Overview of methods for variant calling from next-generation sequence data

Thomas Keane,
Vertebrate Resequencing Informatics,
Wellcome Trust Sanger Institute
Email: tk2@sanger.ac.uk

Variant calling from next-generation sequence data

- **Data Formats + Workflows**
- **SNP Calling**
- **Short Indels**
- **Structural Variation**
- **Experimental Design**

Data Production Workflow

Next-gen sequencing experiments

- Several, tens or hundreds of samples
- One or more sequencing libraries per sample
- Sample could constitute several libraries

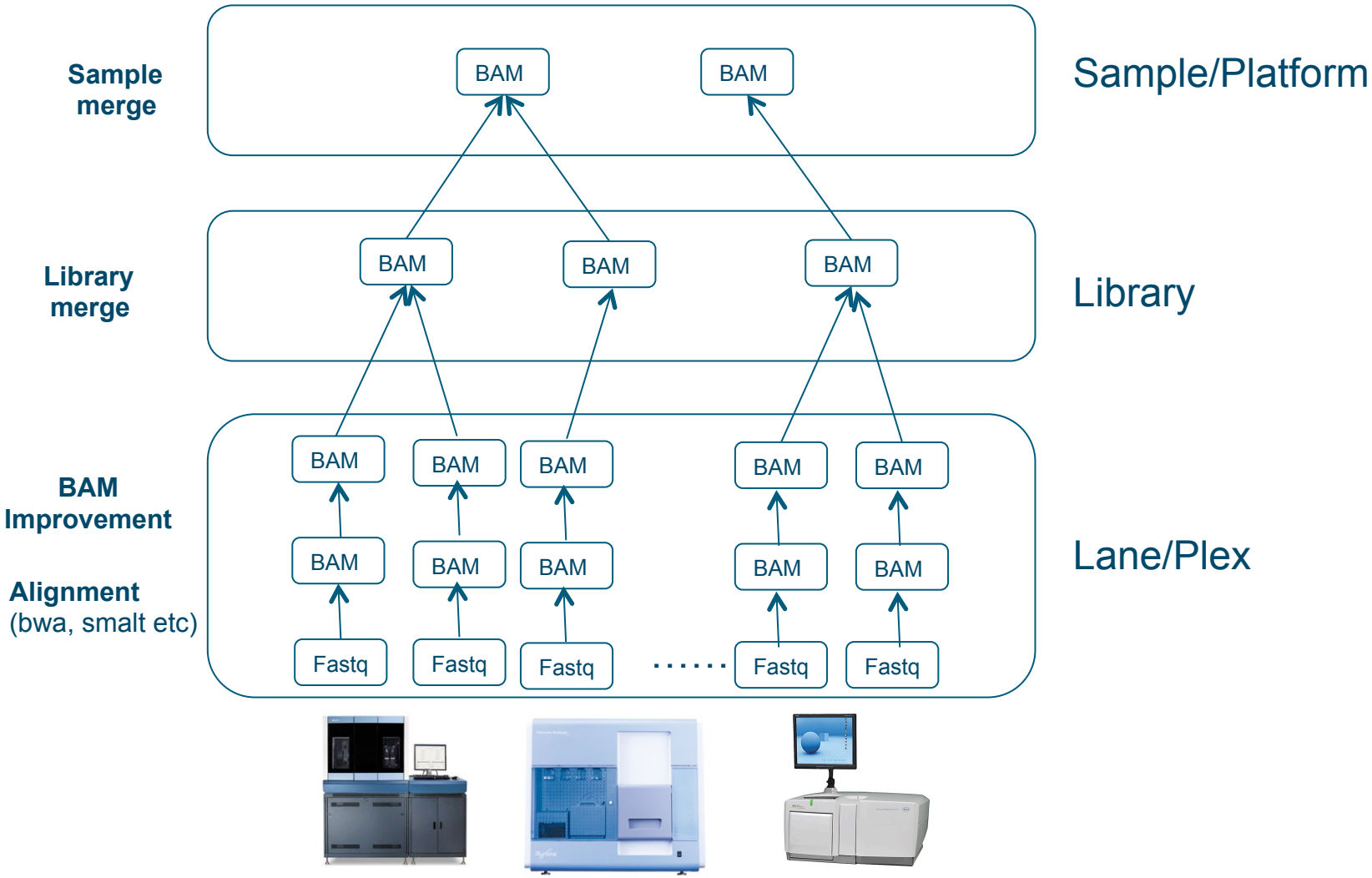
Reference based analysis or de novo assembly

How the data is processed can have consequences on quality of variant calling

Alignment of the reads onto the reference is just the first step

- QC of data is very important for good calls
  - Biases in the library or sequence data will produce unexpected results or miss variant calls
  - E.g. GC bias
- How the data is processed prior to variant calling is important
  - Mapping -> improvement -> merging -> variant calling

Data Production Workflow



SAM/BAM Format

Proliferation of alignment formats over the years: Cigar, psl, gff, xml etc.

SAM (Sequence Alignment/Map) format

- Single unified format for storing read alignments to a reference genome

BAM (Binary Alignment/Map) format

- Binary equivalent of SAM
- Developed for fast processing/indexing

Advantages

- Can store alignments from most aligners
- Supports multiple sequencing technologies
- Supports indexing for quick retrieval/viewing
- Compact size (e.g. 112Gbp Illumina = 116Gbytes disk space)
- Reads can be grouped into logical groups e.g. lanes, libraries, individuals/ genotypes
- Supports second best base call/quality for hard to call bases

Possibility of storing raw sequencing data in BAM as replacement to SRF & fastq

Read Entries in SAM

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Heng Li , Bob Handsaker , Alec Wysoker , Tim Fennell , Jue Ruan , Nils Homer , Gabor Marth , Goncalo Abecasis , Richard Durbin , and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics,* 25:2078-2079

Extended Cigar Format

Cigar has been traditionally used as a compact way to represent a sequence alignment

Operations include

- M - match or mismatch
- I - insertion
- D - deletion

SAM extends these to include

- S - soft clip
- H - hard clip
- N - skipped bases
- P – padding

E.g. Read: `ACGCA-TGCAGTtagacgt`

Ref: `ACTCAGTG--GT`

Cigar: 5M1D2M2I2M7S

What is the cigar line?

E.g. Read: `tgtcgtcACGCATG---CAGTtagacgt`

Ref: `ACGCATGCGGCAGT`

Cigar:

Read Group Tag

Each lane has a unique RG tag

RG tags

- ID: SRR/ERR number
- PL: Sequencing platform
- PU: Run name
- LB: Library name
- PI: Insert fragment size
- SM: Individual
- CN: Sequencing center

1000 Genomes BAM File

```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fdfd811849cc2fadebc929bb925902e5
@SQ     SN:4    LN:191154276    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:23dccd106897542ad87d2765d28a19a1
@SQ     SN:5    LN:180915260    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0740173db9ffd264d728f32784845cd7
@SQ     SN:6    LN:171115067    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1d3a93a248d92a729ee764823acbbc6b
@SQ     SN:7    LN:159138663    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:618366e953d6aaad97dbe4777c29375e
@SQ     SN:8    LN:146364022    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:96f514a9929e410c6651697bded59aec
@SQ     SN:9    LN:141213431    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:3e273117f15e0a400f01055d9f393768
@SQ     SN:10   LN:135534747    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:988c28e000e84c26d552359af1ea2e1d
@SQ     SN:11   LN:135006516    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98c59049a2df285c76ffb1c6db8f8b96
@SQ     SN:12   LN:133851895    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:51851ac0e1a115847ad36449b0015864
@SQ     SN:13   LN:115169878    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:283f8d7892baa81b510a015719ca7b0b
@SQ     SN:14   LN:107349540    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98f3cae32b2a2e9524bc19813927542e
@SQ     SN:15   LN:102531392    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:e5645a794a8238215b2cd77acb95a078
@SQ     SN:16   LN:90354753     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fc9b1a7b42b97a864f56b348b06095e6
@SQ     SN:17   LN:81195210     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:351f64d4f4f9ddd45b35336ad97aa6de
@SQ     SN:18   LN:78077248     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:b15d4b2d29dde9d3e4f93d1d0f2cbc9c
@SQ     SN:19   LN:59128983     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1aacd71f30db8e561810913e0b72636d
@SQ     SN:20   LN:63025520     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0dec9660ec1efaaf33281c0d5ea2560f
@SQ     SN:21   LN:48129895     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:2979a6085bfe28e3ad6f552f361ed74d
@SQ     SN:22   LN:51304566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a718acaa6135fdca8357d5bfe94211dd
@SQ     SN:X    LN:155270560    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:7e0e2e580297b7764e31dbc80c2540dd
@SQ     SN:Y    LN:59373566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1fa3474750af0948bdf97d5a0ee52e51
@SQ     SN:MT   LN:16569        AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:c68f52674c9fb33aef52dcf399755519
@SQ     SN:GL000207.1   LN:4262 AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:f3814841f1939d3ca19072d9e89f3fd7
@RG     ID:ERR000047    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000048    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000071    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000091    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000094    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000105    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000377    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001126    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001127    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001128    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001180    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001181    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR005185    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR008994    PL:ILLUMINA     LB:HUMsgR22AFDFAAPE     PI:333  DS:SRP000546    SM:NA18582      CN:BGI
@RG     ID:ERR009030    PL:ILLUMINA     LB:HUMsgR22AFDFAAPE     PI:333  DS:SRP000546    SM:NA18582      CN:BGI
@PG     ID:bwa  VN:0.5.5
ERR001127.3207020       163     5       9998    0       45M     =       10089   136     AACTAACCCTAACCCTAACCCTAACCATAACCCTAACCCTAACAC    /3/@<></>+<?=??A=?3@@A>??@9A>?11A9=@%%@A=?:$8
XT:A:R  XN:i:3  SM:i:0  AM:i:0  X0:i:2  X1:i:4    XM:i:2  XO:i:0  XG:i:0  RG:Z:ERR001127  NM:i:5  MD:Z:0N0N0N23C16C1      OQ:Z:=>7IIII1I0IIIIIIII0IIIIEII1IIH))?0=I%%II=?-$4
```

```
samtools view –h mybam.bam
```

SAM/BAM Tools

Well defined specification for SAM/BAM

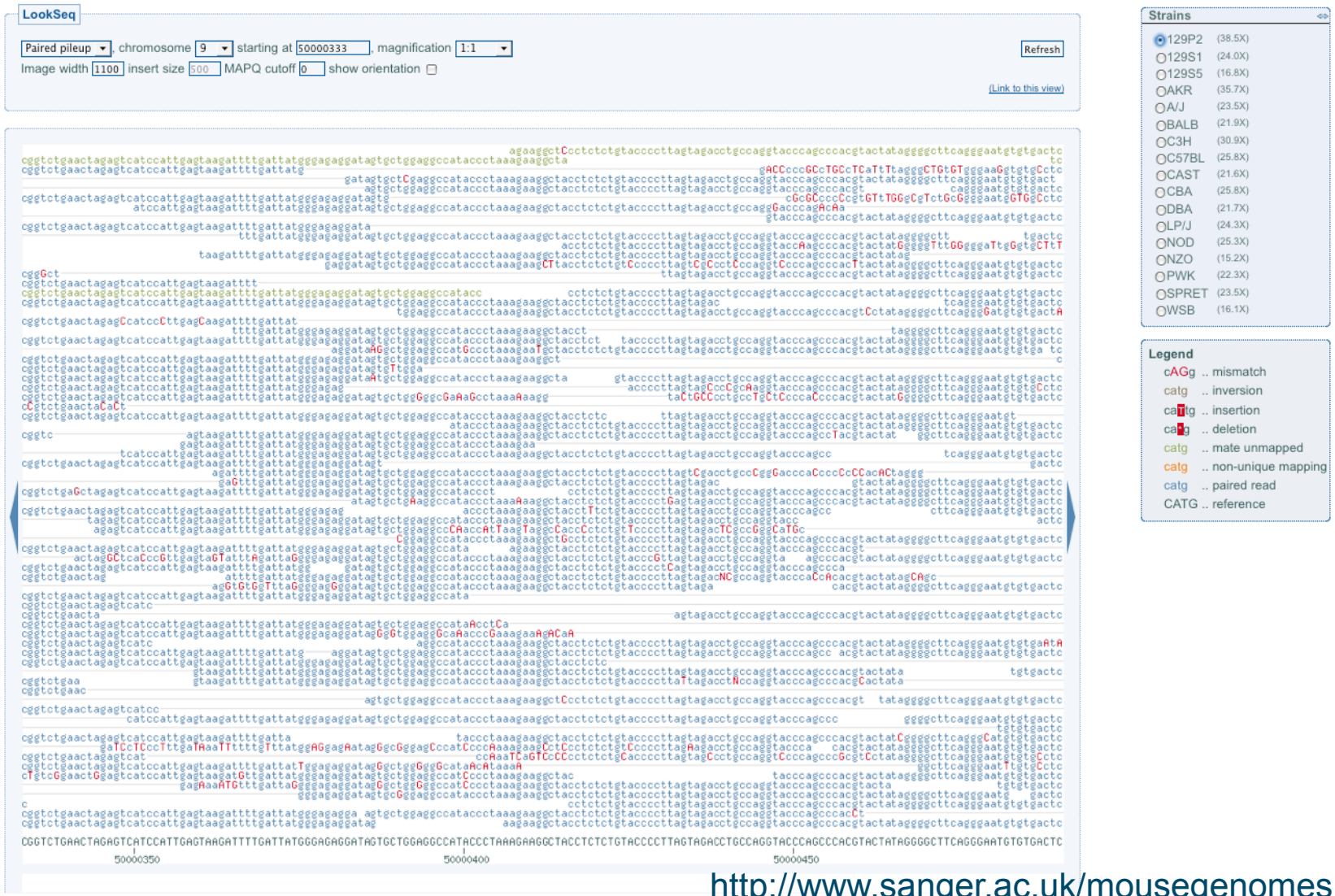
Several tools and programming APIs for interacting with SAM/BAM files

- Samtools - Sanger/C (http://samtools.sourceforge.net)
  - Convert SAM <-> BAM
  - Sort, index, BAM files
  - Flagstat – summary of the mapping flags
  - Merge multiple BAM files
  - Rmdup – remove PCR duplicates from the library preparation
- Picard - Broad Institute/Java (http://picard.sourceforge.net)
  - MarkDuplicates, CollectAlignmentSummaryMetrics, CreateSequenceDictionary, SamToFastq, MeanQualityByCycle, FixMateInformation…….
- Bio-SamTool – Perl (http://search.cpan.org/~lds/Bio-SamTools/)
- Pysam – Python (http://code.google.com/p/pysam/)

BAM Visualisation

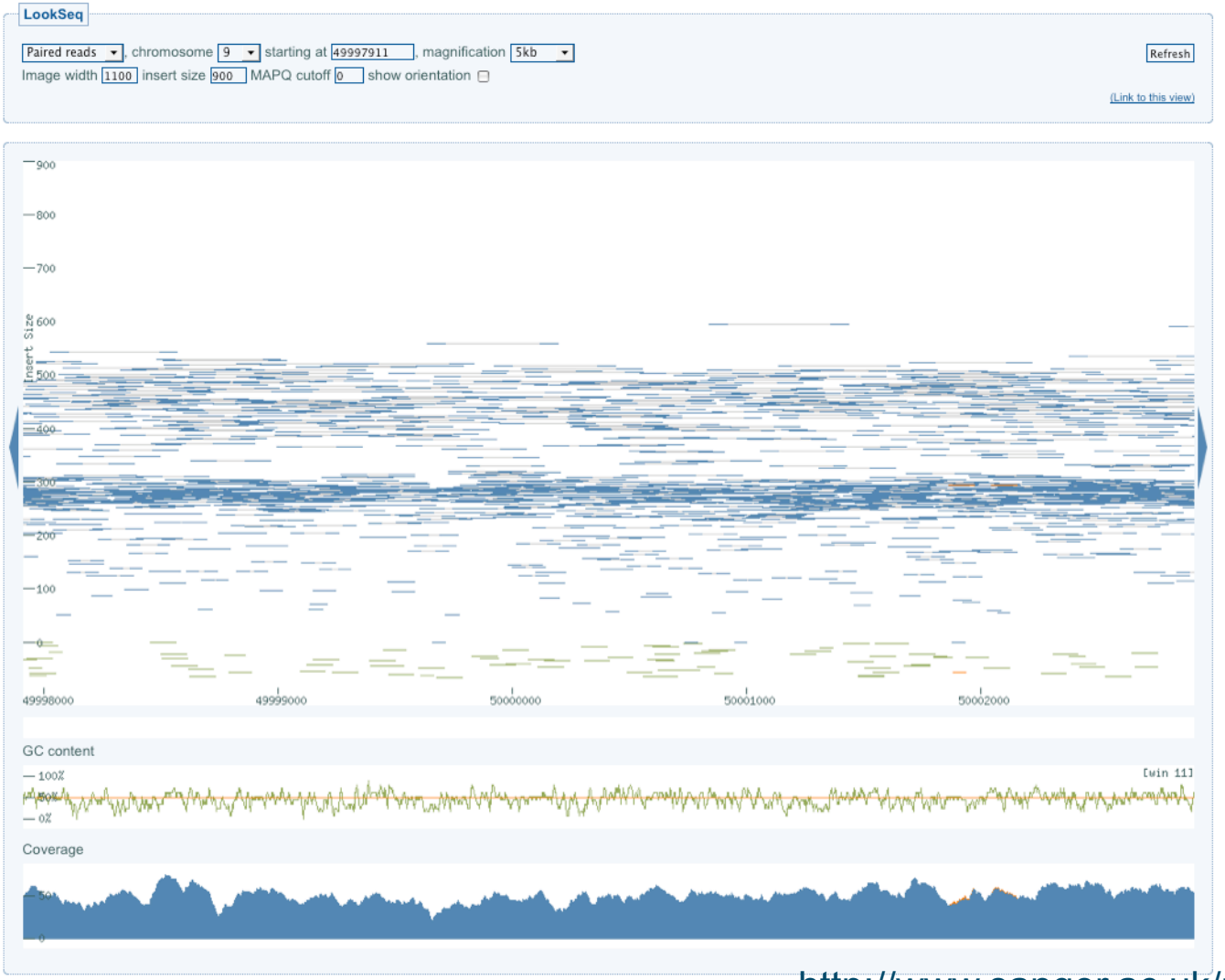
- BamView, LookSeq, Gap5: http://www.sanger.ac.uk/Software
- IGV: http://www.broadinstitute.org/igv/v1.3
- Tablet: http://bioinf.scri.ac.uk/tablet/

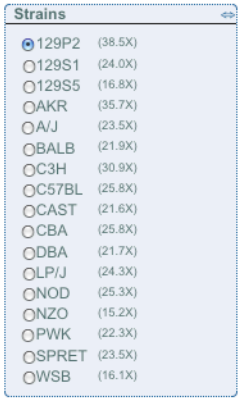
BAM Visualisation

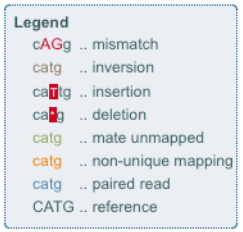


http://www.sanger.ac.uk/mousegenomes

BAM Visualisation

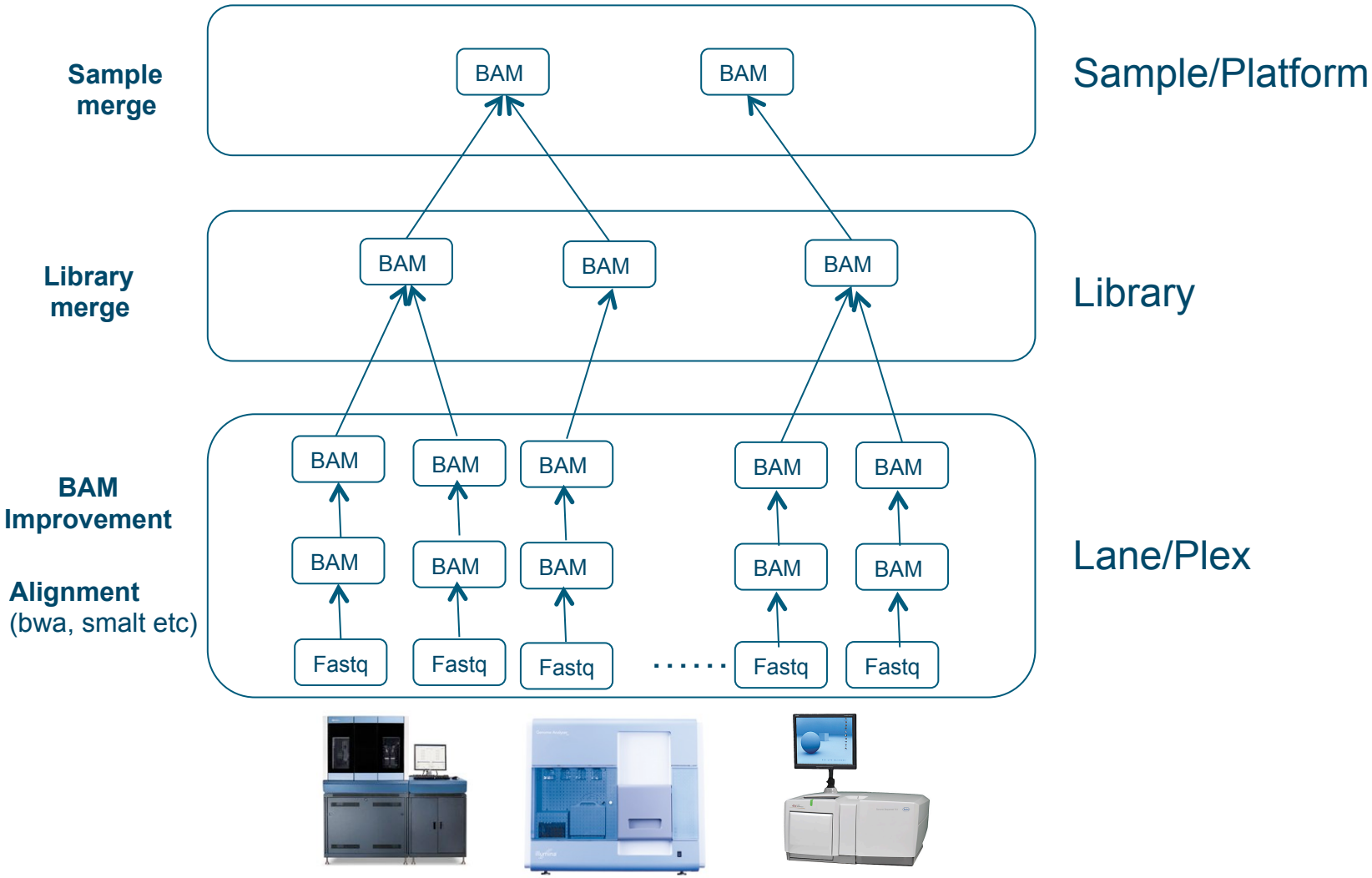






http://www.sanger.ac.uk/mousegenomes

Data Production Workflow



BAM Improvement

Lane level operation carried out after alignment

Input: BAM

Process 1: Realignment

Process 2: Base Quality Recalibration

Output: (improved) BAM

Realignment

Short indels in the sample relative to the reference can pose difficulties for alignment programs

Indels occurring near the ends of the reads are often not aligned correctly

- Excess of SNPs rather than introduce indel into alignment

Realignment algorithm

- Input set of known indel sites and a BAM file
- At each site, model the indel haplotype and the reference haplotype
- Given the information on a known indel
  - Which scenario are the reads more likely to be derived from?
- New BAM file produced with read cigar lines modified where indels have been introduced by the realignment process

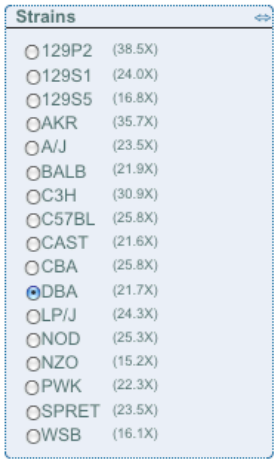
Software

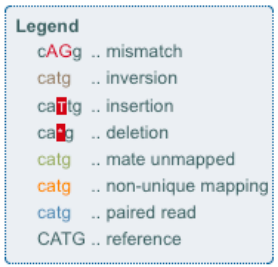
- Implemented in GATK from Broad (IndelRealigner function)

What sites?

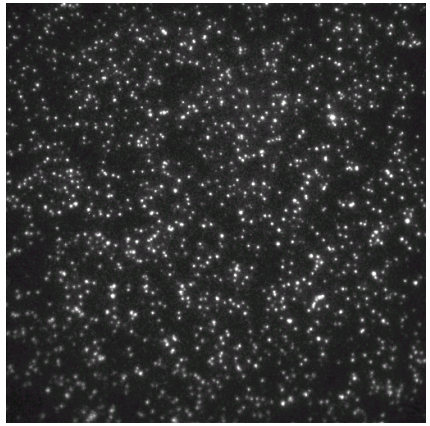
- Previously published indel sites, dbSNP, 1000 genomes, generate a rough/ high confidence indel set

Realignment





Base Quality Recalibration



Each base call has an associated base call quality

- What is the chance that the base call is incorrect?
  - Illumina evidence: intensity values + cycle
- Phred values (log scale)
  - Q10 = 1 in 10 chance of base call incorrect
  - Q20 = 1 in 100 chance of base call incorrect
- Accurate base qualities essential measure in variant calling

Rule of thumb: Anything less than Q20 is not useful data

Typically phred values max. out at Q35-40

Illumina sequencing

- Control lane or spiked control used to generate a quality calibration table
- If no control – then use pre-computed calibration tables

Quality recalibration

- 1000 genomes project sequencing carried out on multiple platforms at multiple different sequencing centres
- Are the quality values comparable across centres/platforms given they have all been calibrated using different methods?

Base Quality Recalibration

Original recalibration algorithm

- Align subsample of reads from a lane to human reference
- Exclude all known dbSNP sites
  - Assume all other mismatches are sequencing errors
- Compute a new calibration table bases on mismatch rates per position on the read

Pre-calibration sequence reports Q25 base calls

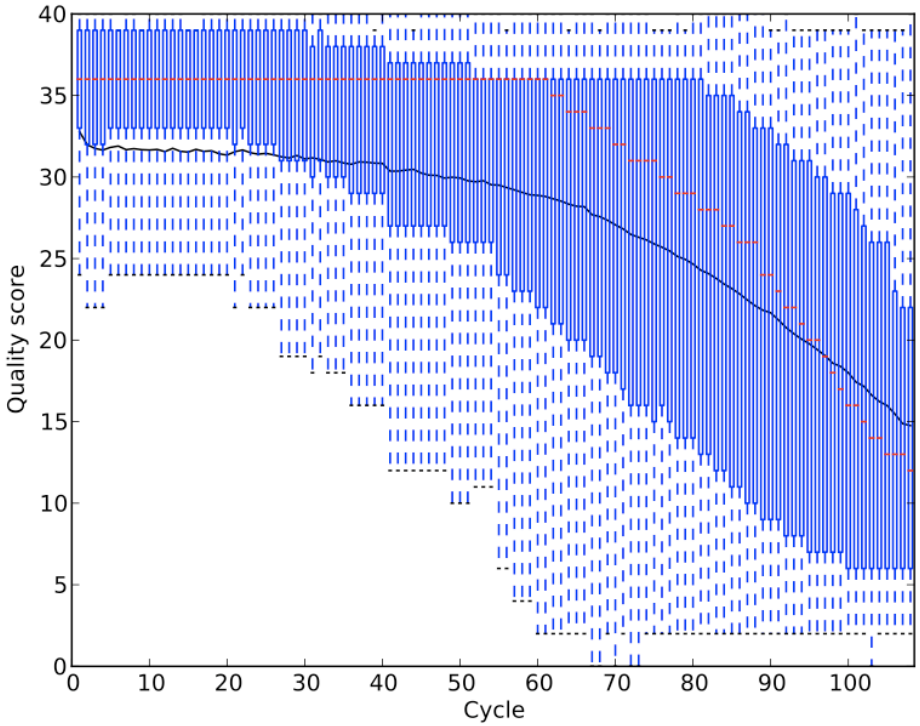
- After alignment - it may be that these bases actually mismatch the reference at a 1 in 100 rate, so are actually Q20

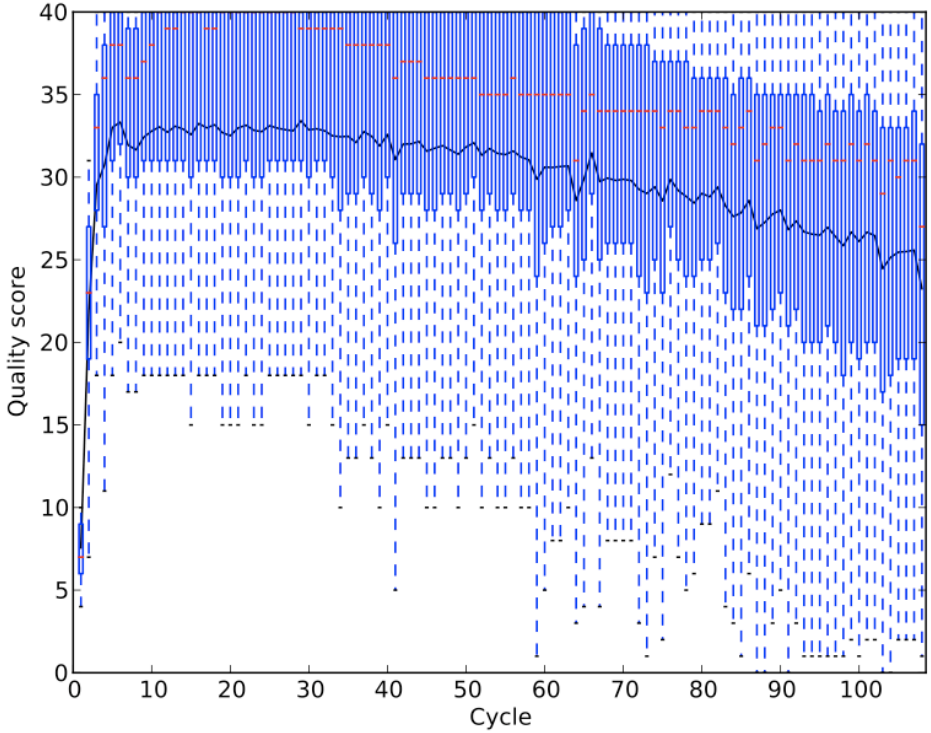
Recent improvements – GATK package

- Reported/original quality score
- The position within the read
- The preceding and current nucleotide (sequencing chemistry effect) observed by the sequencing machine
- Probability of mismatching the reference genome

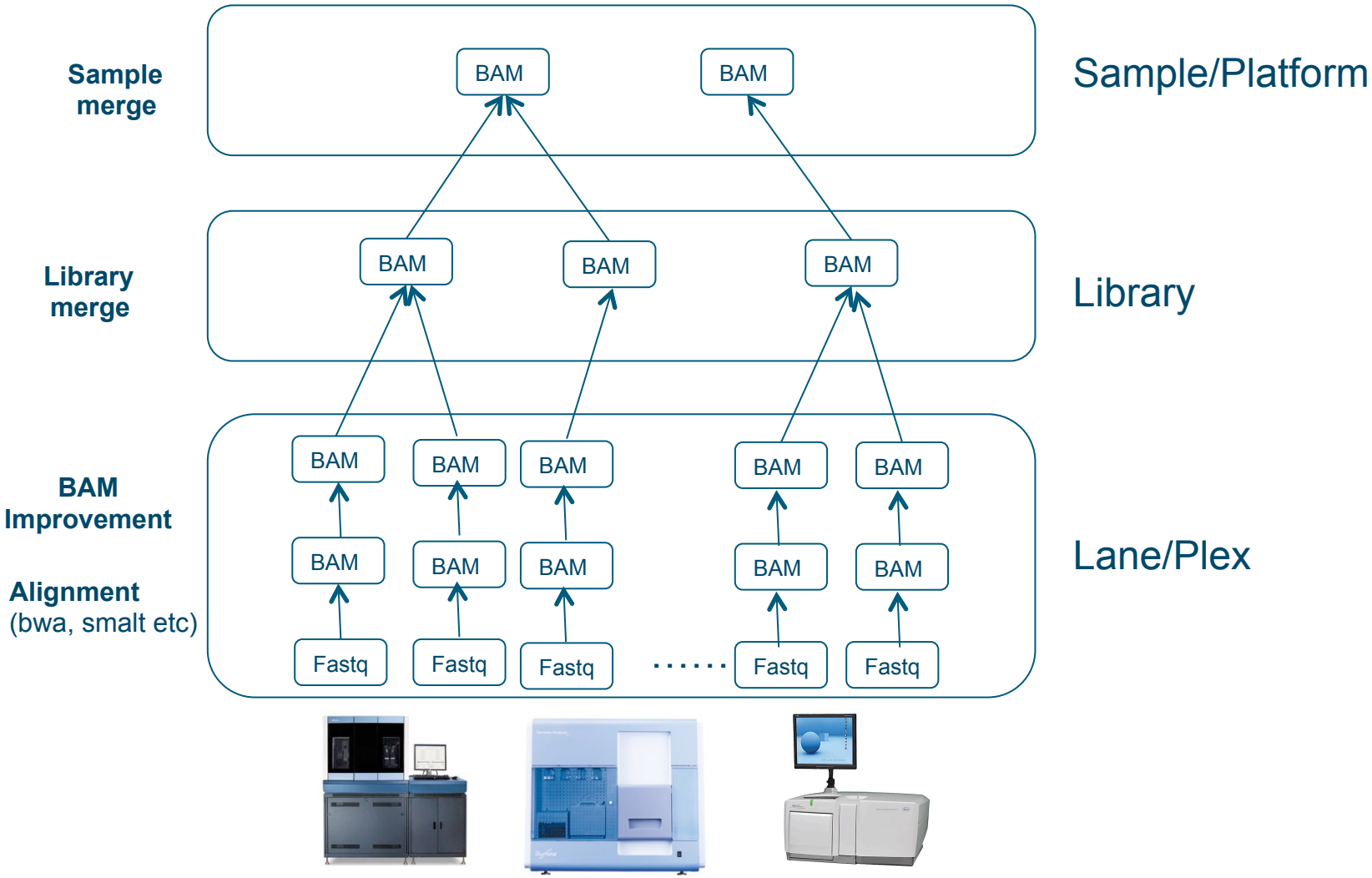
NOTE: requires a reference genome and a catalog of variable sites

Base Quality Recalibration Effects





Data Production Workflow



Library Merge

Library level operation carried out after BAM improvement

Input: Multiple Lane BAMs

Process 1: Merge BAMs (picard - MergeSamFiles)

Process 2: Duplicate fragment identification

Output: BAM

Library Duplicates

All second-gen sequencing platforms are NOT single molecule sequencing

- PCR amplification step in library preparation
- Can result in duplicate DNA fragments in the final library prep.
- PCR-free protocols do exist – require large volumes of input DNA

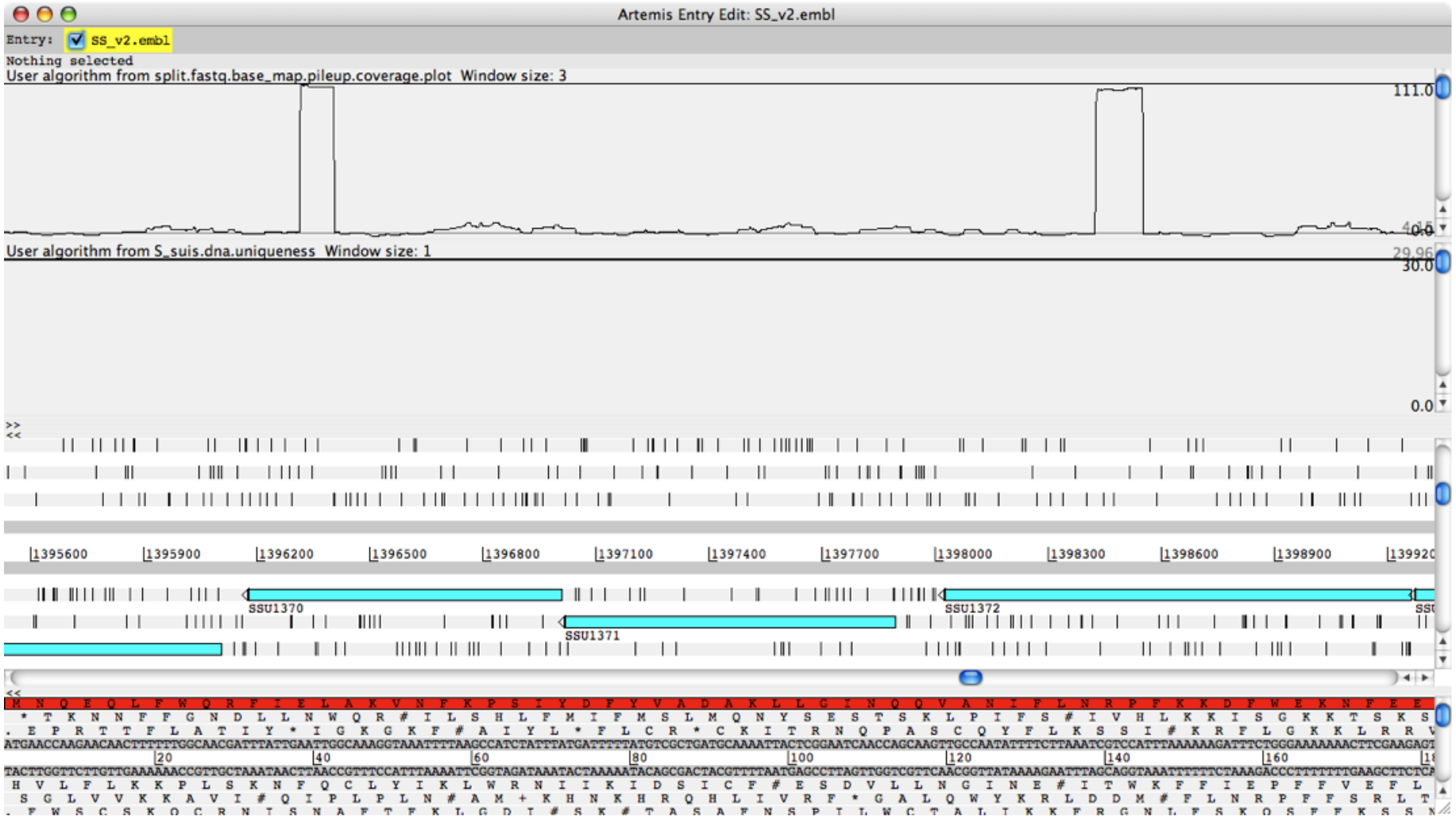
Generally low number of duplicates in good libraries (<3%)

- Align reads to the reference genome
- Identify read-pairs where the outer ends map to the same position on the genome and remove all but 1 copy
  - Samtools: samtools rmdup or samtools rmdupse
  - Picard/GATK: MarkDuplicates

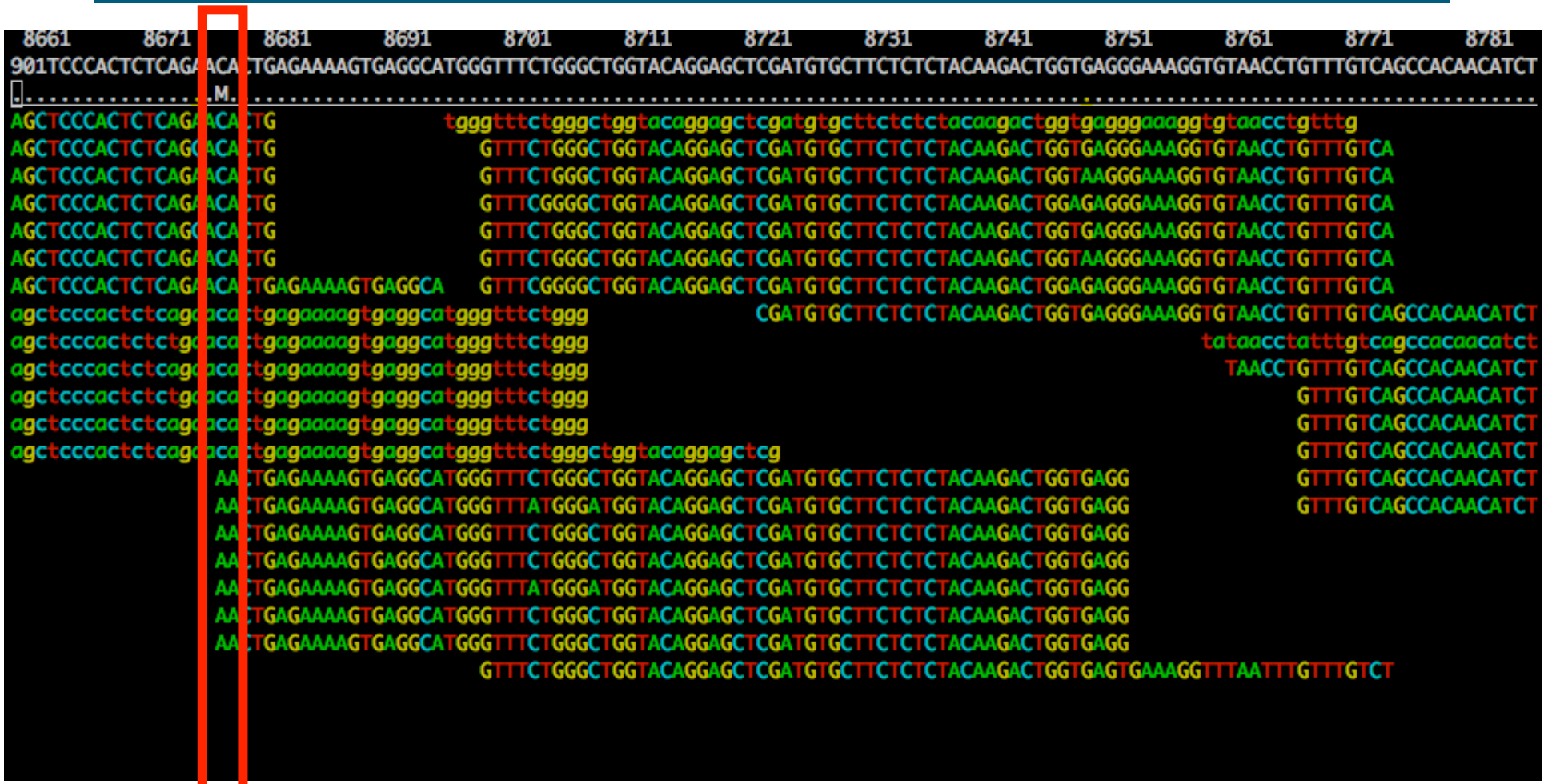
Can result in false SNP calls

- Duplicates manifest themselves as high read depth support

Library Duplicates



Duplicates and False SNPs



NA12005 - chr20:8660-8790

Software Tools

Alignment

- BWA: http://bio-bwa.sourceforge.net/bwa.shtml
- Smalt: http://www.sanger.ac.uk/resources/software/smalt/
- Stampy: http://www.well.ox.ac.uk/project-stampy

BAM Improvement

- Realignment (GATK): http://www.broadinstitute.org/gsa/wiki/index.php/Local_realignment_around_indels
- Recalibration: http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration

Library Merging

- BAM Merging (Picard): http://picard.sourceforge.net/command-line-overview.shtml#MergeSamFiles
- Duplicate Marking/removal (Picard): http://picard.sourceforge.net/command-line-overview.shtml#MarkDuplicates

Variant calling from next-generation sequence data

- Data Formats + Workflows
- **SNP Calling**
- Short Indels
- Structural Variation
- Experimental Design

SNP Calling

SNP – single nucleotide polymorphisms

- Examine the bases aligned to position and look for differences
- Sequence context of the SNP e.g. homopolymer run

SNP Calling vs. genotyping

Homozygous vs heterozygous SNPs

Factors to consider when calling SNPs

- Base call qualities of each supporting base
- Proximity to
  - Small indel
  - Homopolymer run (>4-5bp for 454 and >10bp for illumina)
- Mapping qualities of the reads supporting the SNP
  - Low mapping qualities indicates repetitive sequence
- Read length
  - Possible to align reads with high confidence to larger portion of the genome with longer reads
- Paired reads
- Sequencing depth
  - Few individuals/strains at high coverage vs. low coverage many individuals/strains
    - 1000 genomes is low coverage sequencing across many individuals
    - Population based SNP calling methods

Read Length & Callable Genome

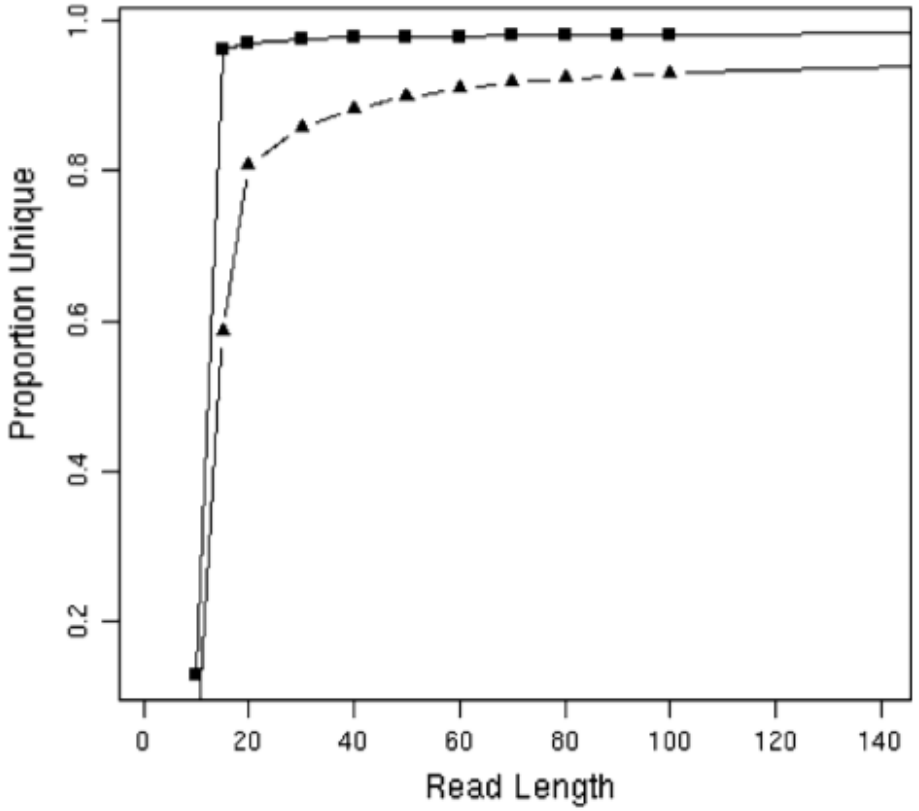
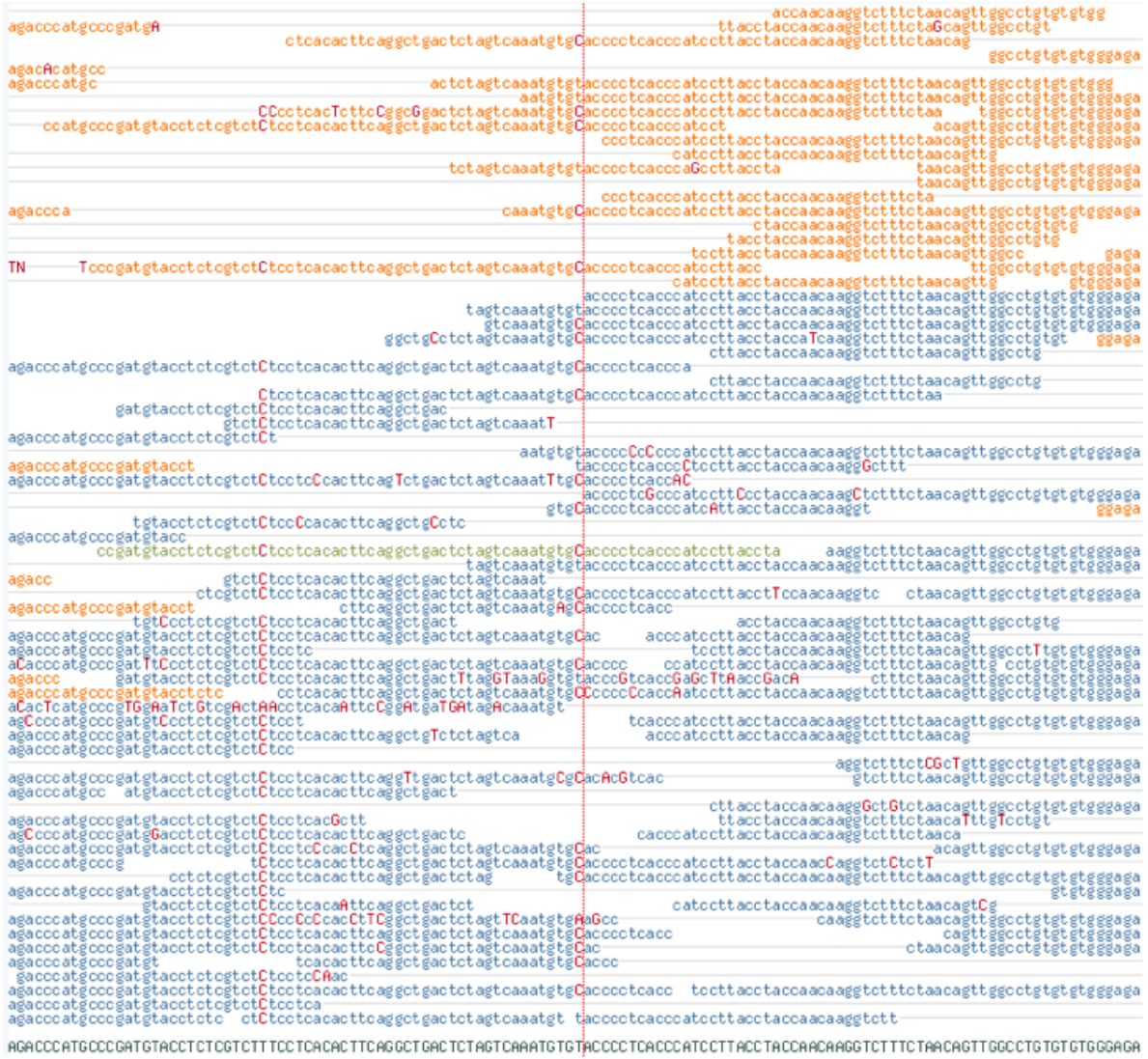


Fig. 1 The proportion of unique sequence in the *Streptococcus suis* (squares) and *Mus musculus* (triangles) genomes for varying read lengths. This graph indicates that read length has a critical affect on the ability to place reads uniquely to the genome

Mouse SNP


LookSeq
Paired pileup , chromosome 3 starting at 120871678 , magnification 1:1
Refresh
Image width 700 insert size 500 MAPQ cutoff 0 show orientation
(Link to this view)
120871700
120871750
Strains
129P2 (38.5X)
129S1 (24.0X)
129S5 (16.8X)
AKR (35.7X)
A/J (23.5X)
BALB (21.9X)
C3H (30.9X)
C57BL (25.8X)
CAST (21.6X)
CBA (25.8X)
DBA (21.7X)
LP/J (24.3X)
NOD (25.3X)
NZO (15.2X)
PWK (22.3X)
SPRET (23.5X)
WSB (16.1X)
Legend
cAGg .. mismatch
catg .. inversion
caTtg .. insertion
ca-g .. deletion
catg .. mate unmapped
catg .. non-unique mapping
catg .. paired read
CATG .. reference

Is this a SNP?



Variant Call Format (VCF)

VCF is a standardised format for storing DNA polymorphism data

- SNPs, insertions, deletions and structural variants
- With rich annotations

Indexed for fast data retrieval of variants from a range of positions

Store variant information across many samples

Record meta-data about the site

- dbSNP accession, filter status, validation status,

Very flexible format

- Arbitrary tags can be introduced to describe new types of variants
- No two VCF files are necessarily the same
  - User extensible annotation fields supported
- Same event can be expressed in multiple ways by including different numbers

Currently v4.1

VCF Format

Header section and a data section

Header

- Arbitrary number of meta-information lines
- Starting with characters '##'
- Column definition line starts with single '#'

Mandatory columns

- Chromosome (CHROM)
- Position of the start of the variant (POS)
- Unique identifiers of the variant (ID)
- Reference allele (REF)
- Comma separated list of alternate non-reference alleles (ALT)
- Phred-scaled quality score (QUAL)
- Site filtering information (FILTER)
- User extensible annotation (INFO)

Example VCF – SNPs/indels

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

VCF Trivia 1

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

What version of the human reference genome was used?

What does the DB INFO tag stand for?

What does the ALT column contain?

At position 17330, what is the total depth? What is the depth for sample NA00002?

At position 17330, what is the genotype of NA00002?

Which position is a tri-allelic SNP site?

What sort of variant is at position 1234567? What is the genotype of NA00002?

More information

SNP Calling + Genotyping

- Samtools
  - http://bioinformatics.oxfordjournals.org/content/25/16/2078.long
  - http://samtools.sourceforge.net
- GATK
  - http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

VCF

- http://bioinformatics.oxfordjournals.org/content/27/15/2156.full
- http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
- http://vcftools.sourceforge.net

Variant calling from next-generation sequence data

- Data Formats + Workflows
- SNP Calling
- **Short Indels**
- Structural Variation
- Experimental Design

Short indel Calling

Small insertions and deletions observed in the alignment of the read relative to the reference genome

BAM format

- I or D character denote indel in the read

Simple method

- Call indels based on the I or D events in the BAM file
  - Samtools varFilter

Factors to consider when calling indels

- Misalignment of the read
  - Alignment scoring - often cheaper to introduce multiple SNPs than an indel
  - Sufficient flanking sequence either side of the read
- Homopolymer runs either side of the indel
- Length of the reads
- Homozygous or heterozygous

Example Indel

Is this an indel?

Is this an indel?

Indel Discovery and Local Realignment

Simple models for calling indels based on the initial alignments show high false positives and negatives

More sophisticated algorithms been developed

- E.g. Dindel, GATK

Example Algorithm overview

- Scan for all I or D operations across the input BAM file
- Foreach I or D operation
  - Create new haplotype based on the indel event
  - Realign the reads onto the alternative reference
  - Count the number of reads that support the indel in the alternative reference
  - Make the indel call

Issues

- Very computationally intensive if testing every possible indel
  - Alternatively test a subset of known indels (i.e. genotyping mode)

Variant calling from next-generation sequence data

- Data Formats + Workflows
- SNP Calling
- Short Indels
- **Structural Variation**
- Experimental Design

Genomic Structural Variation

Large DNA rearrangements (>100bp)

Frequent causes of disease

- Referred to as genomic disorders
- Mendelian diseases or complex traits such as behaviors
  - E.g. increase in gene dosage due to increase in copy number
- Prevalent in cancer genomes

Many types of genomic structural variation (SV)

- Insertions, deletions, copy number changes, inversions, translocations & complex events

Comparative genomic hybridization (CGH) traditionally used to for copy number discovery

- CNVs of 1–50 kb in size have been under-ascertained

Next-gen sequencing revolutionised field of SV discovery

- Parallel sequencing of ends of large numbers of DNA fragments
- Examine alignment distance of reads to discover presence of genomic rearrangments
- Resolution down to ~100bp

Structural Variation

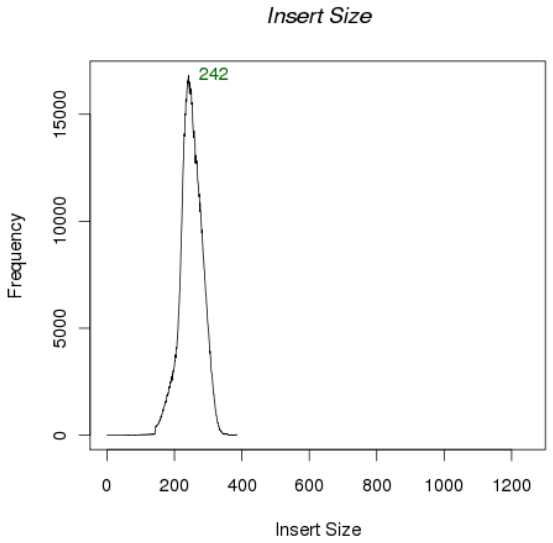
Several types of structural variations (SVs)

- Large Insertions/deletions
- Inversions
- Translocations
- Copy number variations

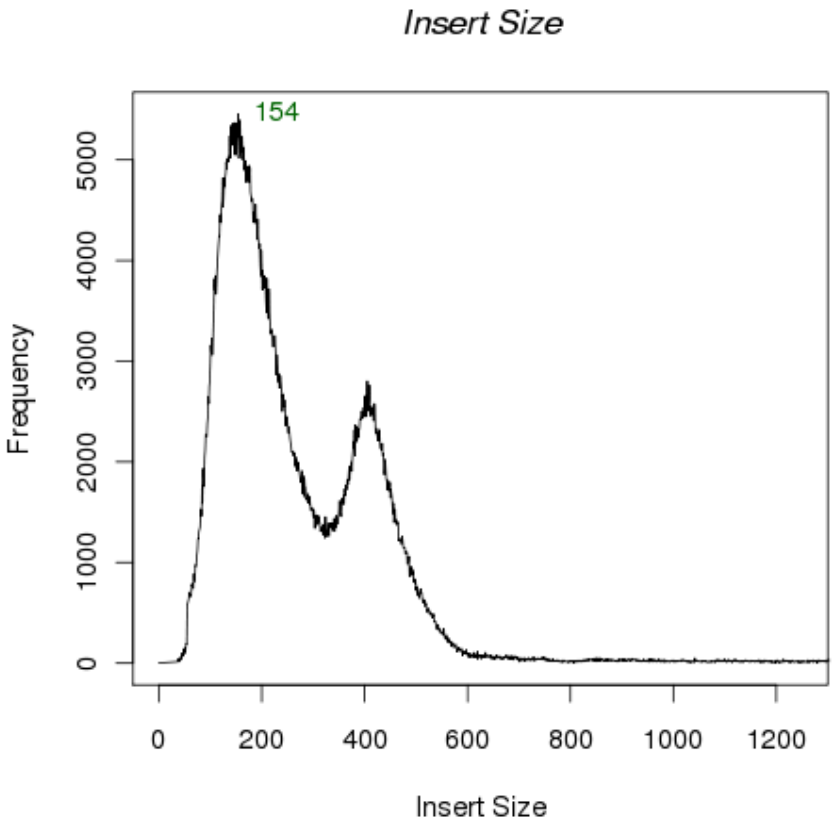
Read pair information used to detect these events

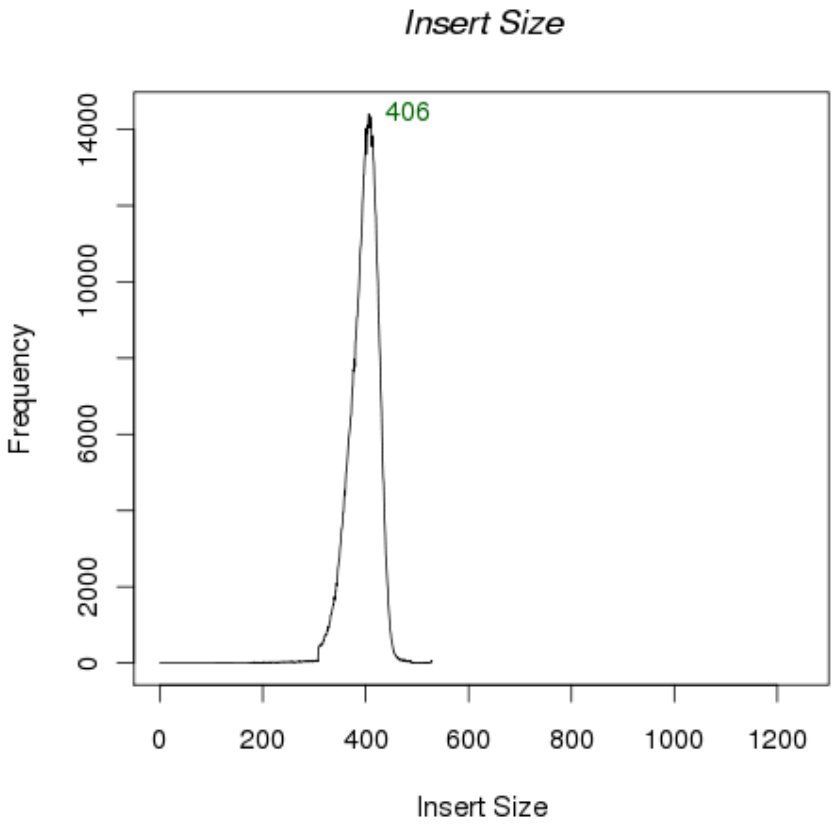
- Paired end sequencing of either end of DNA fragment
  - Observe deviations from the expected fragment size
- Presence/absence of mate pairs
- Read depth to detect copy number variations
- Several SV callers published recently

Run several callers and produce large set of partially overlapping calls

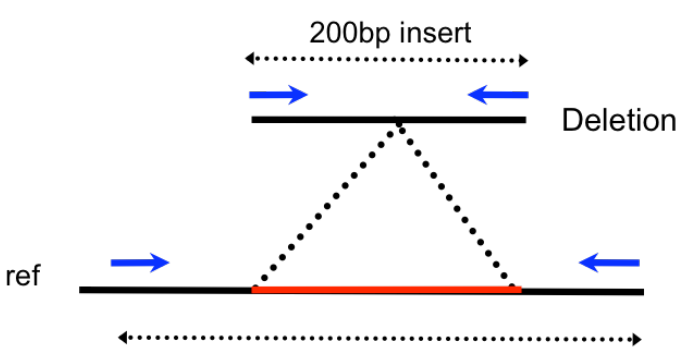


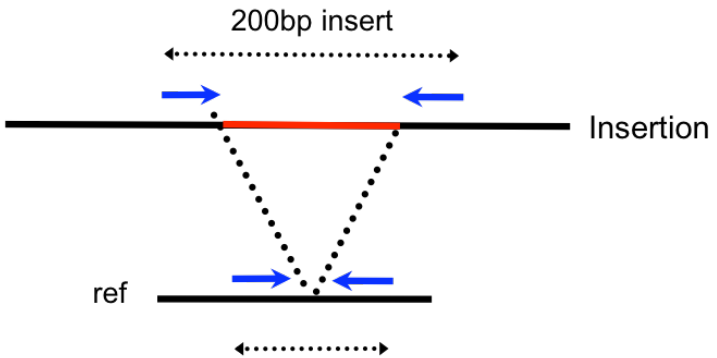
N.B. Fragment Size QC

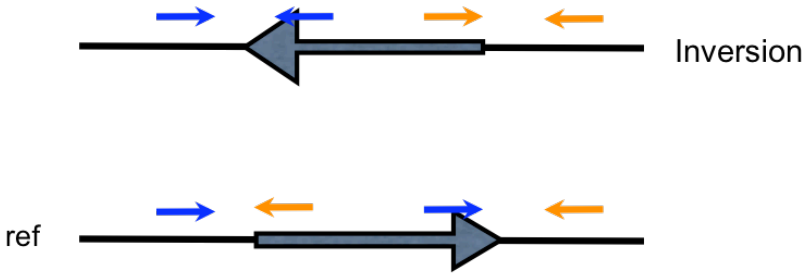


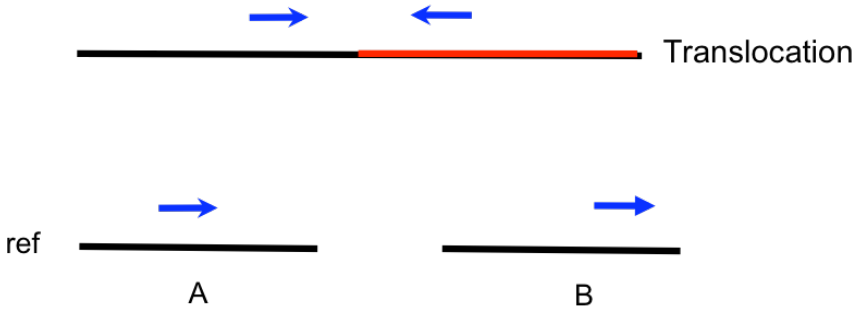


SV Types


200bp insert
Deletion
ref


200bp insert
Insertion
ref

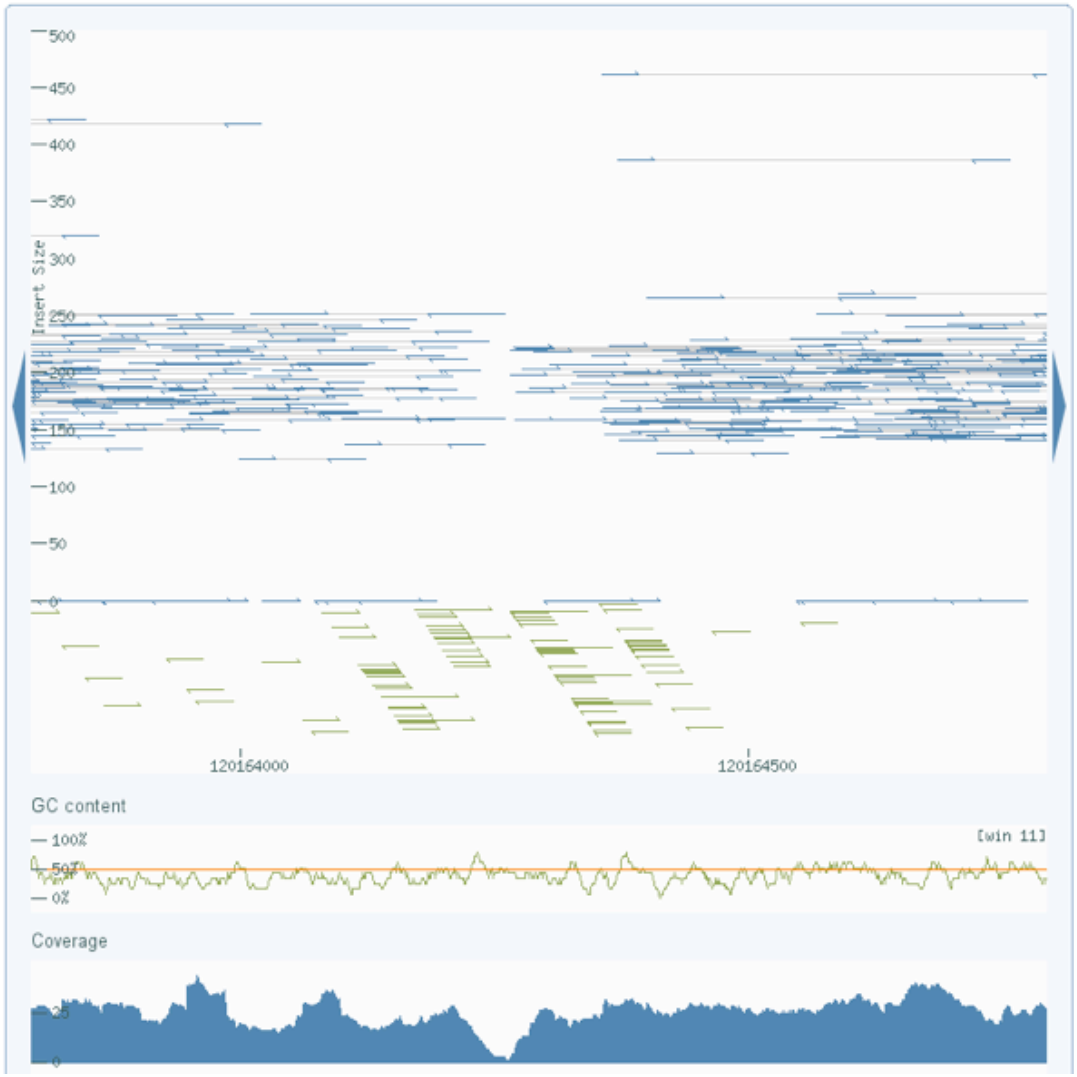

Inversion
ref


Translocation
ref
A
B

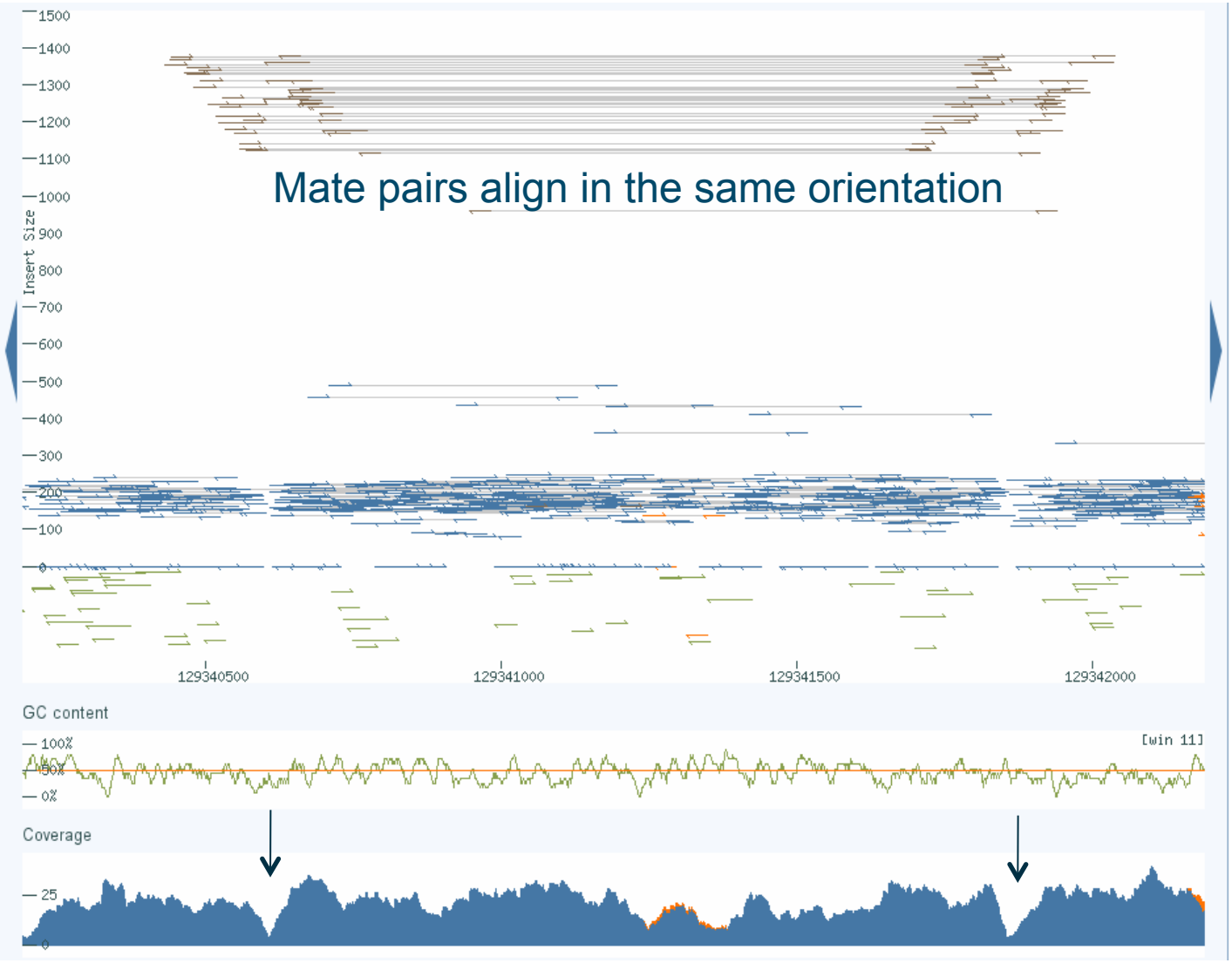
What is this?


Mouse Genomes Project
LookSeq
Paired reads , chromosome 17 starting at 41591345 , magnification 10kb
Refresh
Image width 700 insert size 2500 MAPQ cutoff 0
(Link to this view)
Insert Size
2500
2000
1500
1000
500
0
41595000
41600000
GC content
100%
50%
0%
[win 11]
Coverage
30
20
10
0

What is this?


500
450
400
350
300
250
200
150
100
50
0
Insert Size
120164000
120164500
GC content
100%
50%
0%
[win 11]
Coverage
25
0

What is this?


Mate pairs align in the same orientation
Insert Size
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
129340500
129341000
129341500
129342000
GC content
100%
50%
0%
[win 11]
Coverage
25
0

Mobile Element Insertions

Transposons are segments of DNA that can move within the genome

- A minimal ‘genome’ – ability to replicate and change location
- Relics of ancient viral infections

Dominate landscape of mammalian genomes

- **38-45%** of rodent and primate genomes
- Genome size proportional to number of TEs

Class 1 (RNA intermediate) and 2 (DNA intermediate)

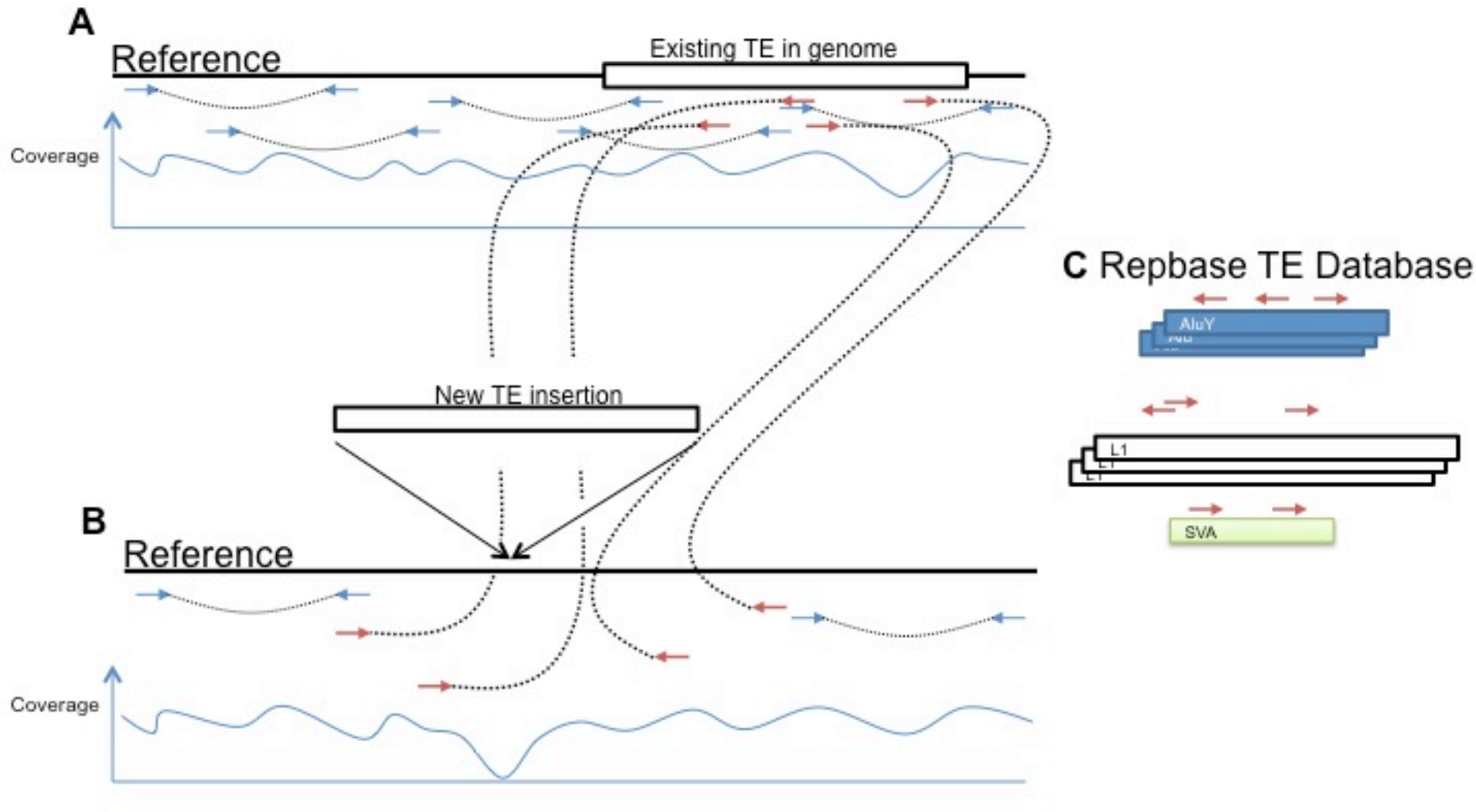
Potent genetic mutagens

- Disrupt expression of genes
- Genome reorganisation and evolution
- Transduction of flanking sequence

Species specific families

- Human: Alu, L1, SVA
- Mouse: SINE, LINE, ERV

Mobile Element Insertions



SVMerge

Initially developed for mouse genomes project

- Several software packages currently available to discover SVs

Various approaches using information from anomalously mapped read pairs OR read depth analysis

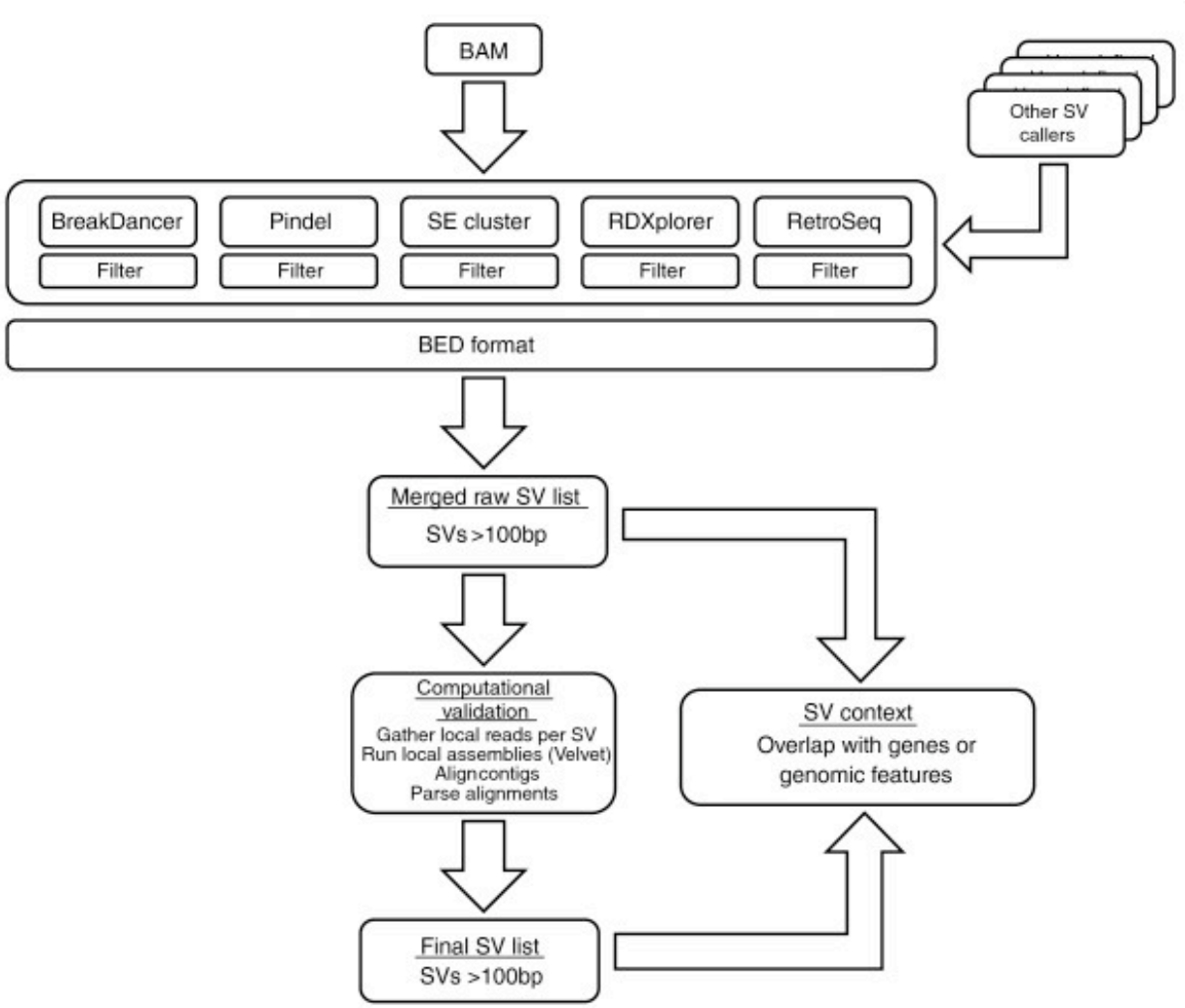
No single SV caller is able to detect the full range of structural variants

- Paired-end mapping information, for example, cannot detect SVs where the read pairs do not flank the SV breakpoints
- Insertion calls made using the split-mapping approach are also size-limited because the whole insertion breakpoint must be contained within a read
- Read-depth approaches can identify copy number changes without the need for read-pair support, but cannot find copy number neutral events

SVMerge, a meta SV calling pipeline, which makes SV predictions with a collection of SV callers

- Input is a BAM file per sample
- Run callers individually + outputs sanitized into standard BED format
- SV calls merged, and computationally validated using local *de novo* assembly
- Primarily a SV discovery/calling + validation tool

SVMerge Workflow



Wong et al (2010)

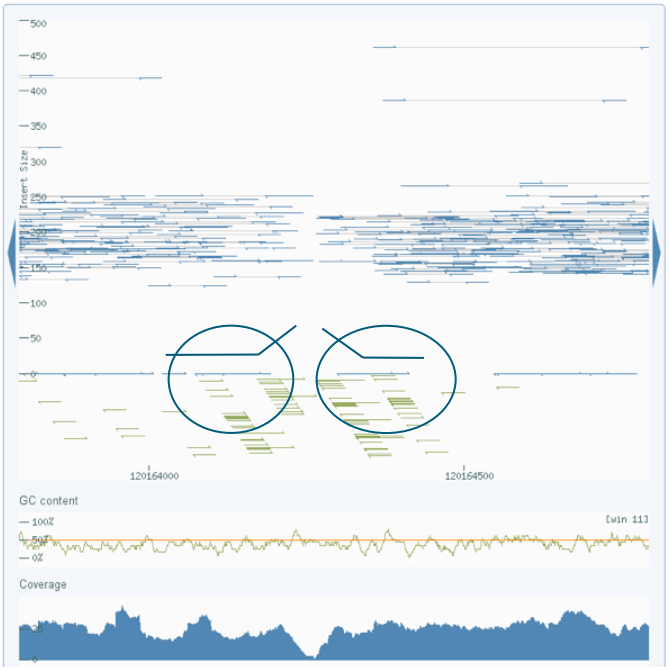
Local Assembly Validation

Key to the approach is the computational validation step

- Local assembly and breakpoint refinement
- All SV calls (except those lacking read pair support e.g. CNG/CNL)

Algorithm

- Gather mapped reads, and any unmapped mate-pairs (<1kb of a insertion breakpoint, <2kb of all other SV types)
- Run local velvet assembly
- Realign the contigs produced with exonerate
- Detect contig breaks proximal to the breakpoint (s)



VCF for Structural Variants

```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly file">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=MEINFO,Number=4,Type=String,Description="Mobile element info of the form NAME,START,END,POLARITY">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM  POS ID REF ALT          QUAL  FILTER  INFO                                                                                  FORMAT       NA00001
2 321682    .  T   <DEL>         6     PASS    IMPRECISE;SVTYPE=DEL;END=321887;SVLEN=-105;CIPOS=-56,20;CIEND=-10,62                  GT:GQ        0/1:12
2 14477084  .  C   <DEL:ME:ALU>  12    PASS    IMPRECISE;SVTYPE=DEL;END=14477381;SVLEN=-297;MEINFO=AluYa5,5,307,+;CIPOS=-22,18;CIEND=-12,32  GT:GQ        0/1:12
3 9425916   .  C   <INS:ME:L1>   23    PASS    IMPRECISE;SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22;MIINFO=L1HS,1,6025,-         GT:GQ        1/1:15
3 12665100  .  A   <DUP>         14    PASS    IMPRECISE;SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND=-500,500           GT:GQ:CN:CNQ ./.:0:3:16.2
4 18665128  .  T   <DUP:TANDEM>  11    PASS    IMPRECISE;SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;CIEND=-10,10                  GT:GQ:CN:CNQ ./.:0:5:8.3
```

VCF Trivia 2

```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly file">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=MEINFO,Number=4,Type=String,Description="Mobile element info of the form NAME,START,END,POLARITY">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM  POS ID REF ALT          QUAL  FILTER  INFO                                                                                  FORMAT       NA00001
2 321682    .  T   <DEL>         6     PASS    IMPRECISE;SVTYPE=DEL;END=321887;SVLEN=-105;CIPOS=-56,20;CIEND=-10,62                  GT:GQ        0/1:12
2 14477084  .  C   <DEL:ME:ALU>  12    PASS    IMPRECISE;SVTYPE=DEL;END=14477381;SVLEN=-297;MEINFO=AluYa5,5,307,+;CIPOS=-22,18;CIEND=-12,32  GT:GQ        0/1:12
3 9425916   .  C   <INS:ME:L1>   23    PASS    IMPRECISE;SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22;MIINFO=L1HS,1,6025,-         GT:GQ        1/1:15
3 12665100  .  A   <DUP>         14    PASS    IMPRECISE;SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND=-500,500           GT:GQ:CN:CNQ ./.:0:3:16.2
4 18665128  .  T   <DUP:TANDEM>  11    PASS    IMPRECISE;SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;CIEND=-10,10                  GT:GQ:CN:CNQ ./.:0:5:8.3
```

How many different types of SV events can be described in this file?
How many different types of insertion are possible? What are they?
What is the confidence interval of the breakpoint for the SV at Chr3 position 12665100?
What does the CN format tag mean?
What is the estimate of copy number at Chr3 position 12665100 for NA00001?

Software Tools

Breakdancer
- Insertions, deletions, inversions, translocations
- http://gmt.genome.wustl.edu/breakdancer/current/

Pindel
- Insertions and deletions
- https://trac.nbic.nl/pindel/

Genome STRIP
- Calling across low coverage populations + genotyping
- http://www.broadinstitute.org/gsa/wiki/index.php/Genome_STRiP

RetroSeq
- Mobile element insertion discovery
- https://github.com/tk2/RetroSeq

Breakpoint assembly
- Tigra: http://genome.wustl.edu/software/tigra_sv
- SVMerge: http://svmerge.sourceforge.net/

Integrating Calls
- SVMerge: http://svmerge.sourceforge.net/

Variant calling from next-generation sequence data

- Data Formats + Workflows
- SNP Calling
- Short Indels
- Structural Variation
- **Experimental Design**

Experimental Design

Choosing right sequencing technology to get optimal results for experiment

Experiment 1: *"I want to determine the genome of a new fungi species with no closely related reference genome"*

- Whole-genome sequencing
- De novo assembly with no reference
  - Longer reads might be more useful – 454?
- Mixture of fragment sizes
  - 200, 500, 3kb, 5kb, 10kb
  - Short range pairing information and long range information for scaffolding

Experiment 2: *"I want to measure the relative expression level differences of one yeast species under different environmental conditions"*

- Sequence the transcriptome (RNA-seq)
- Illumina or SOLiD sequencing for high depth
  - Multiplex the sequencing into a singlle lane
- Measure the relative expression levels by aligning
  - e.g. use Cufflinks to detect differential expression across the samples

Experimental Design

Experiment 4: *"I want to catalog all of the structural variants in a human cancer cell vs. the normal cell for as little cost as possible"*

- Fragment coverage vs. sequence coverage
- SVs are called from discordant read pairs – long range information
  - Sequence coverage not important
  - Require fragment coverage
- Sequence multiple paired libraries with short read length
  - E.g. 1000bp in total capacity
  - 100bp reads = 5RPs = 5 fragments x 500bp per fragment = 2.5Kbp fragment coverage
  - 40bp reads = 25RPs = 25 fragments x 500bp per fragment = 12.5Kbp fragment coverage
  - More fragments sequenced = more independent sources of evidence to call structural variants

Experiment 5: "*I have 3 patients with a rare condition and want to find the causative variant*"

- High depth sequencing (20x?) per patient. Illumina or SOLiD or complete genomics
- Exome sequencing – 1 lane per patient
- SNPs + short indels
- Exclude all common variation (dbSNP + 1000Genomes)
- Is there a shared truncating variant? If not – is there a shared truncating structural variant?

Q&A

Questions from you?

Slides: http://www.slideshare.net/thomaskeane/