MPI EVA: High-throughput sequencing of ancient and modern DNA samples

for studying the genetic history of humankind…

MPI for Evolutionary Anthropology
Evolutionary Genetics / Bioinformatics group
Martin Kircher

May 16 2010, CSHL

pictures by Illumina, Roche, MPI-EVA

NGS/Sequencing pipelines ...

http://sarahpalintruthsquad.files.wordpress.com/2008/09/alaska-pipeline.jpg

NGS/Sequencing pipelines ...

http://sarahpalintruthsquad.files.wordpress.com/2008/09/alaska-pipeline.jpg

http://startupblog.files.wordpress.com/2008/09/pipeline1.jpg

http://www.undergroundsolutions.com/images/engineering/PipeIronCorroded.png

unknown source

Sequencing ...

Sequencing instruments at MPI EVA

2 x

ABI 3730/3730xl

2 x

454 FLX Titanium

5 x

Illumina Genome Analyzer

ABI SOLiD

HeliScope

> 170 Illumina runs over the last 2.5 years

Stored by run on NFS (currently about 40Tb, including files of intermediate processing steps)

(Incomplete) Meta-data in XML, text files and encoded in file names

Processed with different program versions...

Data partially duplicated in user home directories

Non-uniform collection:

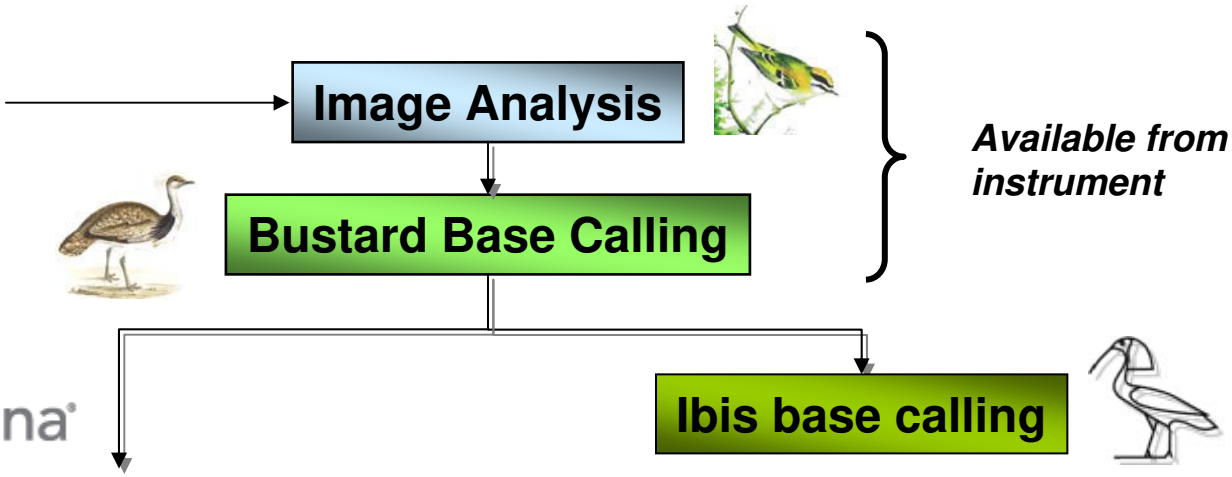
- **Varying read length**
- **Single read, Paired ends, With/without sample index**
- **Different instrument versions**
- **Experimental protocols (shotgun, RNAseq, smallRNAs, ChipSeq, ...),**
- **Different samples / species and different people analyzing the data**

...

PIPELINE

pictures by Illumina, Roche, ABI, Helicos

EVA data analysis pipeline



- → HTML report generation (cluster densities, intensity development)
- → ELAND control read (φX) mapping
- → Chastity filtering
- → HTML report generation (alignment, filtering ➔ data quality)

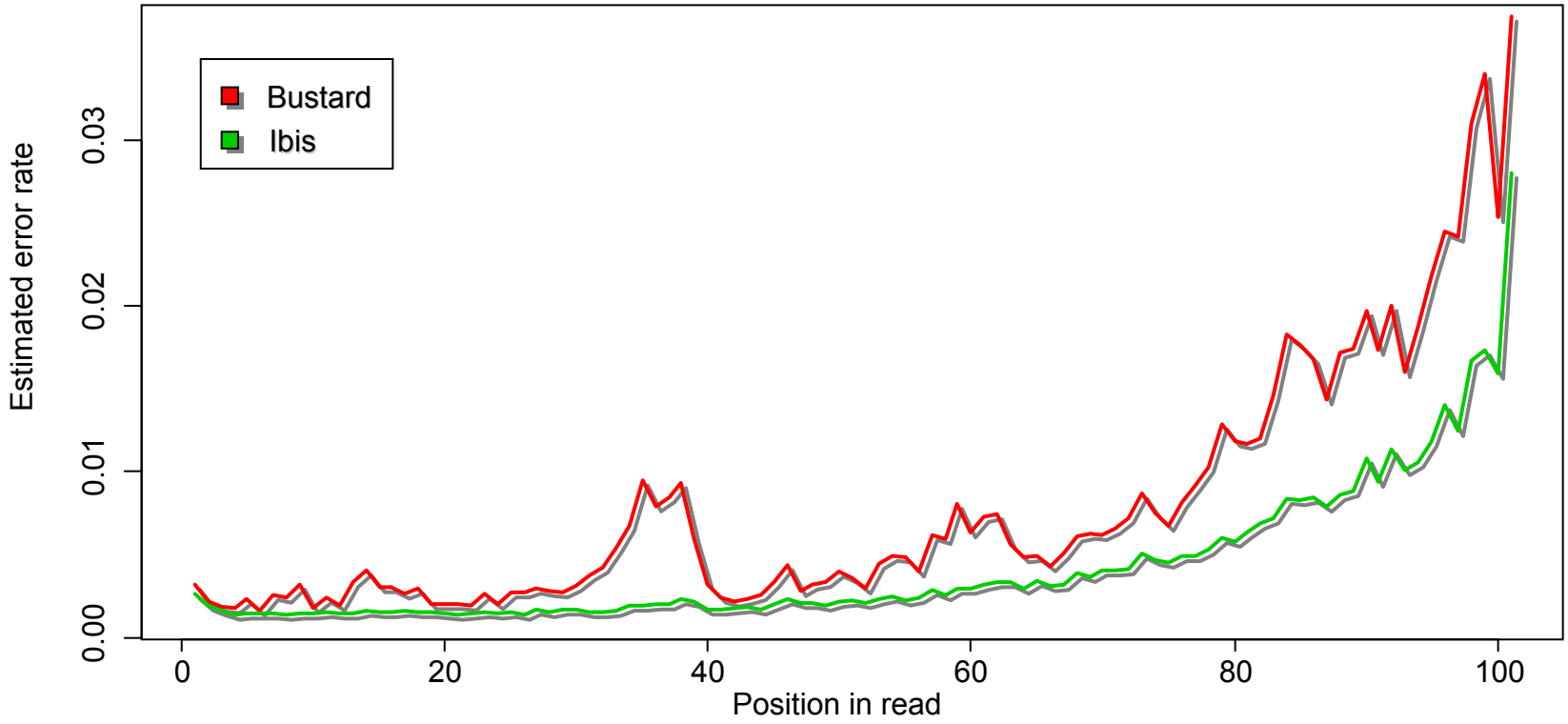
- → Error profile extraction
- → Index sequence handling
- → Adapter trimming / PE read merging / chimera filtering
- → Quality filtering
- → Complexity filtering
- → Bowtie / Tophat / BWA alignment

Intensity files

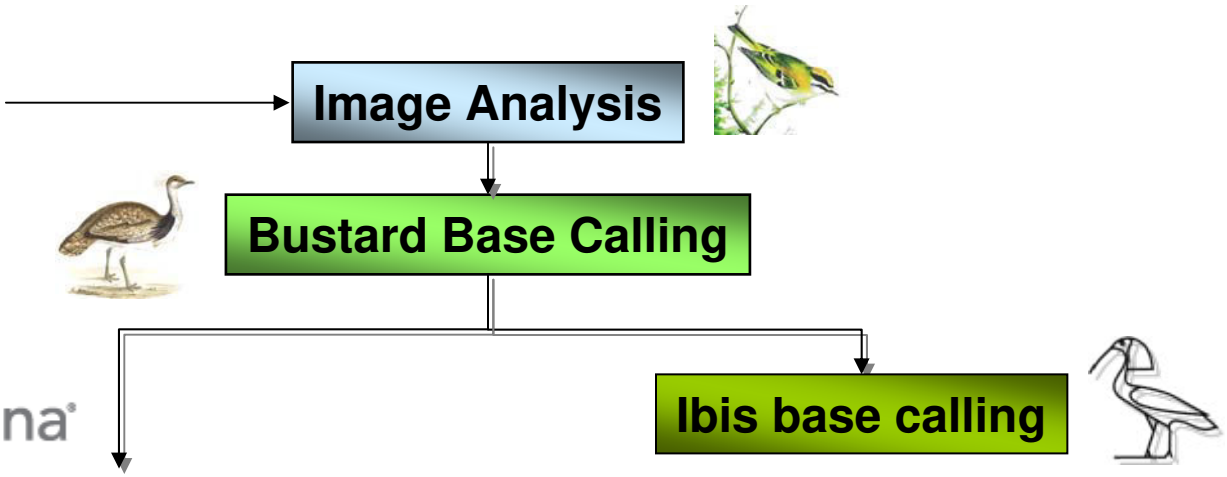
Lane	Tile	X	Y	Cycle 1				Cycle 2				Cyc	
4	20	853	419	-28.8	-27.1	23.7	696.7	156.0	537.9	4.3	72.7	17.0	-0.5
4	20	157	530	171.4	619.2	182.3	136.9	566.1	564.0	3.5	5.9	150.8	165.4
4	20	170	1889	-150.2	5.9	-34.0	609.9	236.5	688.8	-100.7	-39.3	-5.1	391.7
4	20	386	1680	1345.5	821.3	77.8	-16.0	-70.5	-203.1	996.3	513.4	837.4	524.8
4	20	702	1599	-46.7	-6.4	3.6	537.1	-39.1	-54.9	856.6	412.6	246.0	780.2
4	20	887	739	55.9	106.4	953.4	481.3	0.0	0.0	0.0	0.0	1033.8	722.3
4	20	1229	894	1272.7	728.7	-63.2	-63.2	1072.6	604.6	-15.0	131.8	139.8	172.1
4	20	1417	87	1092.7	651.8	-12.9	-27.1	254.5	118.7	-86.4	548.1	102.6	480.0
4	20	1163	220	-85.2	28.0	637.9	380.5	664.0	375.2	10.7	25.1	84.5	402.1
4	20	1147	2015	933.3	540.1	-73.1	-112.1	-79.1	-40.8	704.1	285.0	134.4	424.4
4	20	854	1797	203.7	791.5	-15.4	-31.5	1149.6	660.5	-112.9	-33.1	938.7	693.3
4	20	759	565	982.9	712.4	-104.8	-68.4	-17.7	-21.7	811.1	473.2	231.2	678.2
4	20	41	1833	852.1	385.5	87.5	63.1	-48.1	83.5	17.1	534.5	-31.8	-135.9
4	20	197	1417	142.5	716.9	-164.1	33.8	-108.3	-219.3	-30.4	711.7	4.5	-24.6
4	20	88	1135	-63.8	-90.1	30.1	778.9	-80.0	-4.8	42.7	733.2	-26.4	-47.1
4	20	483	24	151.5	452.3	-31.2	6.8	727.2	470.6	-11.0	-56.3	144.1	492.5
4	20	1768	1918	75.6	422.8	-15.4	-40.0	180.4	397.4	-51.4	0.4	1024.9	590.8
4	20	890	1588	36.5	122.1	22.2	438.4	40.2	15.2	914.3	402.7	25.0	559.2
4	20	770	523	0.0	0.0	0.0	0.0	-169.1	-39.5	882.7	574.1	911.5	603.8
4	20	1433	1910	-117.9	-66.0	994.1	471.4	76.5	505.7	-6.3	17.7	832.8	477.1
4	20	433	1675	-88.5	-18.8	1266.4	475.3	900.7	646.3	-19.7	-97.6	960.4	566.3
4	20	675	729	969.0	636.9	5.8	-7.8	1090.1	789.1	-23.0	-153.4	878.7	544.6
4	20	387	1704	658.1	387.2	0.6	40.1	932.9	507.5	-65.9	24.4	900.0	520.8

IBIS – Improved Base Identification System

Kircher M, Stenzel U, Kelso J: Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. 2009 Aug 14;10(8):R83. PMID: 19682367; Download http://bioinf.eva.mpg.de/ibis/



EVA data analysis pipeline



→ HTML report generation

densities, intensity deve

→ ELAND control read (φX) mapping

→ Chastity filtering

→ HTML report generation (alignment, filtering ➔ data quality)

User-defined via webform -> creating processing cmdline

→ Error profile extraction

→ Index sequence handling

→ Adapter trimming / PE read merging / chimera filtering

→ Quality filtering

→ Complexity filtering

→ Bowtie / Tophat / BWA alignment

Multiplex Paired End: 100506_SOLEXA-GA03_0001_PEi_KP_chimp1 (lane 3, 4, 5, 6)

Read length: 101 101
Index read length: 7
Library type: Illumina Multiplex library

Run processing:

COMMENT: Generally we run our in-house base caller (Ibis) for all Illumina sequencing runs. If you do not want to have the results for Ibis or additionaly would like to have the output of the standard base caller Bustard, you have to change the below options. If you just would like to retrieve sequences for all clusters without any further processing (except splitting by index, if applicable), please change the lower radio button to "Only raw sequences".

Bustard base calling: No Bustard base calling results

Ibis base calling: (•) Yes ( ) No

Data processing: (•) Complete processing ( ) Only raw sequences

2

er R2	Sample owner	Status

100506_SOLEXA-GA03_0001_PEi_KP_chimp1

Please provide contact information:

Your Email: martin.kircher@eva.mpg.de

Please define run and lane(s) for which parameters should be applied:

Attention: This defines parameters of the run, not the library. In case your libr
have an index read, but the run has: define that the run has an index read.

Type of the run: ( ) Single Read (•) Paired End

With index read: (•) Yes ( ) No

Lane(s): 1 2 **3** **4** **5** **6** 7 8

1

Next Back

Lane	Sample name	conc. (pM)	Seq.primer R1	Index Primer	Seq.Prir
1	Agnagui	11	Genomic R1	Index Seq	MP R2

Multiplex Paired End: 100506_SOLEXA-GA03_0001_PEi_KP_chimp1 (lane 3, 4, 5, 6)

The library protocol and run allow for an index read, specify the index sequences used:

COMMENT: If no index is specified the index read information will be dropped. Therefore, do not select an index if only one index is present in the library or you are for some other reason not interested in the index read. If control sequences have been spiked into your lane, don't forget to include this index at the very bottom of the list.

HINT: Press CTRL for selecting multiple disconnected ranges of index sequences.

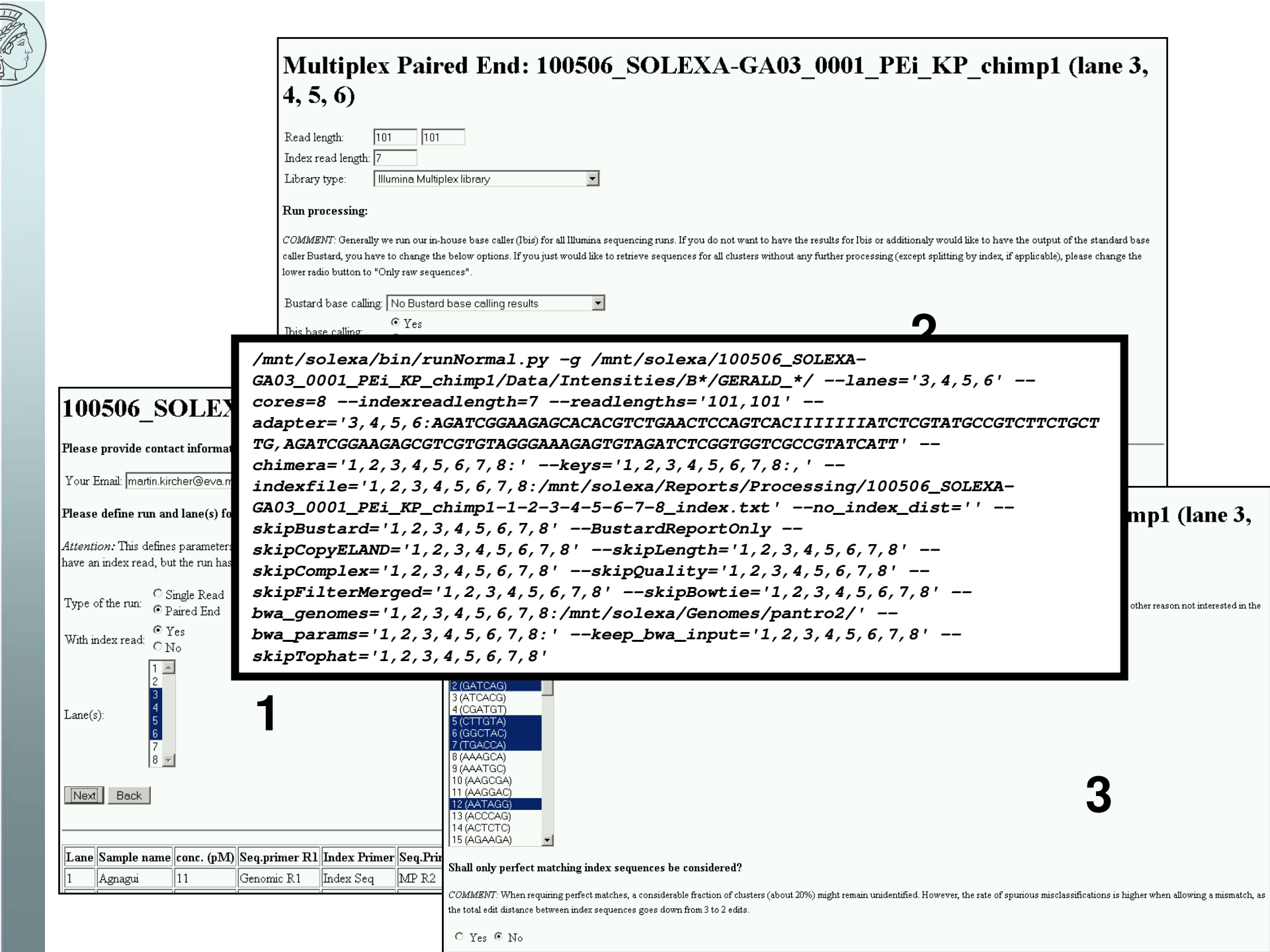
1 (ACAGTG)
2 (GATCAG)
3 (ATCACG)
4 (CGATGT)
5 (CTTGTA)
6 (GGCTAC)
7 (TGACCA)
8 (AAAGCA)
9 (AAATGC)
10 (AAGCGA)
11 (AAGGAC)
12 (AATAGG)
13 (ACCCAG)
14 (ACTCTC)
15 (AGAAGA)

3

Shall only perfect matching index sequences be considered?

COMMENT: When requiring perfect matches, a considerable fraction of clusters (about 20%) might remain unidentified. However, the rate of spurious misclassifications is higher when allowing a mismatch, as the total edit distance between index sequences goes down from 3 to 2 edits.

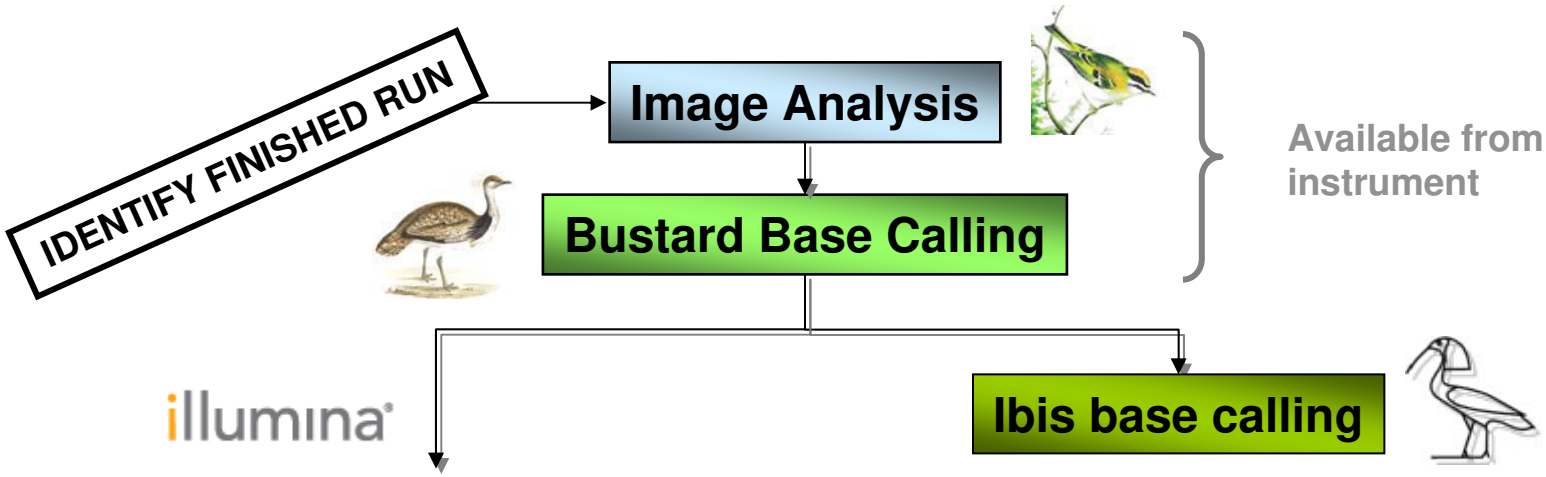
( ) Yes (•) No


Multiplex Paired End: 100506_SOLEXA-GA03_0001_PEi_KP_chimp1 (lane 3, 4, 5, 6)
Read length: 101 101
Index read length: 7
Library type: Illumina Multiplex library
Run processing:
COMMENT: Generally we run our in-house base caller (Ibis) for all Illumina sequencing runs. If you do not want to have the results for Ibis or additionaly would like to have the output of the standard base caller Bustard, you have to change the below options. If you just would like to retrieve sequences for all clusters without any further processing (except splitting by index, if applicable), please change the lower radio button to "Only raw sequences".
Bustard base calling: No Bustard base calling results
Ibis base calling: Yes
2
/mnt/solexa/bin/runNormal.py -g /mnt/solexa/100506_SOLEXA-GA03_0001_PEi_KP_chimp1/Data/Intensities/B*/GERALD_*/ --lanes='3,4,5,6' --cores=8 --indexreadlength=7 --readlengths='101,101' --adapter='3,4,5,6:AGATCGGAAGAGCACACGTCTGAACTCCAGTCACIIIIIIIATCTCGTATGCCGTCTTCTGCTTG,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT' --chimera='1,2,3,4,5,6,7,8:' --keys='1,2,3,4,5,6,7,8:,' --indexfile='1,2,3,4,5,6,7,8:/mnt/solexa/Reports/Processing/100506_SOLEXA-GA03_0001_PEi_KP_chimp1-1-2-3-4-5-6-7-8_index.txt' --no_index_dist='' --skipBustard='1,2,3,4,5,6,7,8' --BustardReportOnly --skipCopyELAND='1,2,3,4,5,6,7,8' --skipLength='1,2,3,4,5,6,7,8' --skipComplex='1,2,3,4,5,6,7,8' --skipQuality='1,2,3,4,5,6,7,8' --skipFilterMerged='1,2,3,4,5,6,7,8' --skipBowtie='1,2,3,4,5,6,7,8' --bwa_genomes='1,2,3,4,5,6,7,8:/mnt/solexa/Genomes/pantro2/' --bwa_params='1,2,3,4,5,6,7,8:' --keep_bwa_input='1,2,3,4,5,6,7,8' --skipTophat='1,2,3,4,5,6,7,8'
100506_SOLEX
Please provide contact informat
Your Email: martin.kircher@eva.m
Please define run and lane(s) fo
Attention: This defines parameters
have an index read, but the run has
Type of the run: Single Read Paired End
With index read: Yes No
Lane(s): 1 2 3 4 5 6 7 8
Next Back
1
mp1 (lane 3,
other reason not interested in the
2 (GATCAG)
3 (ATCACG)
4 (CGATGT)
5 (CTTGTA)
6 (GGCTAC)
7 (TGACCA)
8 (AAAGCA)
9 (AAATGC)
10 (AAGCGA)
11 (AAGGAC)
12 (AATAGG)
13 (ACCCAG)
14 (ACTCTC)
15 (AGAAGA)
3
Lane Sample name conc. (pM) Seq.primer R1 Index Primer Seq.Prir
1 Agnagui 11 Genomic R1 Index Seq MP R2
Shall only perfect matching index sequences be considered?
COMMENT: When requiring perfect matches, a considerable fraction of clusters (about 20%) might remain unidentified. However, the rate of spurious misclassifications is higher when allowing a mismatch, as the total edit distance between index sequences goes down from 3 to 2 edits.
Yes No

GALAXY – a wish list

Tracking for Illumina
(Re-)base calling
Primary data processing
Efficient data storage
Processing from GALAXY and shell (API)

Tracking / Re-base calling



→ HTML report generation (cluster densities, intensity development)

EXTRACT RUN INFO AND CALL GERALD TO OBTAIN RUN QUALITY REPORT

→ Chastity filtering

→ HTML report generation (alignment, filtering ➔ data quality)

→ Error profile extraction

→ Index sequence handling

REDO BASE CALLING
→ OBTAIN QUALITY MEASURE FROM CONTROL READS
→ OBTAIN NEW SEQUENCE FILES

→ Quality filtering

→ Complexity filtering

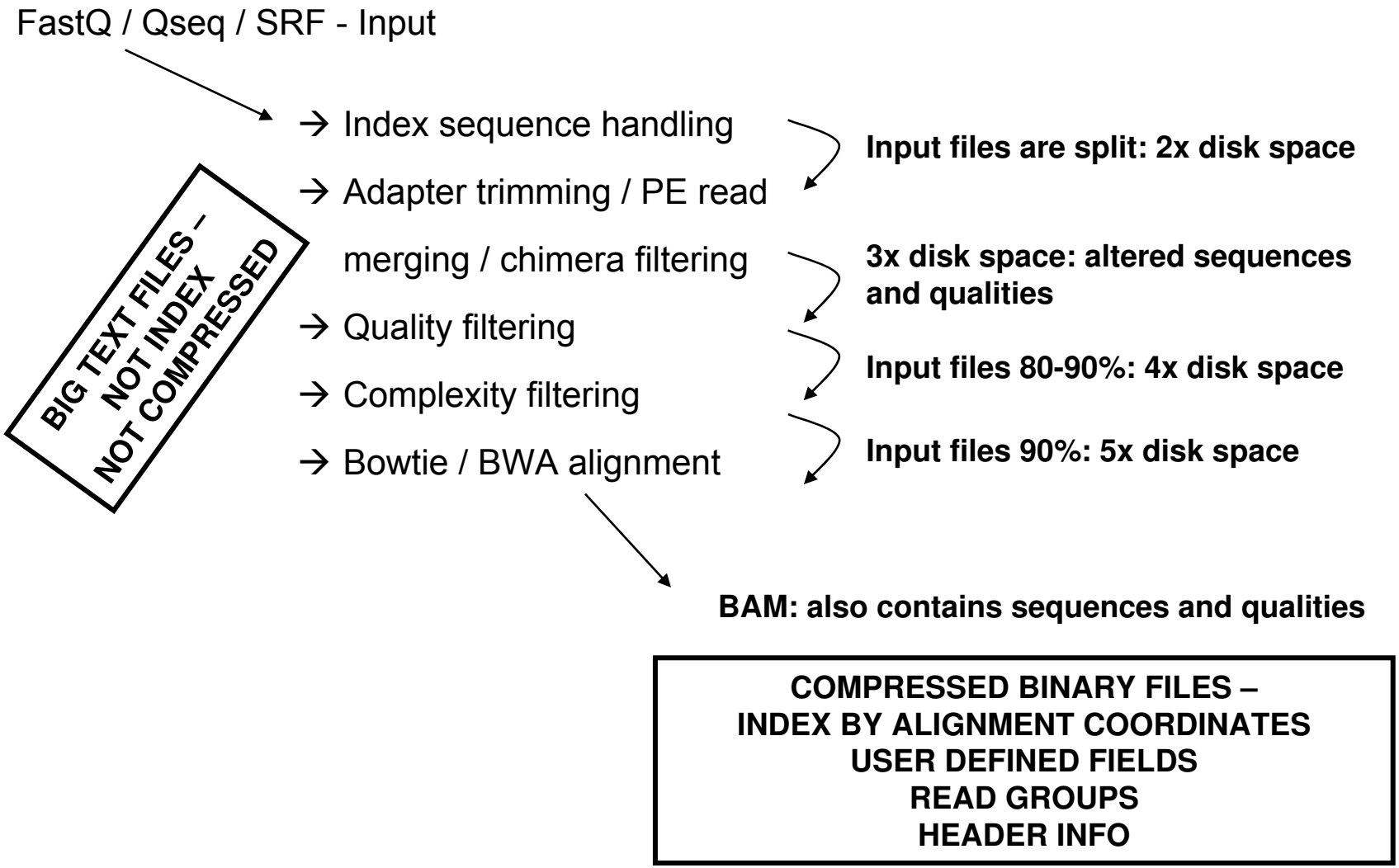
→ Bowtie / BWA alignment

Primary data processing

→ DOCUMENT PRIMARY DATA PROCESSING PARAMETERS
→ ALLOW REPROCESING TO NON-COMPUTER SKILLED PERSON

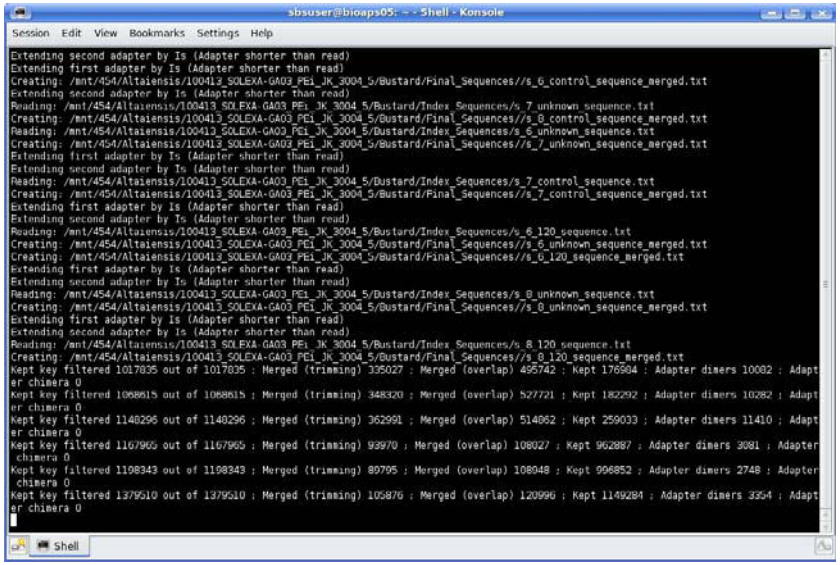
Merging
KeyFilter
Index sort/filter
Run quality assessment
Adapter trimming
Quality filter
Mapping/Alignment
Sequence complexity filter

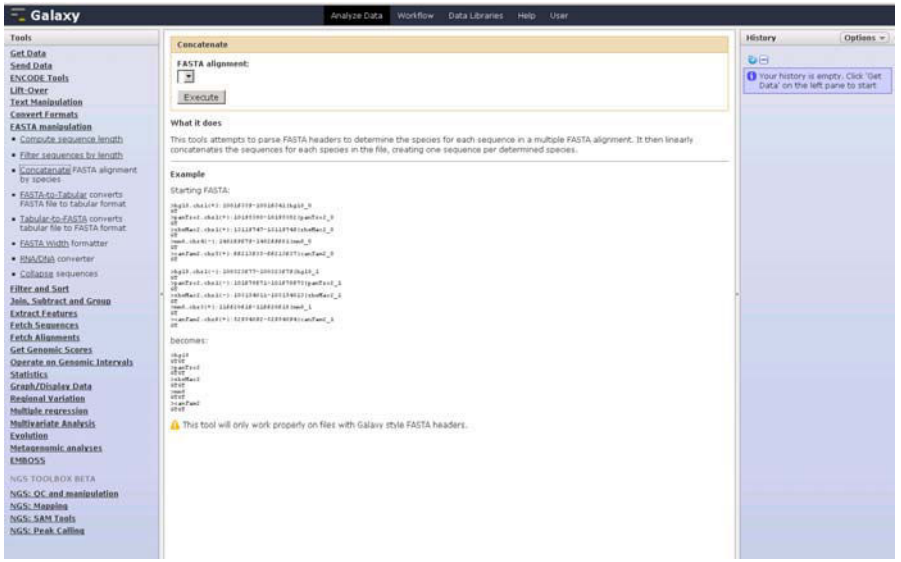
Efficient data storage



GALAXY and shell processing

- Two user groups accessing the same data





- Give galaxy access to shell processed results and use an Galaxy API for documentation of processing

Thank you!

(for providing GALAXY and your attention)

Wish list:

Tracking for Illumina
(Re-)base calling
Primary data processing
Efficient data storage
Processing from GALAXY and shell (API)