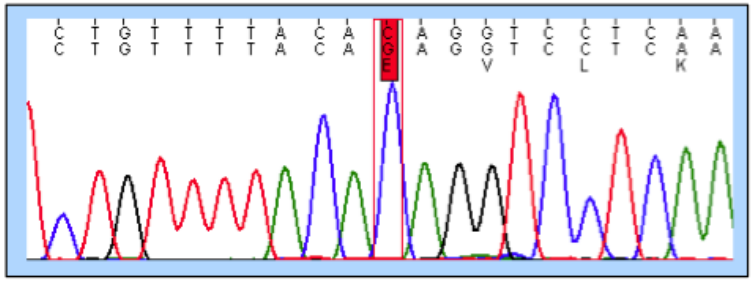
Target region is ~3Mb or ~0.1% of genome, giving enrichment of 130-fold

Identification of a splice site mutation

Parent (heterozygous carrier)
Boundary change at 786115 seen in 3 reads of 37
Boundary change at 786307 seen in 2 reads of 54
Boundary change at 791679 seen in 7 reads of 15
Boundary change at 838687 seen in 1 reads of 63
Boundary change at 838764 seen in 2 reads of 62

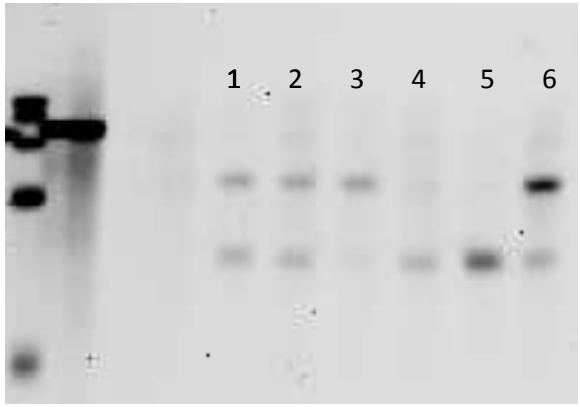


Child (homozygous affected)
Boundary change at 786115 seen in 2 reads of 68
Boundary change at 786307 seen in 1 reads of 96
Boundary change at 791679 seen in 36 reads of 36
Boundary change at 838687 seen in 2 reads of 116
Boundary change at 838764 seen in 5 reads of 170



Mutation segregates with the condition
Affects the splicing of the mRNA
Is not present in 200 ethnically matched controls
Is the only pathogenic change identified in the target region
Highly conserved

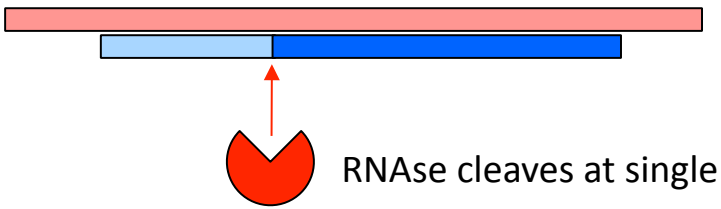
RNAse Protection Assay



Exons 12+13 protected (~245b)

Exon 13 protected (~160b)

1. Heterozygote + ΔAG probe
2. Heterozygote + normal probe
3. Homozygote + ΔAG probe
4. Homozygote + normal probe
5. Normal + ΔAG probe
6. Normal + normal probe

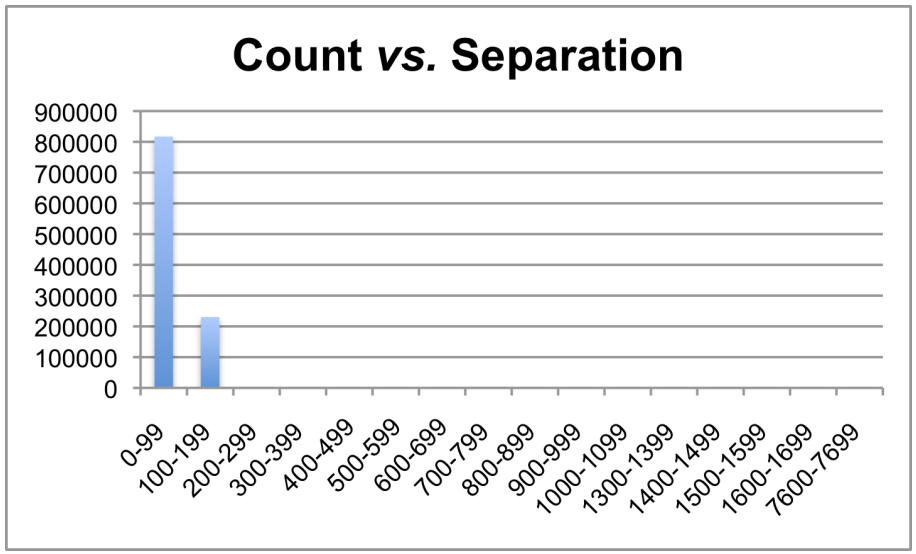


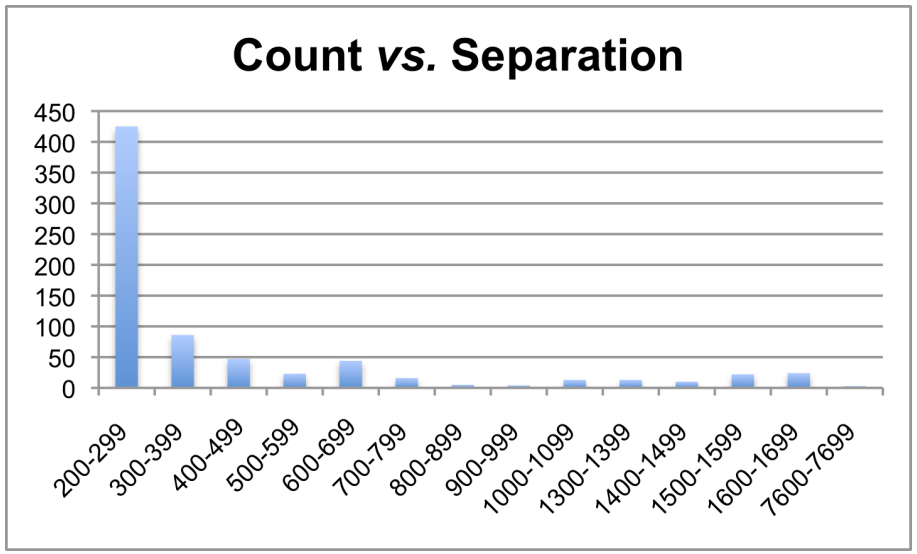
mRNA

Biotin-labelled antisense probe

RNAse cleaves at single stranded regions

Genomic rearrangements by paired read analysis

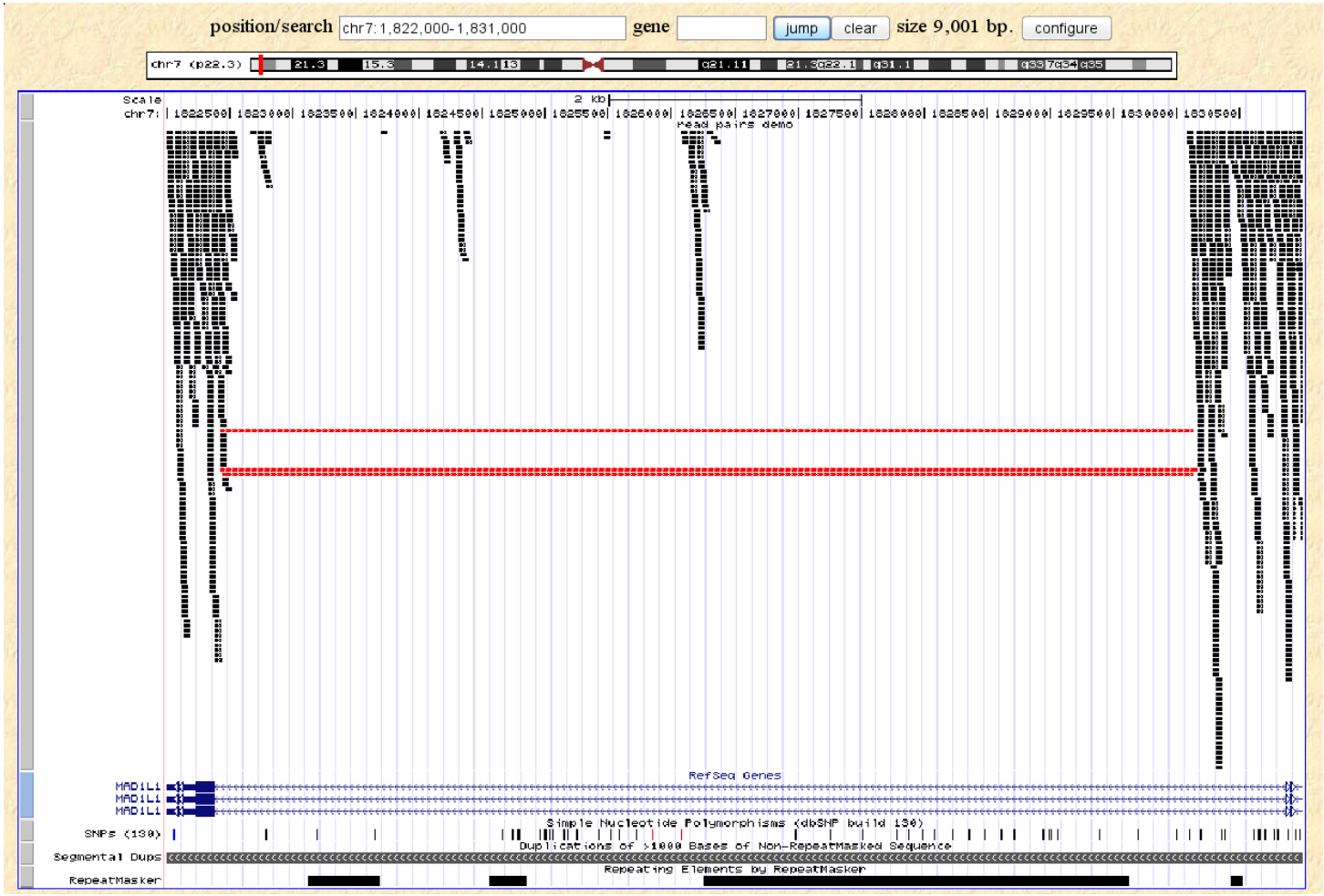




Size of fragment sequenced was ~200b

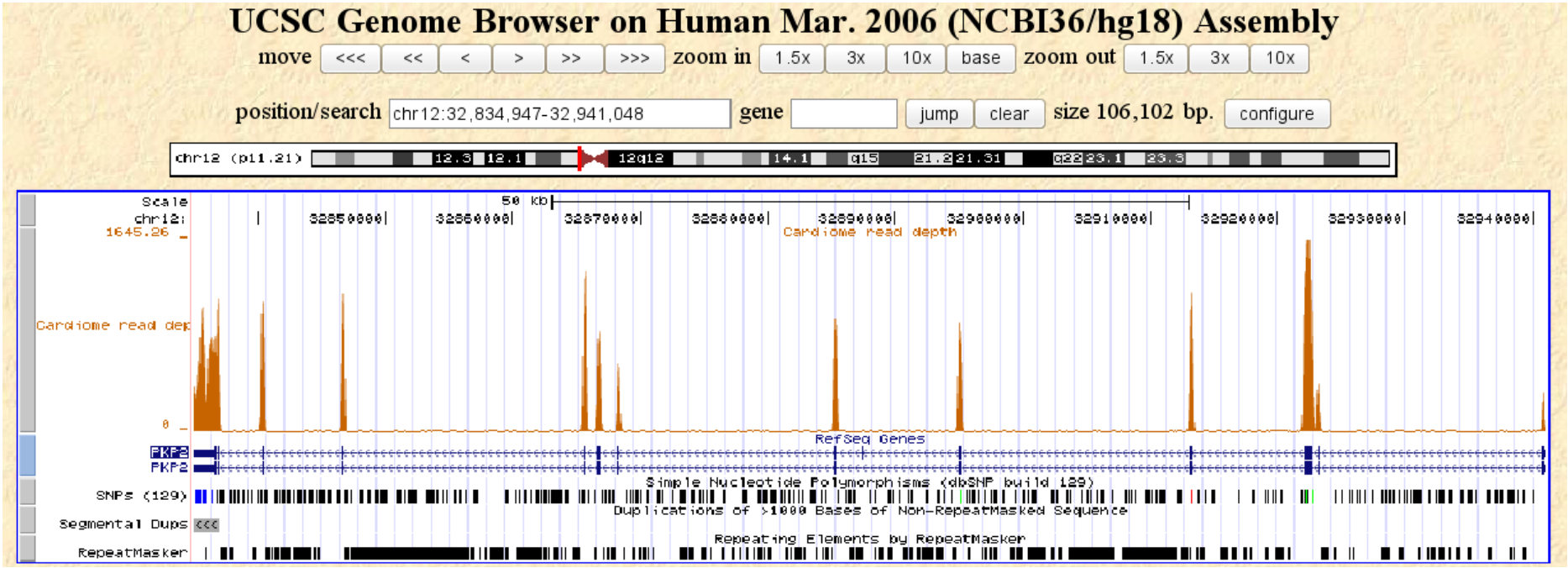
Separations > 200b may indicate large changes

A novel homozygous deletion?



Read pair data in a consanguineous case

Coverage in a well-behaved region



Different alignment programs do not work in the same way.

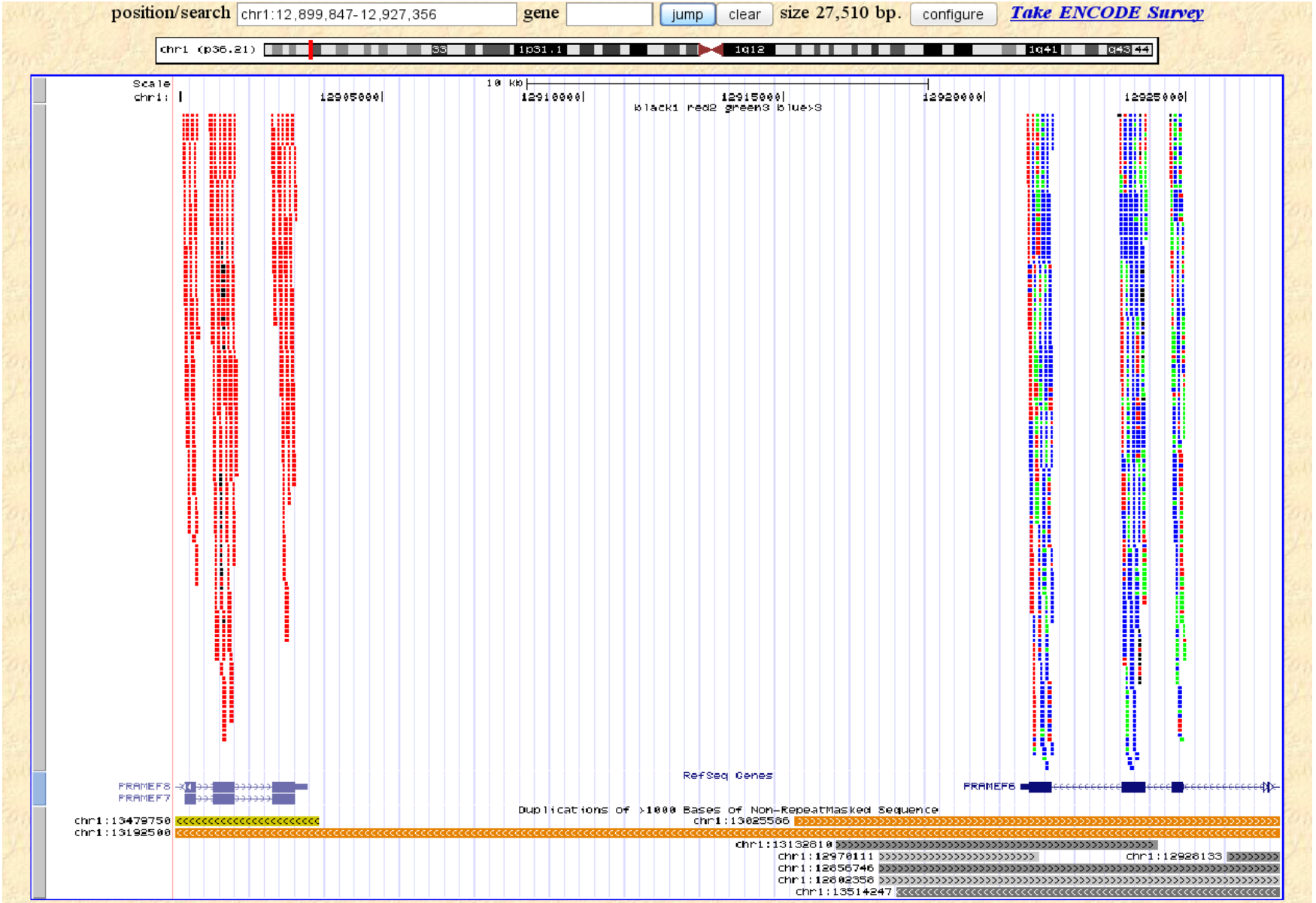
Eland “sorted” files do not include any reads that map to more than one position in the genome. Reads that map to more than one position will be problematic

Used Bowtie aligning to hg18 to map those reads which ELAND would not place on the genome and then analysed those mapping into the LCA9 region.

There were 58 exons within the duplicated region and apparent coverage was very good:
Cumulative exon length was 27714 bases
Of these 358 had zero coverage and 921 had coverage below a depth of 10

But reads were mapped to multiple positions within the duplication and in the majority of these mappings identifying the true position is not possible.
The table below shows the number of reads within the duplication and the number of times each read mapped into the region. Unique is defined as a read having one genomic location with fewer mismatches than other mapped locations.

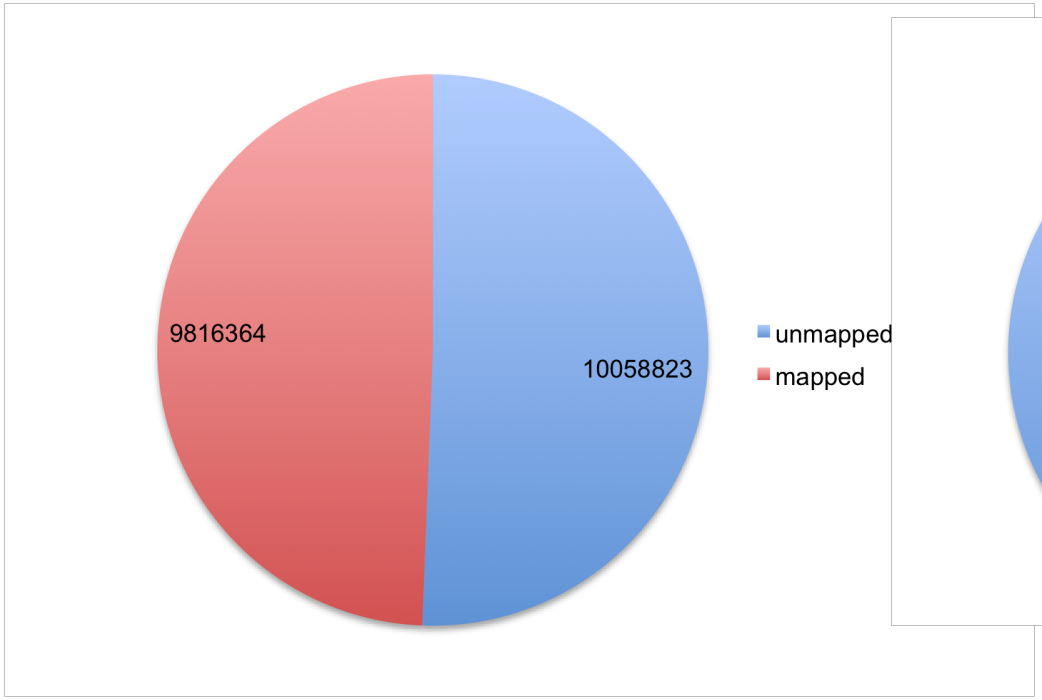
Hits within LCA9	No. of reads	Unique	Not unique
1	5240		
2	15541	1057	14484
3	5353	645	4708
4	13578	1024	12554
5	263	19	244
6	105	0	105



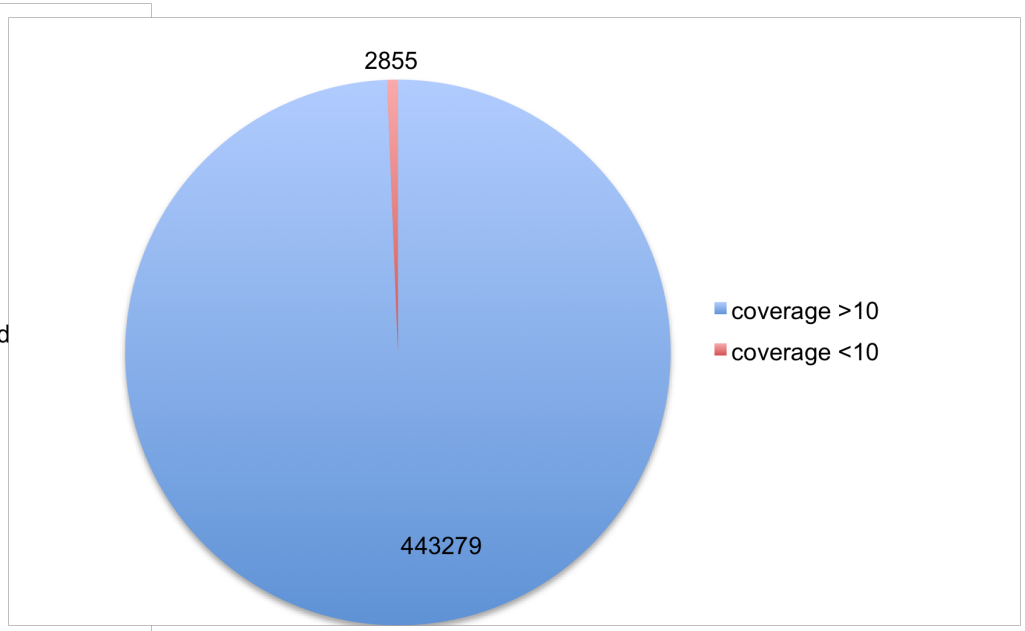
Mapping test whole exome reads onto a small section of the duplicated region in LCA9 using Bowtie onto hg18.
Colour scheme is: Black- maps uniquely within region; Red- maps twice within LCA9 duplication(s);
Green- maps three times; Blue- four (or more) times

Sub-Genomic Targetting

- Focus on regions of interest
- Higher coverage
- Easier informatics
- Fewer “unwanted” genotypes
- Target on the basis of phenotype or pathways
- Compare with exome targetting
- *Work in progress…*



Total reads



Mapped reads

Coverage of a 54-gene pull down experiment.
Data from Eland2 alignment with stats retrieved using
a Python script written by Dr Bruce Hayward
Exons analysed: 1325
Total reads: 19875187
Reads not mapping to an exon: 10058823
Cumulative exon length: 446134 bases
981 bases had zero coverage (as unique reads)
1874 had coverage <10

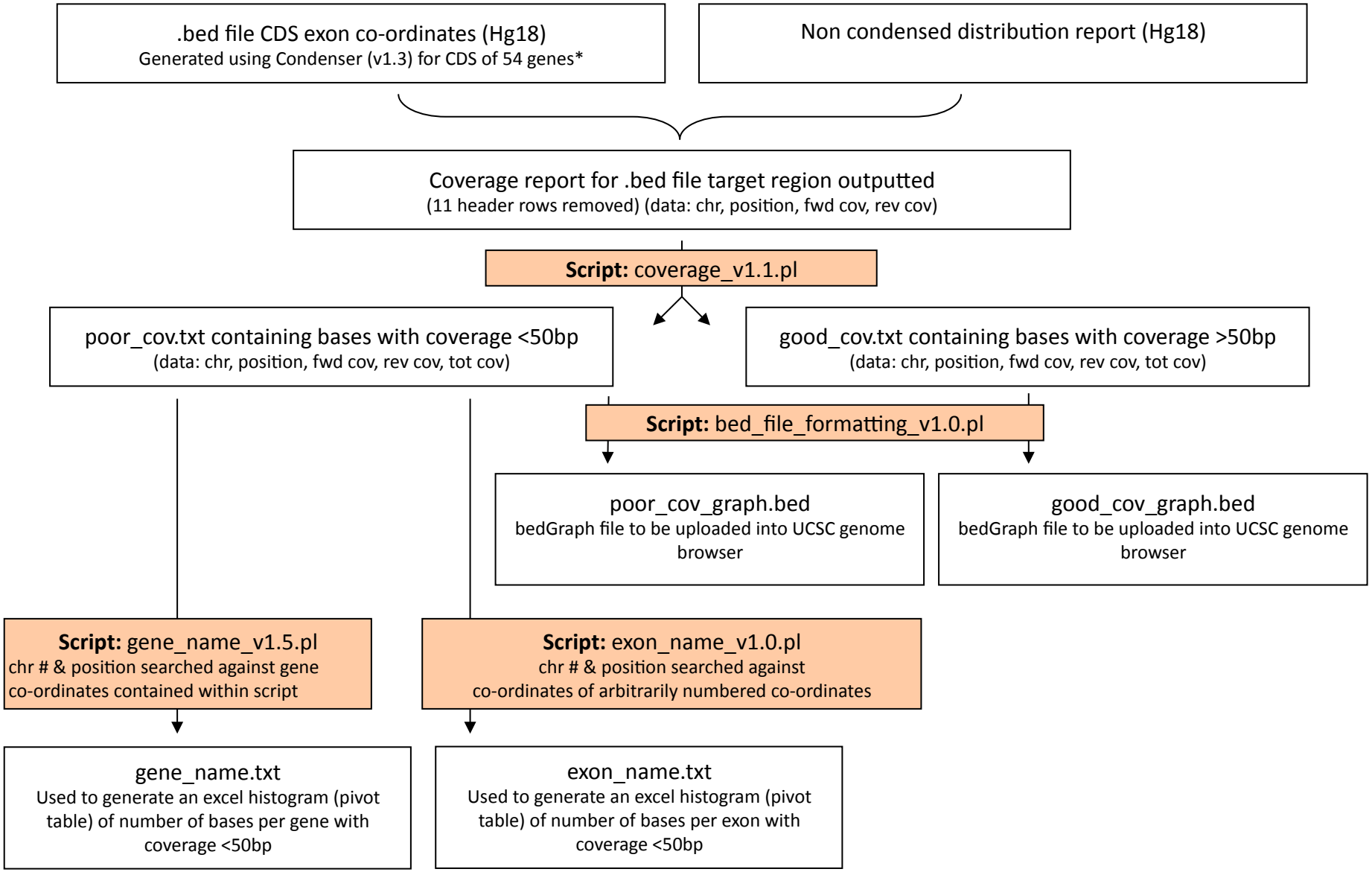
Example script to write a simple file for browsing

```
 #!/usr/bin/perl -w
use strict ;
# GRTaylor April 2009
# take coverage file: this is a subset of the whole genome based on a bed file
# if the bedfile used to define NextGene output file is just exons it is quite a small file, good for comparing exon coverage
# if the bedfile is a gene file it will shaow all coverage within the gene co-ords, so you can see the reads mapping outside the baits (but only witihin the gene list)
# you could look at the whole genome or chromosome, but probably not a good idea to do this at single base resolution
# ignore lines not starting with a number or X or Y (could skip first 11 lines)
# use project name and time as output filename: not implemented
# open forward and reverse coverage files
# split input coverage on tabs and extract values
# foreach line add values in 3rd and 4th columns (forward and revese coverage) = ($sum)
# choose depth of coverage to be reported on  n (selectable ($depth), e.g. 50) could change to > in line 64 to get only high coverage
# write out as bedgraph files using 1 line header and two positions ($position & $pos2) all positions with coverage below the sum
# browser set to Titin just as somewhere to look at example output
```

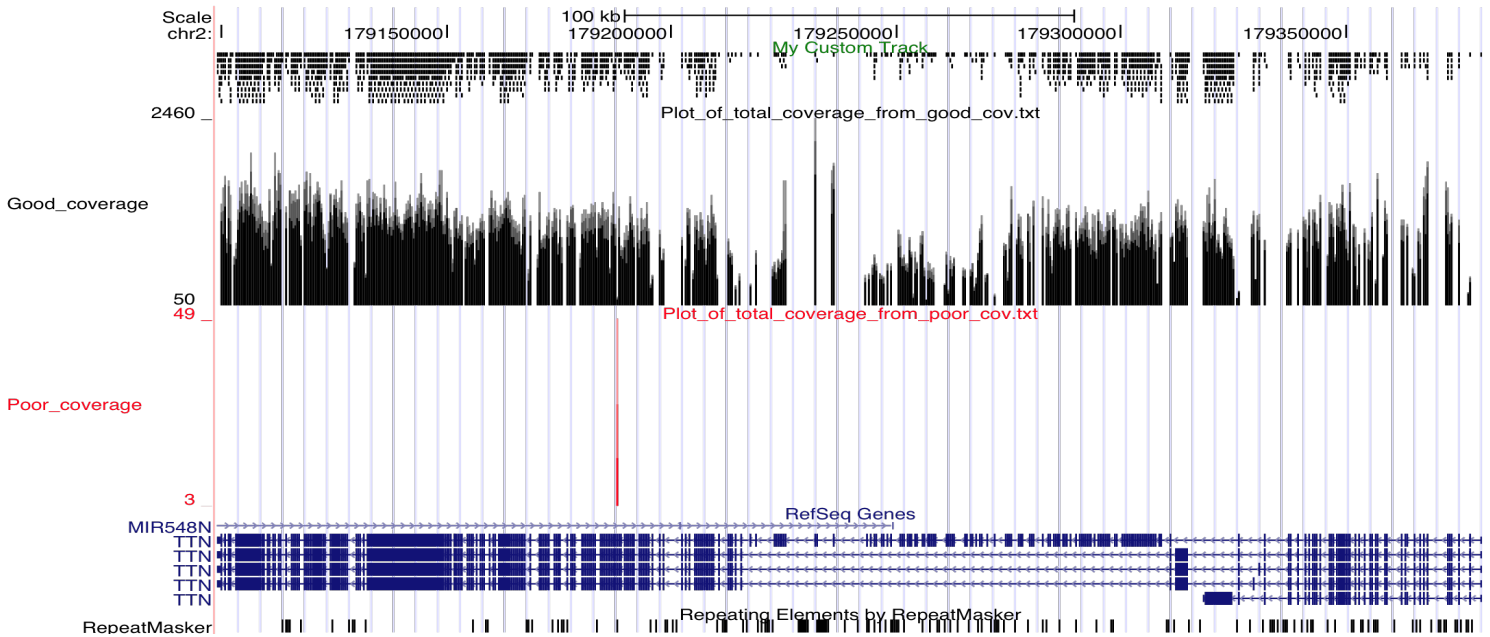
```
print "enter minimum depth of coverage\n";
$depth = <STDIN> ;
chomp $depth ;
open (FOWD, ">", "forward.bed") ;
open (REV, ">", "reverse.bed") ;

# open input file die if no file with error message
print "enter the file name\n";
$filename = <STDIN>;
chomp $filename ;
open (COVERAGE, "<", $filename)
    or die "couldn't open $filename\n" ;

print FOWD "browser position chr2:179390719-179672150 \n";
print FOWD "track type=bedGraph name=forward_reads visibility=full color=200,100,0\n";


print REV "browser position chr2:179390719-179672150 \n";
print REV "track type=bedGraph name=reverse_reads visibility=full color=0,100,200\n";



while ($line = <COVERAGE>) {
    # find $f and $r use split, substring or capture charactors after 2nd and 3rd tabs
    if ($line =~ m/^[\dXY]/) { #line starts with a number replace \d with 12 for chr 12 or m/^[\dXY]/ for X and Y as well
        @line = split /[\t\n]/,$line ; #split on tab or newline
        $chr = $line[0];
        $position = $line[1] ;
        $f = $line[2];
        $r = $line[3] ;
        $sum = $f + $r ; #if $sum > $depth statement
        if ($sum > $depth) {
            $count += 1 ;
            $pos2=$position+1 ;
        print FOWD "chr$chr\t$position\t$pos2\t$f\n" ;
        print REV "chr$chr\t$position\t$pos2\t$r\n" ;

        }
    }
}
close FOWD ;
close REV ;
exit ;
```

NB defined variables (“my etc.” not shown for brevity)

Base coverage pathway (Chris Watson)



Titin *(TTN)*



Example of coverage in a pull down experiment.

Titin is big, chr2: 179390719-179672150 (build 18), 312 exons.

Coverage plotted as bedgraph file and viewd on UCSC brower. Scripting to do this written by clinical scientist trainee Chris Watson.

Summary

- Hands-on NGS requires access to substantial computational support, Unix environment and scripting tools
- Samples could be exported and data returned as variants with some quality markers, avoiding the need for major IT infrastructure (hub and spoke concept: hubs could be academic or commercial providers)
- Cost effective use requires large scale of operation

Acknowledgements

- Sequencing operations
  - Joanne Morgan, Heather Fraser, Clare Logan
- Leeds Autozygosity group
  - Colin Johnson & David Bonthron
- Leeds NHS Genetics
  - Nick Camm, Helen Lindsay, Antigone Tzika, Carol Chu, Paul Roberts & Ruth Charlton
- YCR Pre-cancer genomics group
  - Pamela Rabbitts, Henry Wood & Stefano Berri

Informatics

- Henry Wood, Joanne Morgan, Aengus Stewart, Ian Carr, Stefano Berri, Bruce Hayward, Nick Camm, Chris Watson

Publications

- Morgan JE *et al* Human Mutation 2010 4:484-91 Genetic diagnosis of familial breast cancer using clonal sequencing
- Wood HM *et al* Nucleic Acids Res. 2010 Jun 4. [Epub ahead of print] Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens.