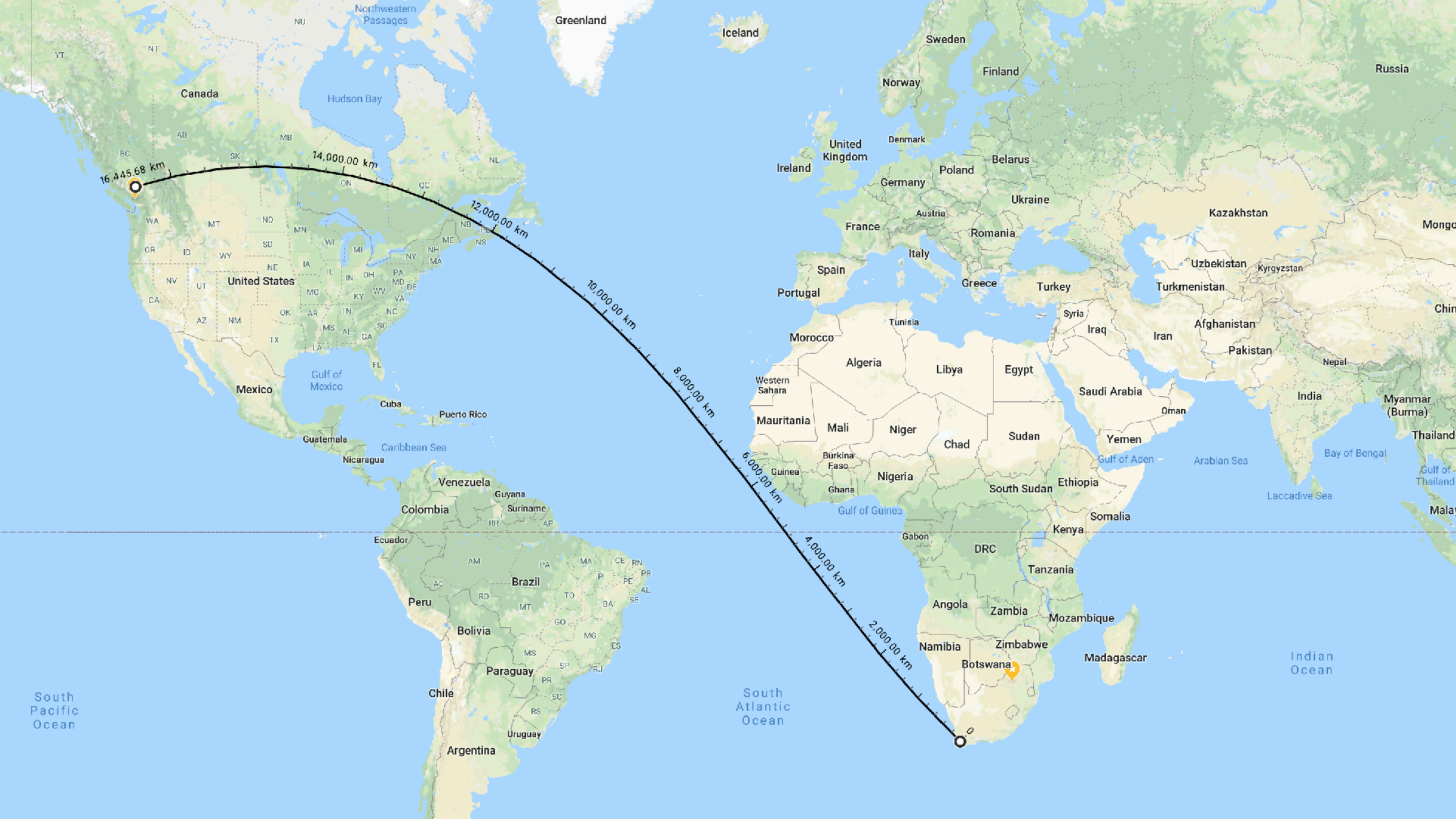
A thousand genomes in a Galaxy server far, far away…

Charles Hefer
Galaxy Africa
3-5 April 2018


16,445.68 km
14,000.00 km
12,000.00 km
10,000.00 km
8,000.00 km
6,000.00 km
4,000.00 km
2,000.00 km
0
Northwestern Passages
Greenland
Iceland
Hudson Bay
Canada
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Caribbean Sea
Guatemala
Nicaragua
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Argentina
Uruguay
South Pacific Ocean
South Atlantic Ocean
Sweden
Finland
Norway
Russia
United Kingdom
Ireland
Denmark
Belarus
Poland
Germany
Ukraine
France
Austria
Romania
Italy
Spain
Portugal
Greece
Turkey
Kazakhstan
Uzbekistan
Kyrgyzstan
Turkmenistan
Syria
Iraq
Iran
Afghanistan
Pakistan
Nepal
India
Myanmar (Burma)
Thailand
Morocco
Tunisia
Algeria
Libya
Egypt
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Saudi Arabia
Oman
Yemen
Gulf of Aden
Arabian Sea
Bay of Bengal
Laccadive Sea
Burkina Faso
Guinea
Ghana
Nigeria
Gulf of Guinea
South Sudan
Ethiopia
Somalia
Kenya
Gabon
DRC
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
Indian Ocean


UBC

Genetic Improvement of Poplar Trees as a Canadian Bioenergy Feedstock

Carl Douglas

Shawn Mansfield

Quentin Cronk

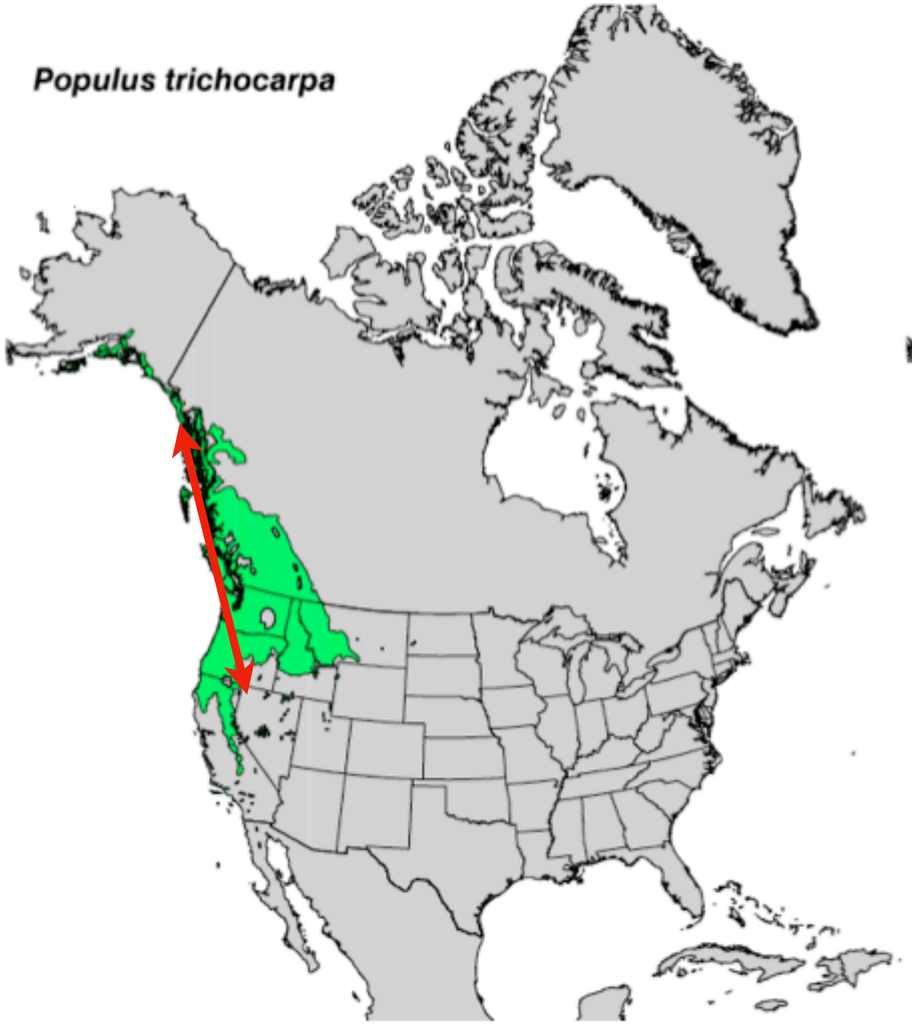
GenomeCanada

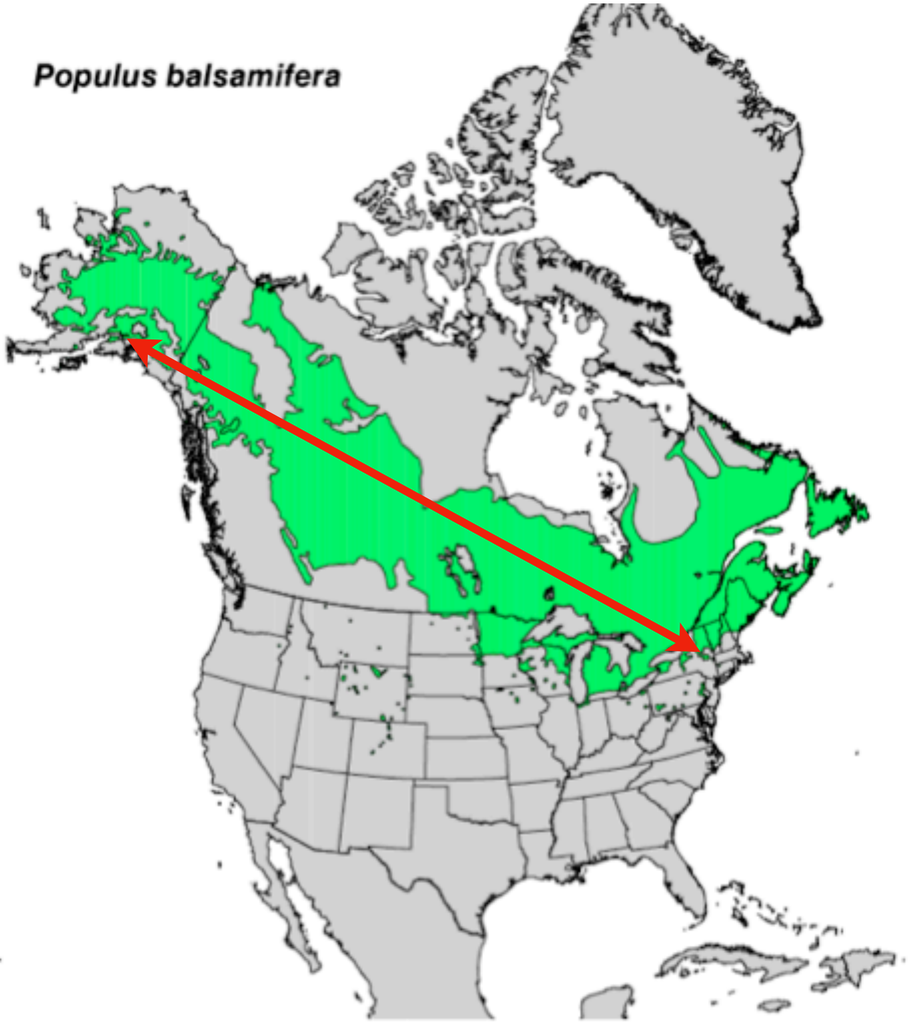
Poplar genome resources

- Poplar - First woody genome sequenced (back in 2006)
  - Followed by *Eucalyptus, Pinus, Picea*
  - Relatively small genome ~460Mbp
  - 19 Chromosomes
  - 41,335 loci and 73,031 proteins (www.phytozome.org)
- Genus consists of 29 species (6 sections)

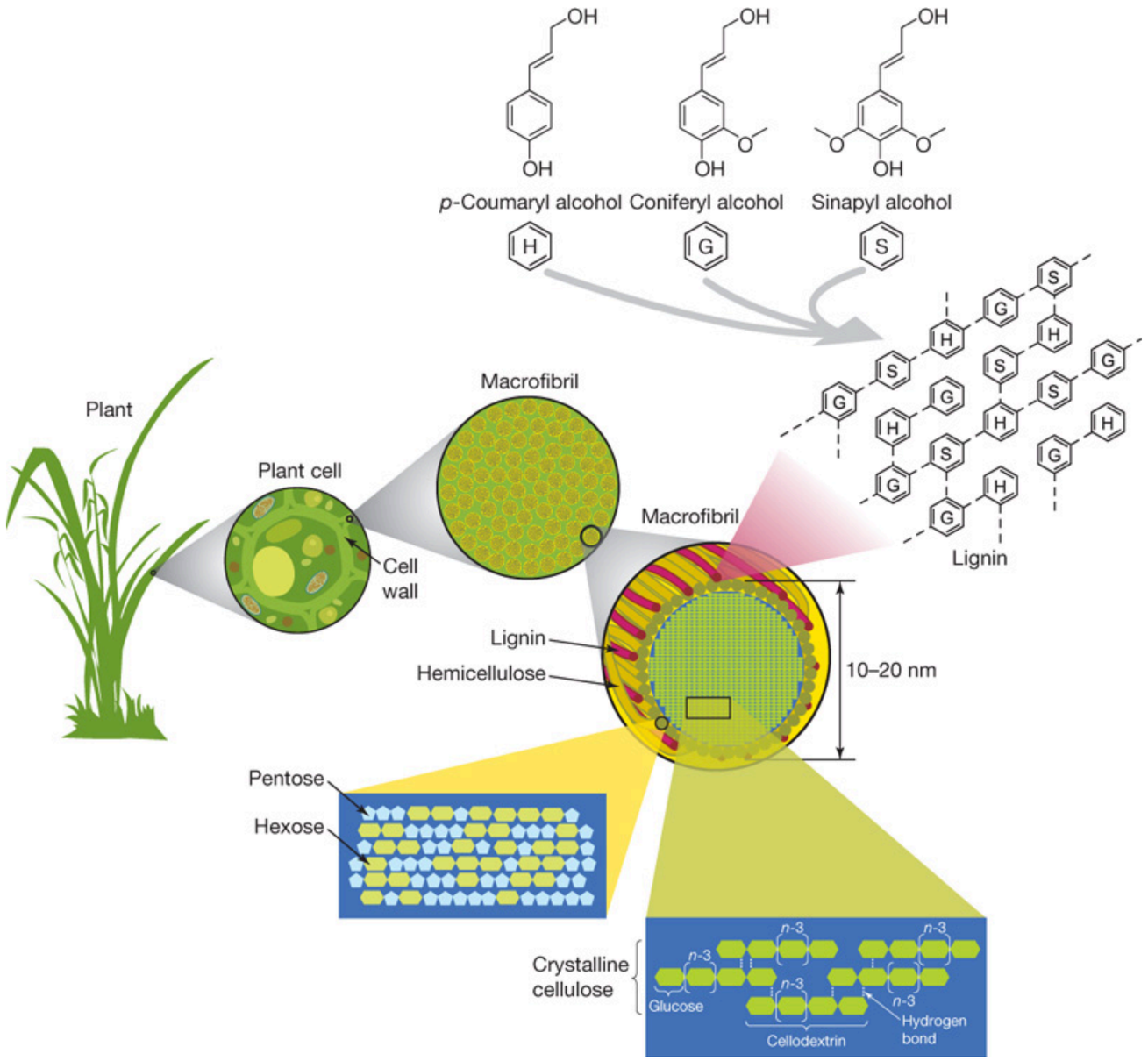
Tuscan *et al*, 2006

Poplar





Lignocellulose as bioenergy feedstock



Poplar as a cellulose crop

Cellulose crop	ton/ha/yr
Sorghum	40 - 55
Sugarcane	80 - 120
Switchgrass	14 - 18
Miscanthus	30 - 41
Corn Stover	3 - 5
Poplar	12 - 24

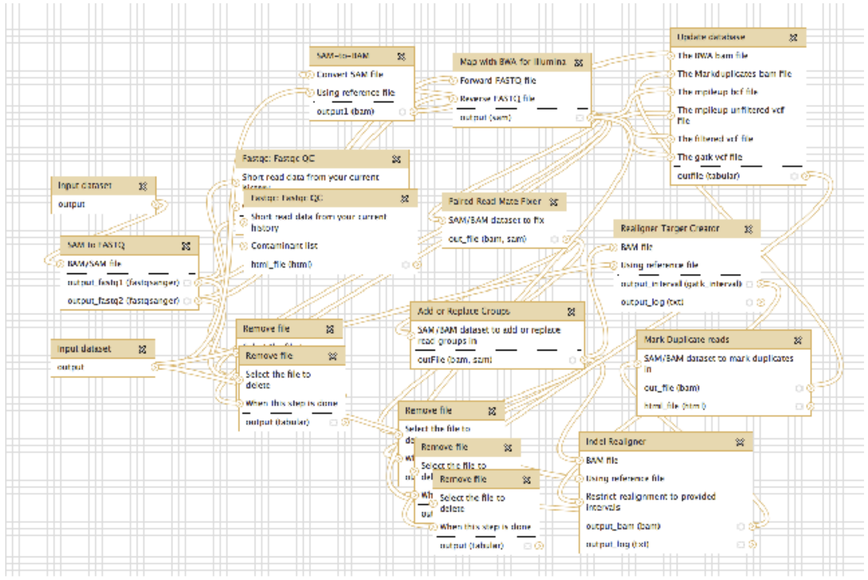
The PopCan project

- Genetic improvement of Poplar sp. for biofuel production
- 2 Main aims
  - Search for adaptive traits within and between species
  - Search for trait associated with:
    - growth
    - wood properties (cellulose to ethanol)
- In essence: Characterize poplar in terms of variation, adaptive traits and functional elements associated with growth and wood properties

P. trichocarpa

Genome Resequencing

- 1,113 samples sequenced over 3 years
  - Between 15X and 30X coverage
  - Illumina HiSeq 2000
- Samples completed in batches
  - rsync from sequencing centre
- Single workflow needed to handle all the samples.
  - Genome resequencing, focus on SNP discovery



Galaxy / WEST

Genome BritishColumbia
Leading ▸ Investing ▸ Connecting

Hardware

east

R910

64 CPUs
512Gb RAM

R910

west

Data analysis

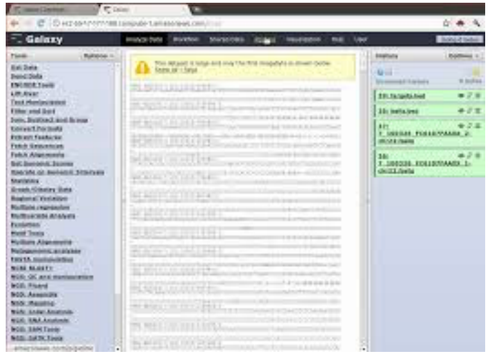
Mirror of “final data”

Mirror of primary data

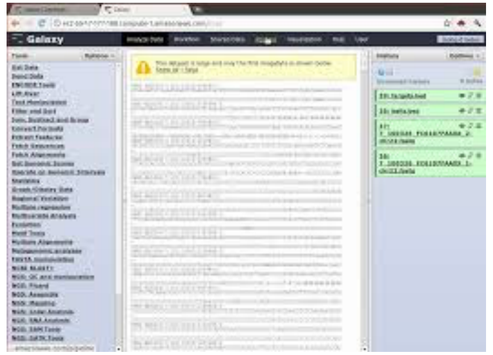
72Tb

72Tb

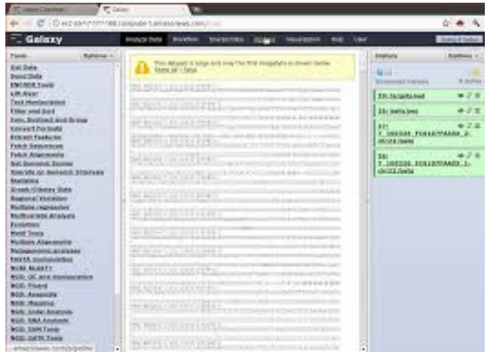
Software



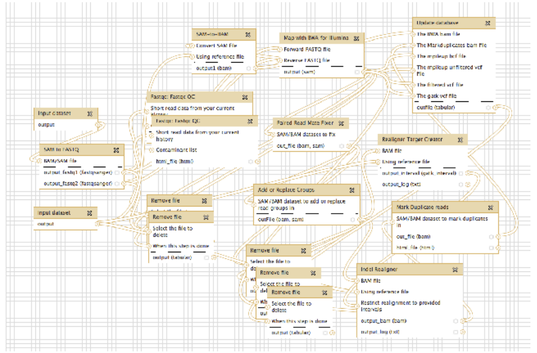
Whole Genome

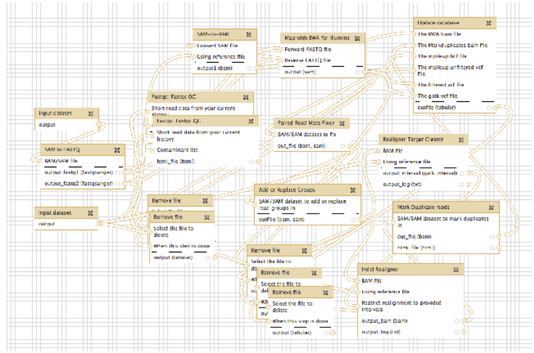


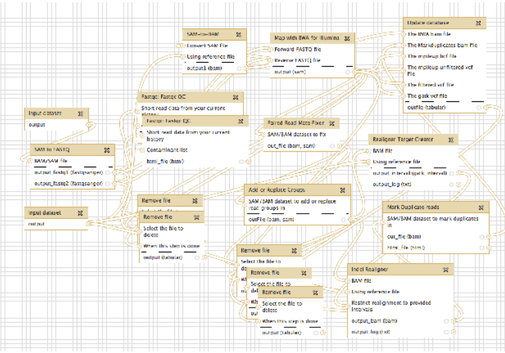
Transcriptome



BiSulfite

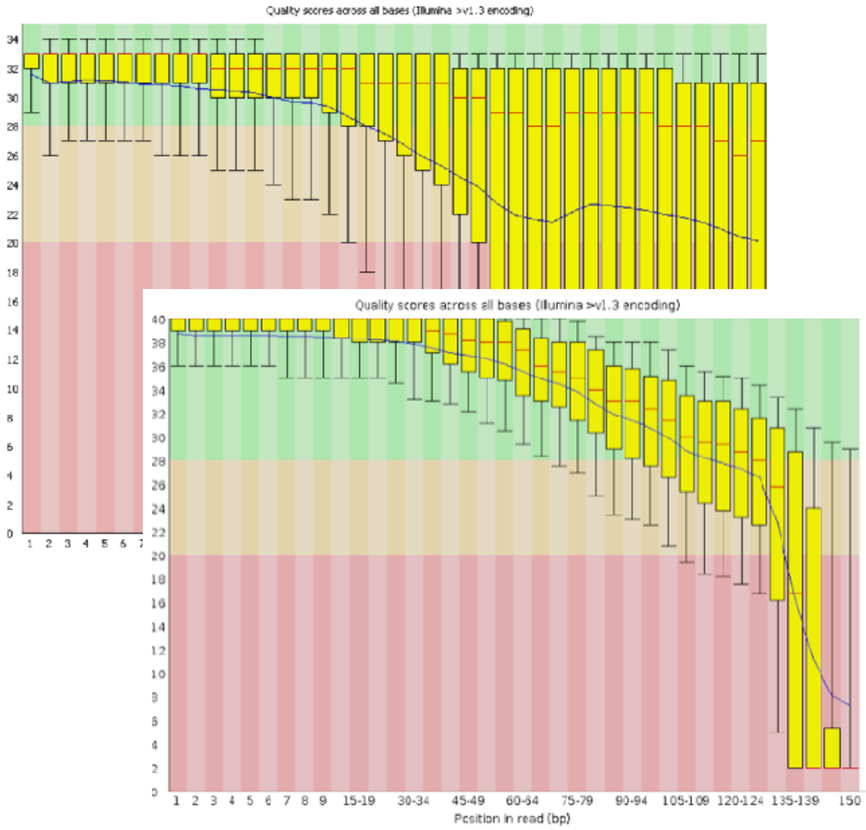




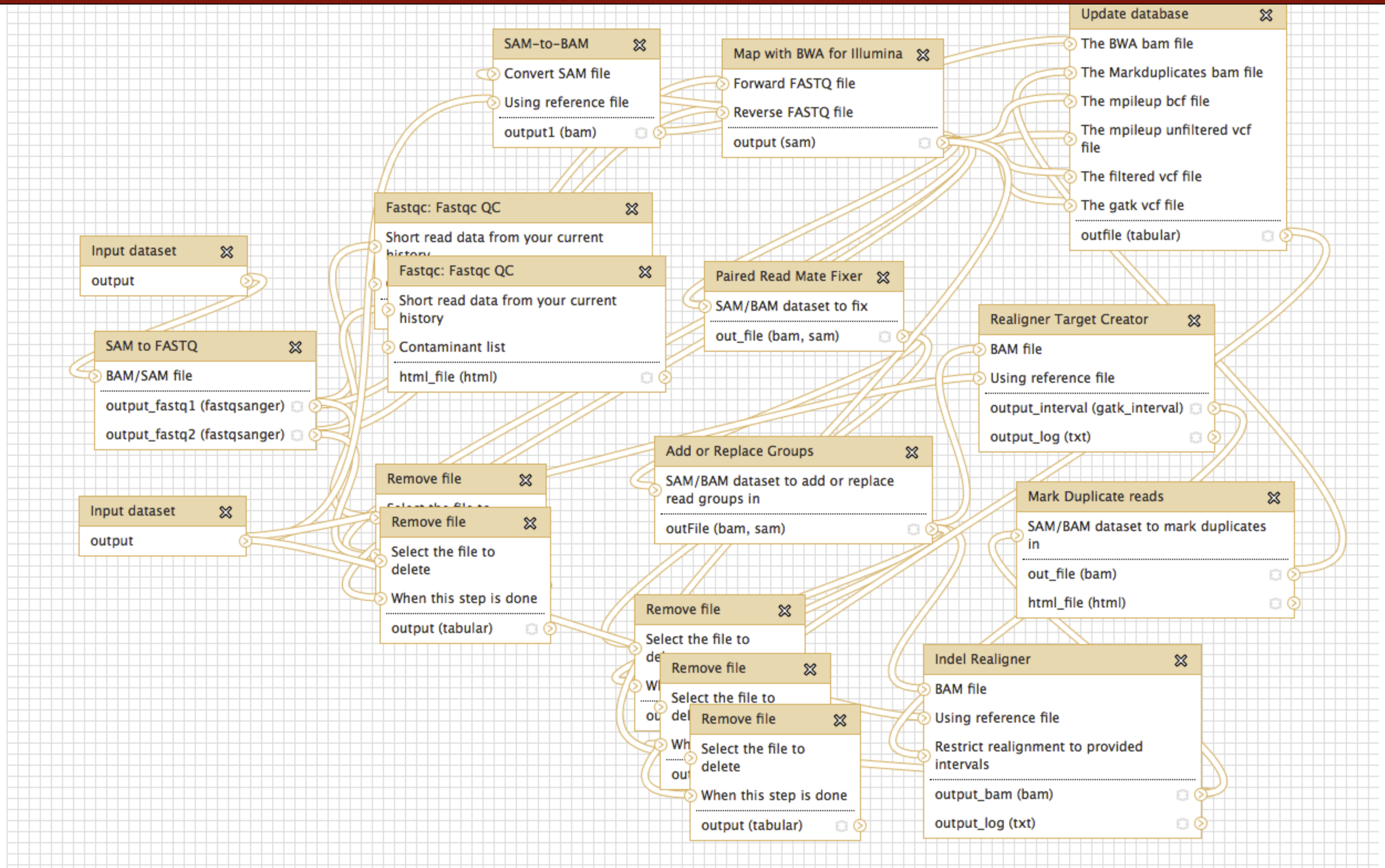


Research team setup

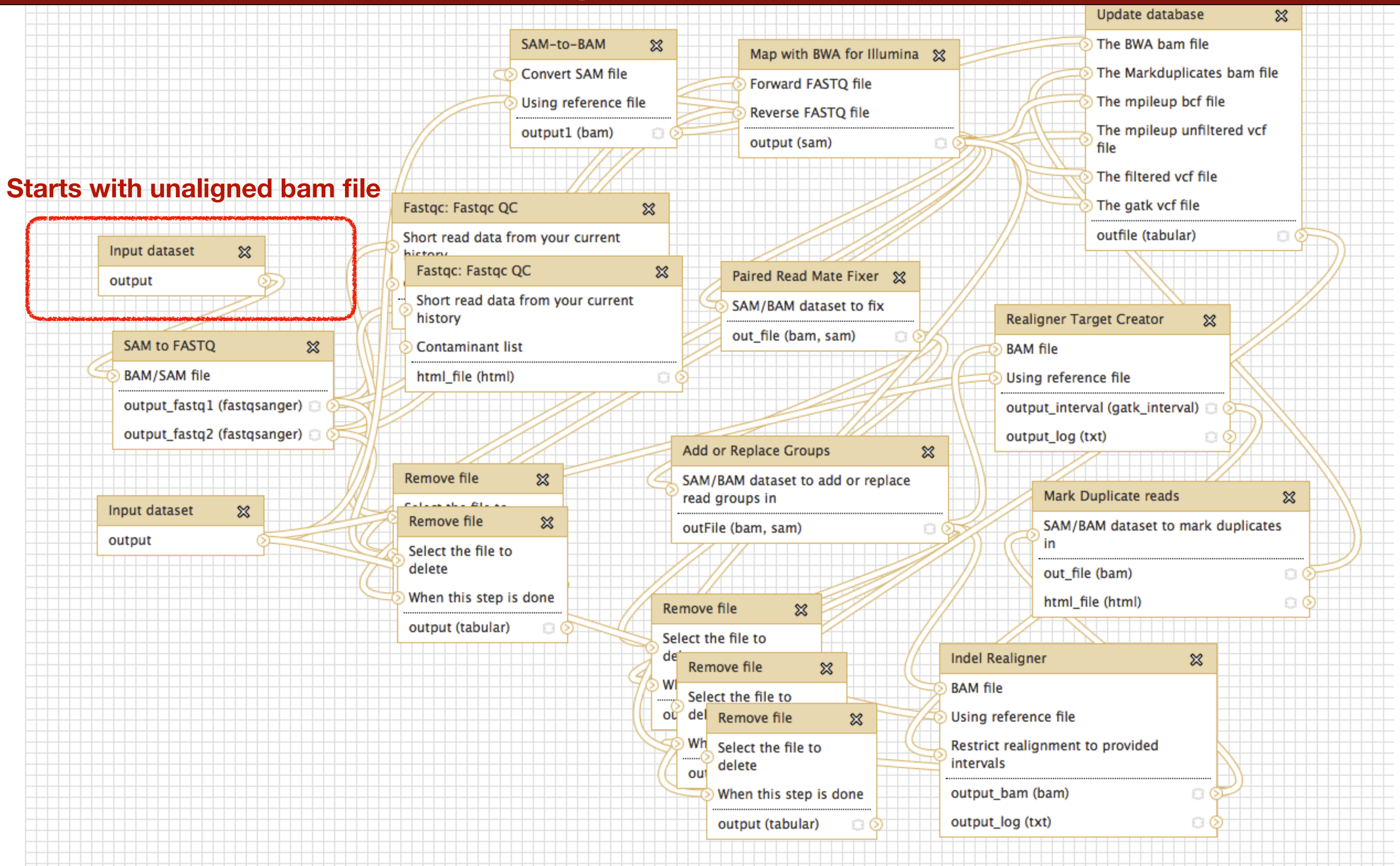
- Small bioinformatics team (1)
- Quick verification of input data
  - Quality control -> within 14 days
  - Data only available for 30 days on FTP server
- Research team setup
  - Very specific samples needed to be processed fast and distributed to postdocs
  - Verification of samples -> IBD from SNPs
  - Duplicate samples



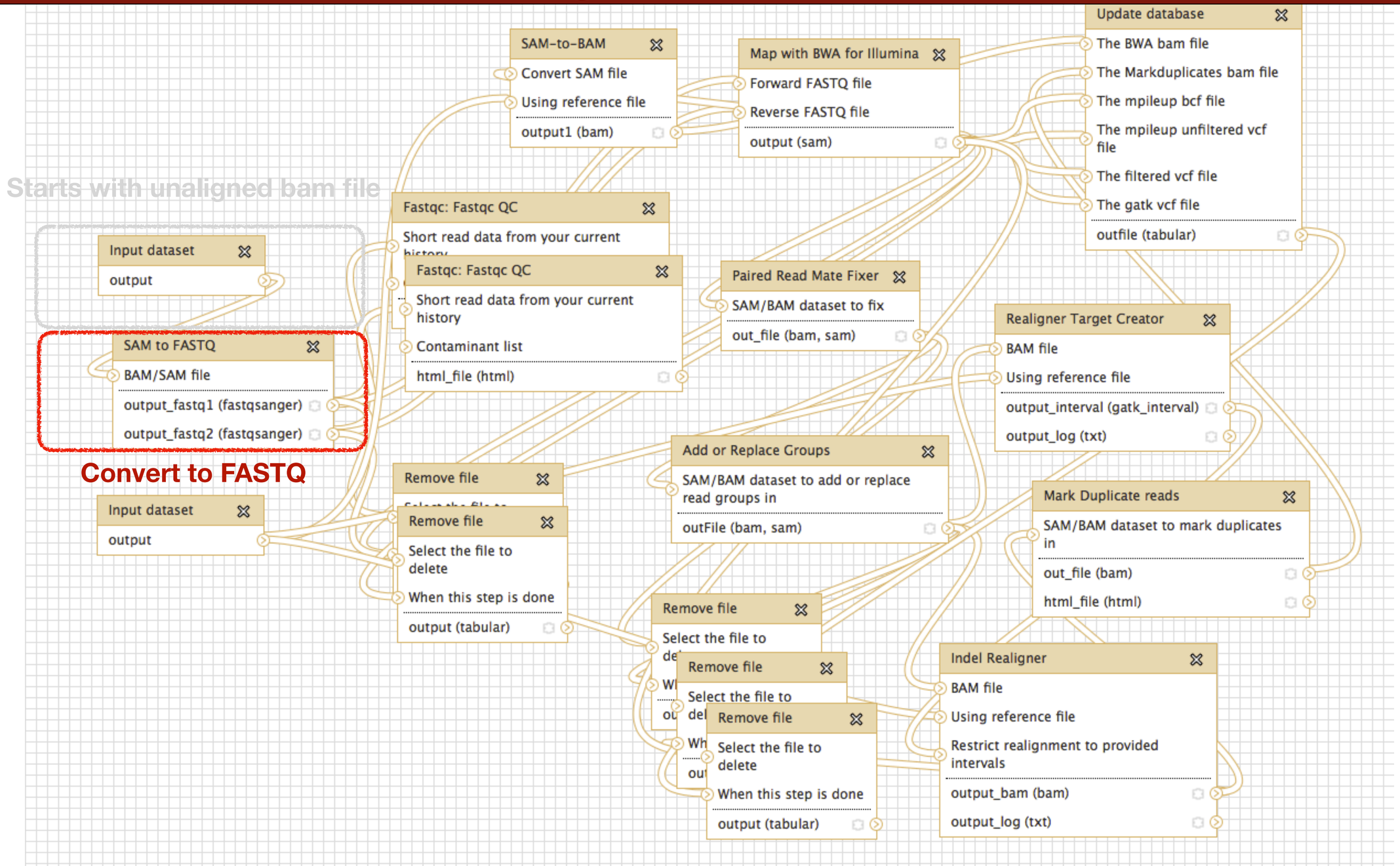
Genome Resequencing pipeline



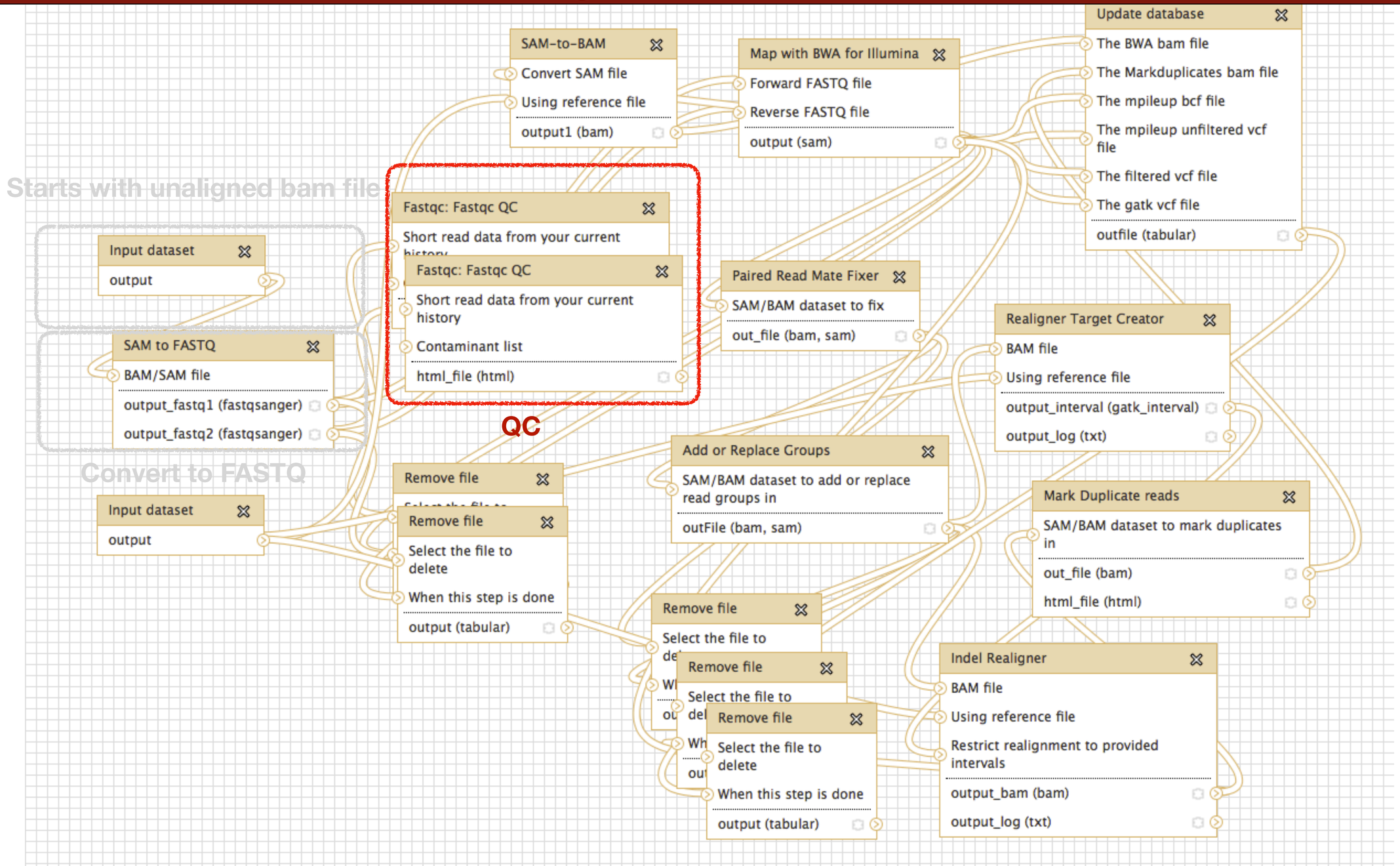
Genome Resequencing pipeline



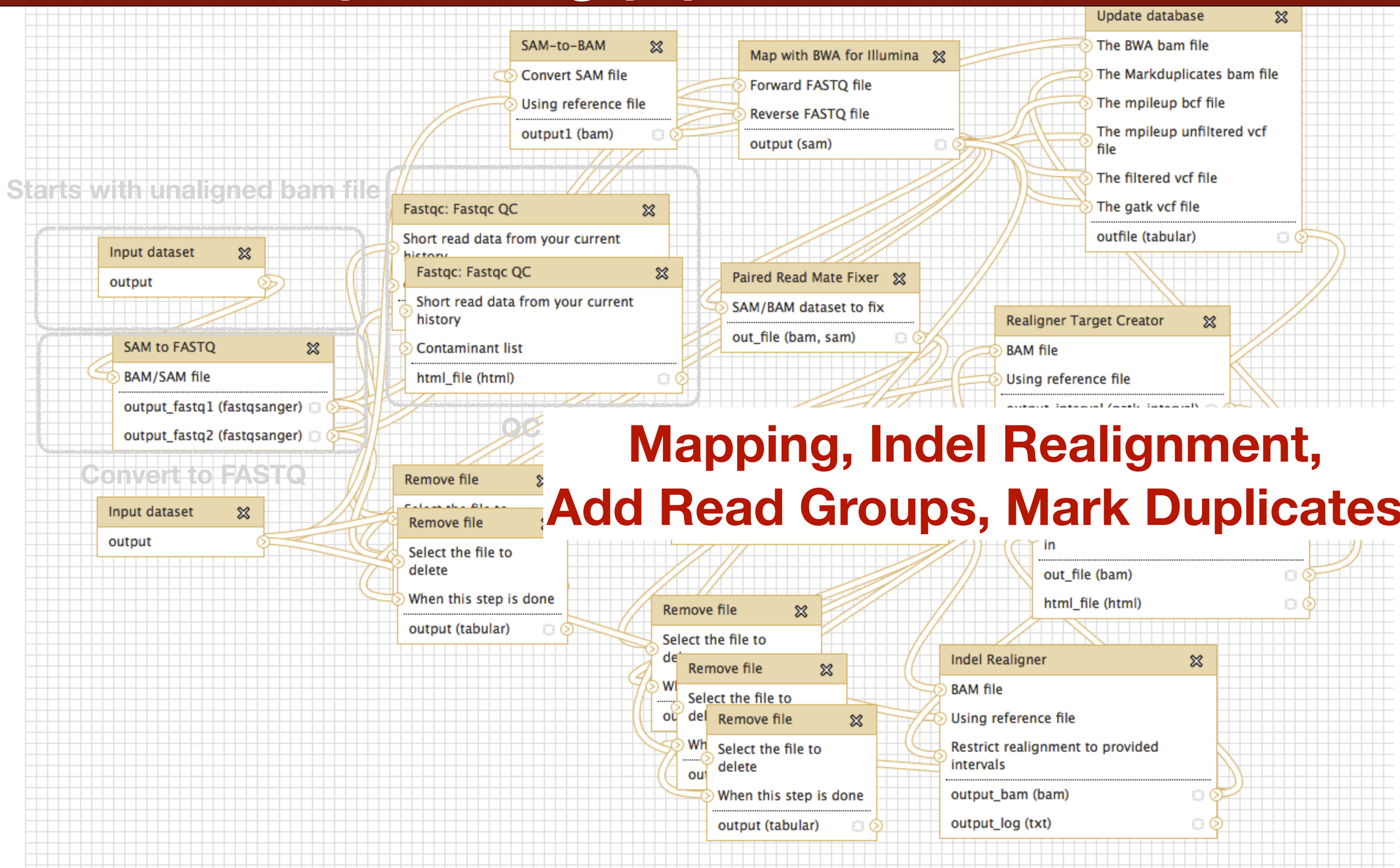
Genome Resequencing pipeline



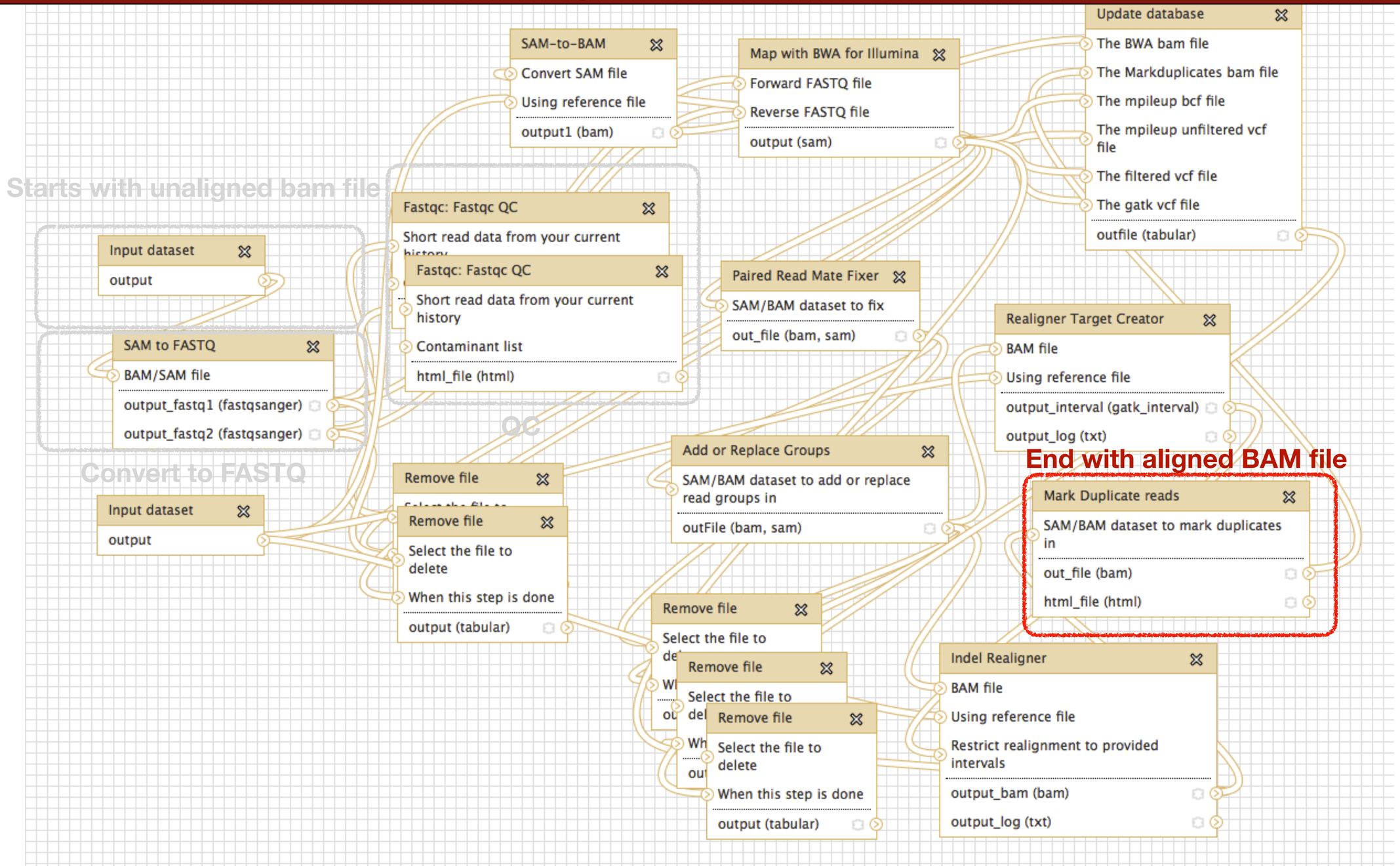
Genome Resequencing pipeline



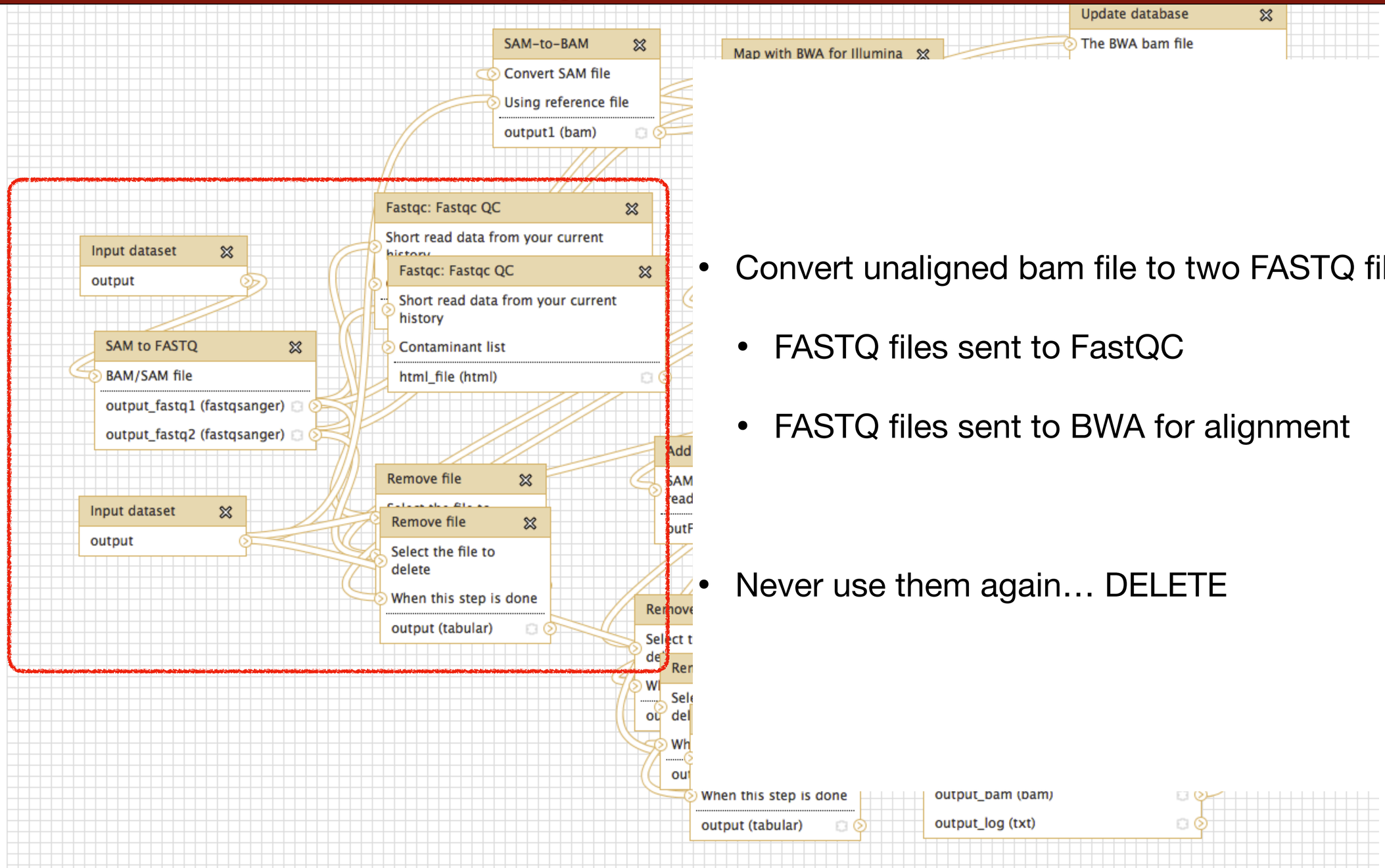
Genome Resequencing pipeline



Genome Resequencing pipeline

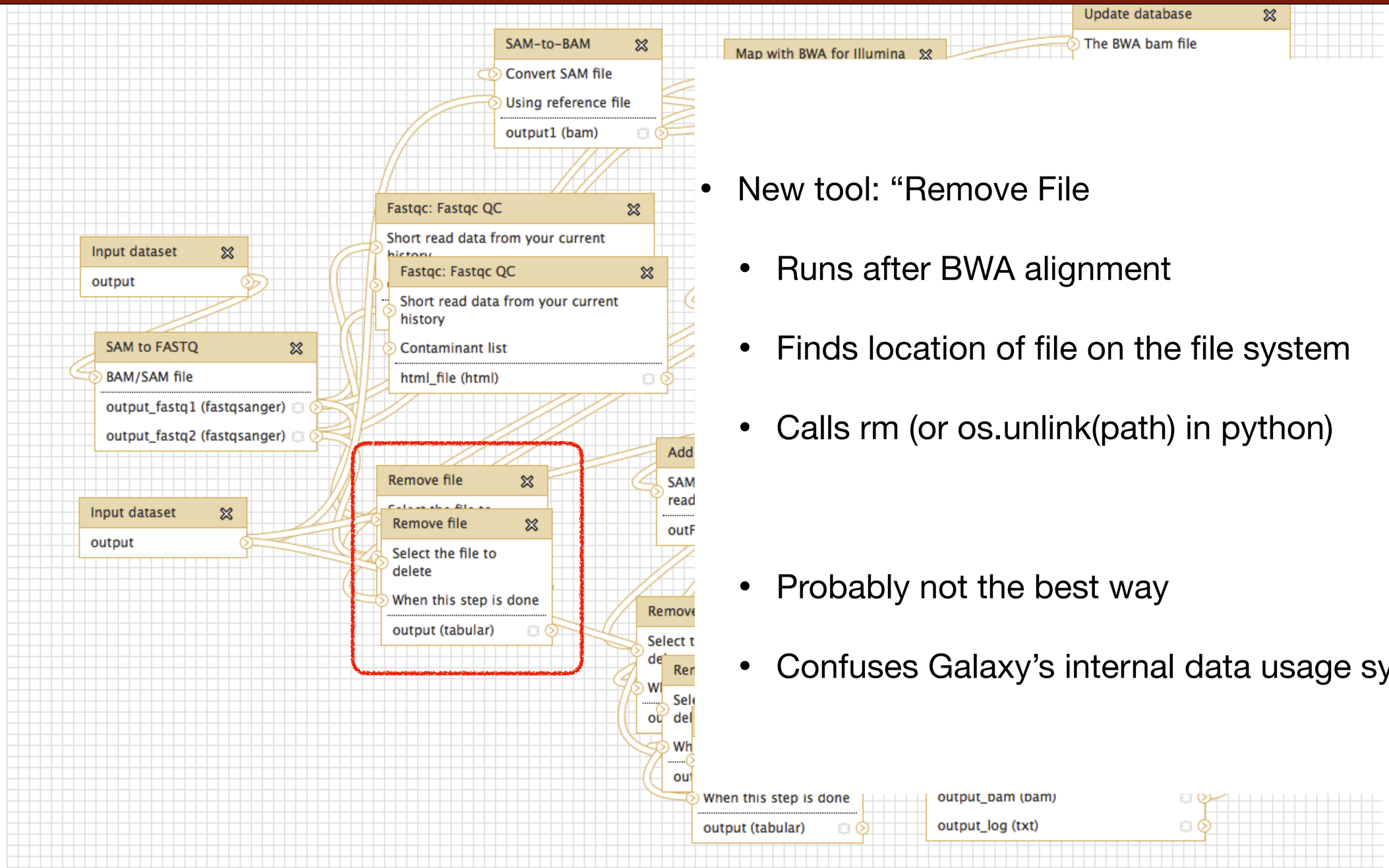


Genome Resequencing pipeline



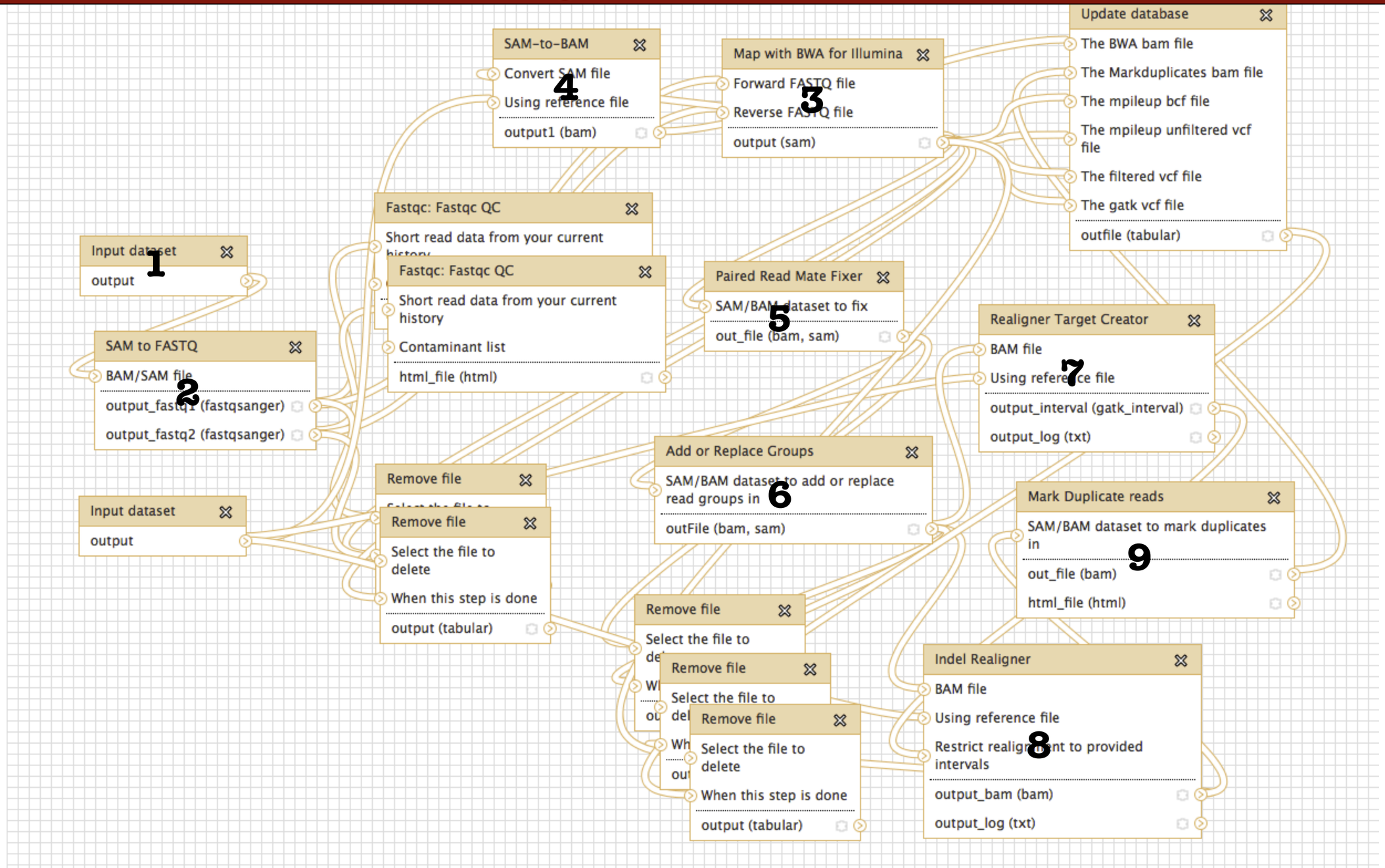
- Convert unaligned bam file to two FASTQ files
  - FASTQ files sent to FastQC
  - FASTQ files sent to BWA for alignment
- Never use them again… DELETE

Genome Resequencing pipeline

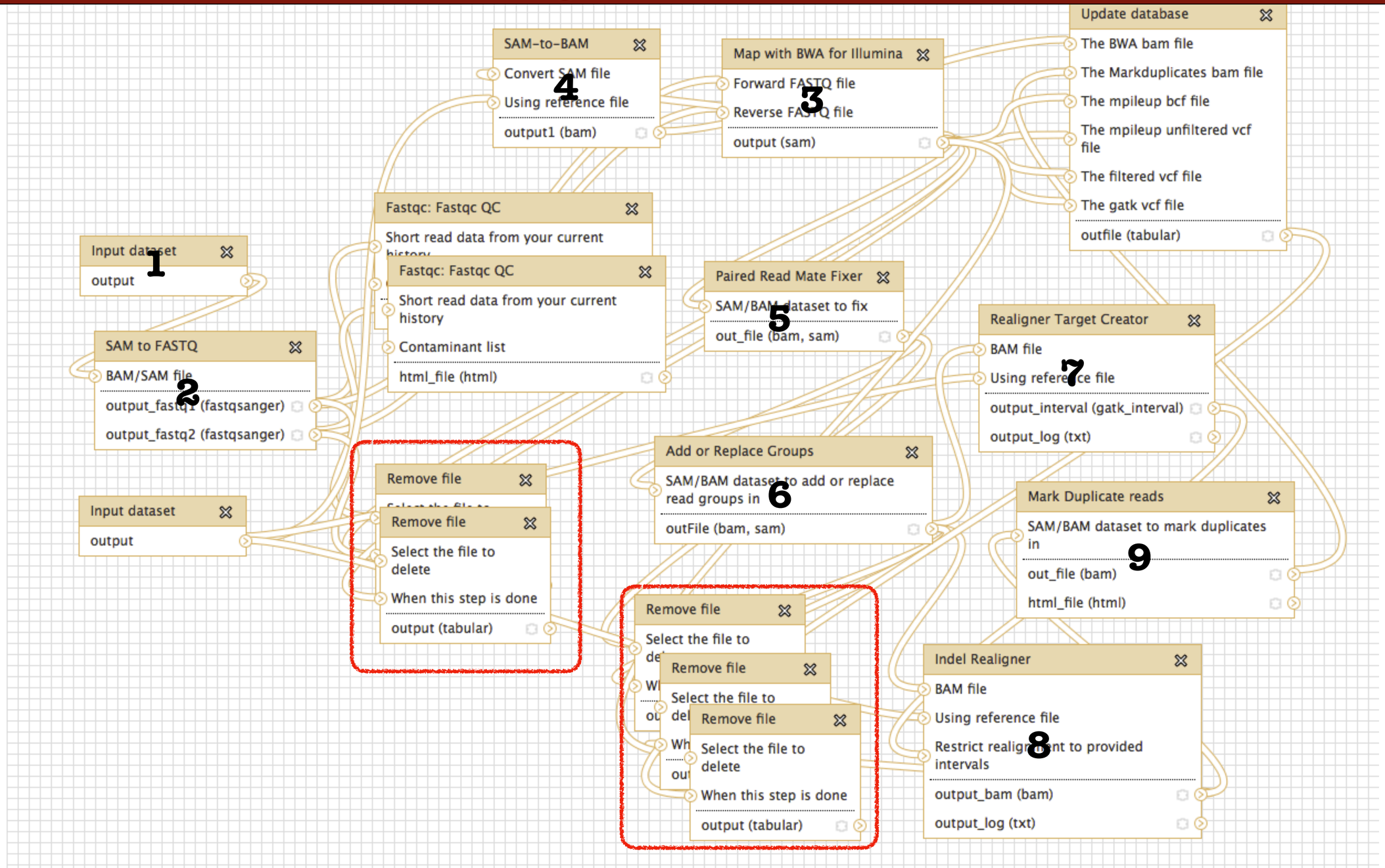


- New tool: “Remove File
  - Runs after BWA alignment
  - Finds location of file on the file system
  - Calls rm (or os.unlink(path) in python)
  - Probably not the best way
  - Confuses Galaxy’s internal data usage system

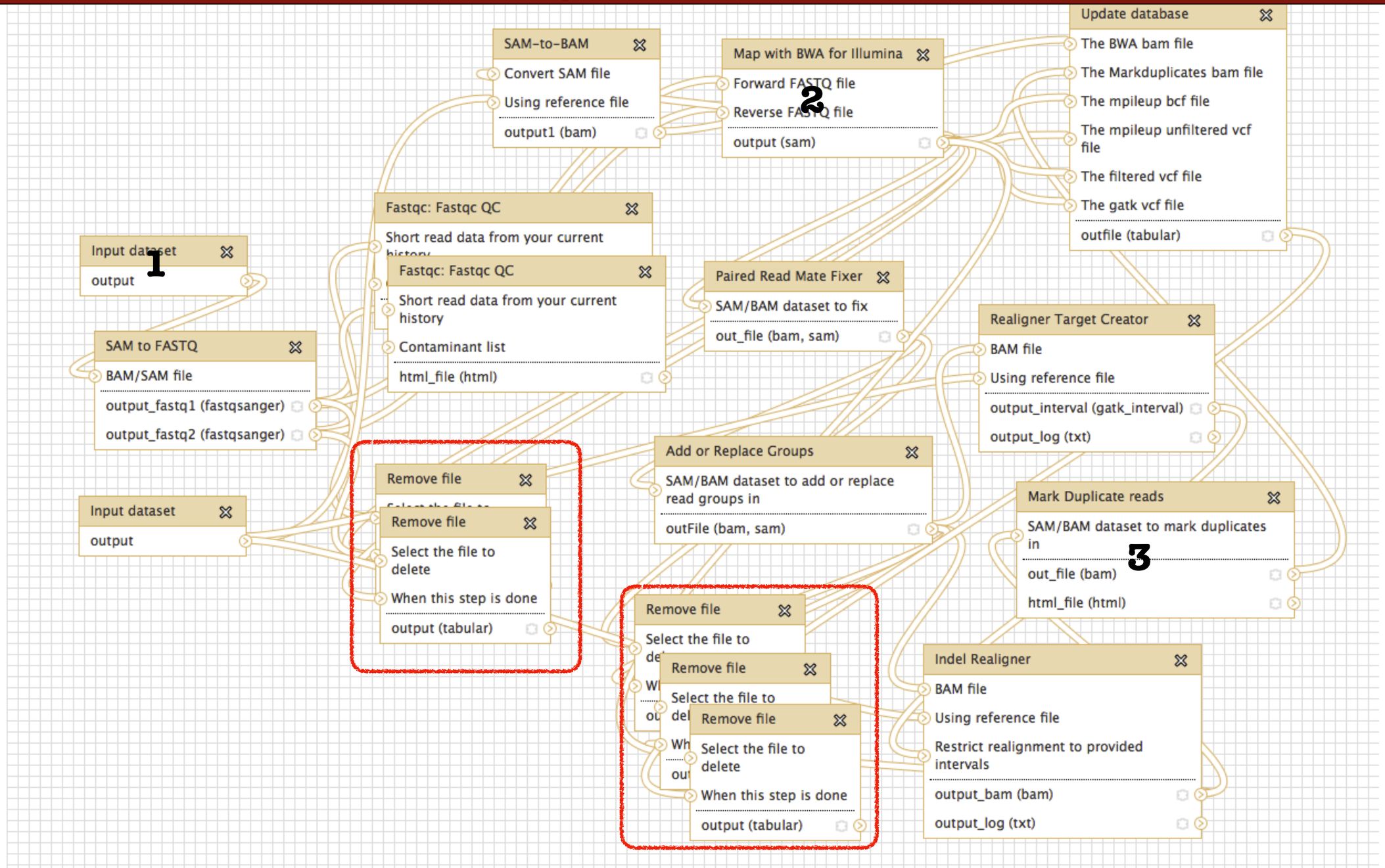
Genome Resequencing pipeline



Genome Resequencing pipeline

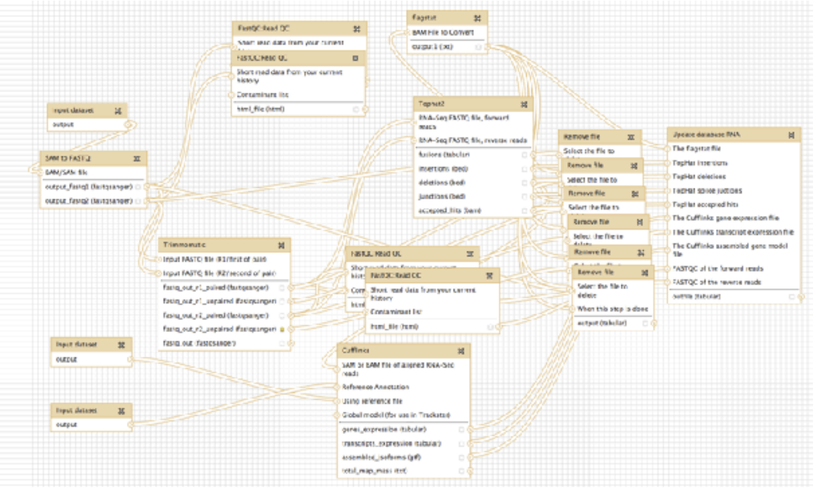


Genome Resequencing pipeline



Transcriptome Sequencing

- 866 transcriptome samples
  - Between 40 and 80 million reads
  - Illumina HiSeq 2000

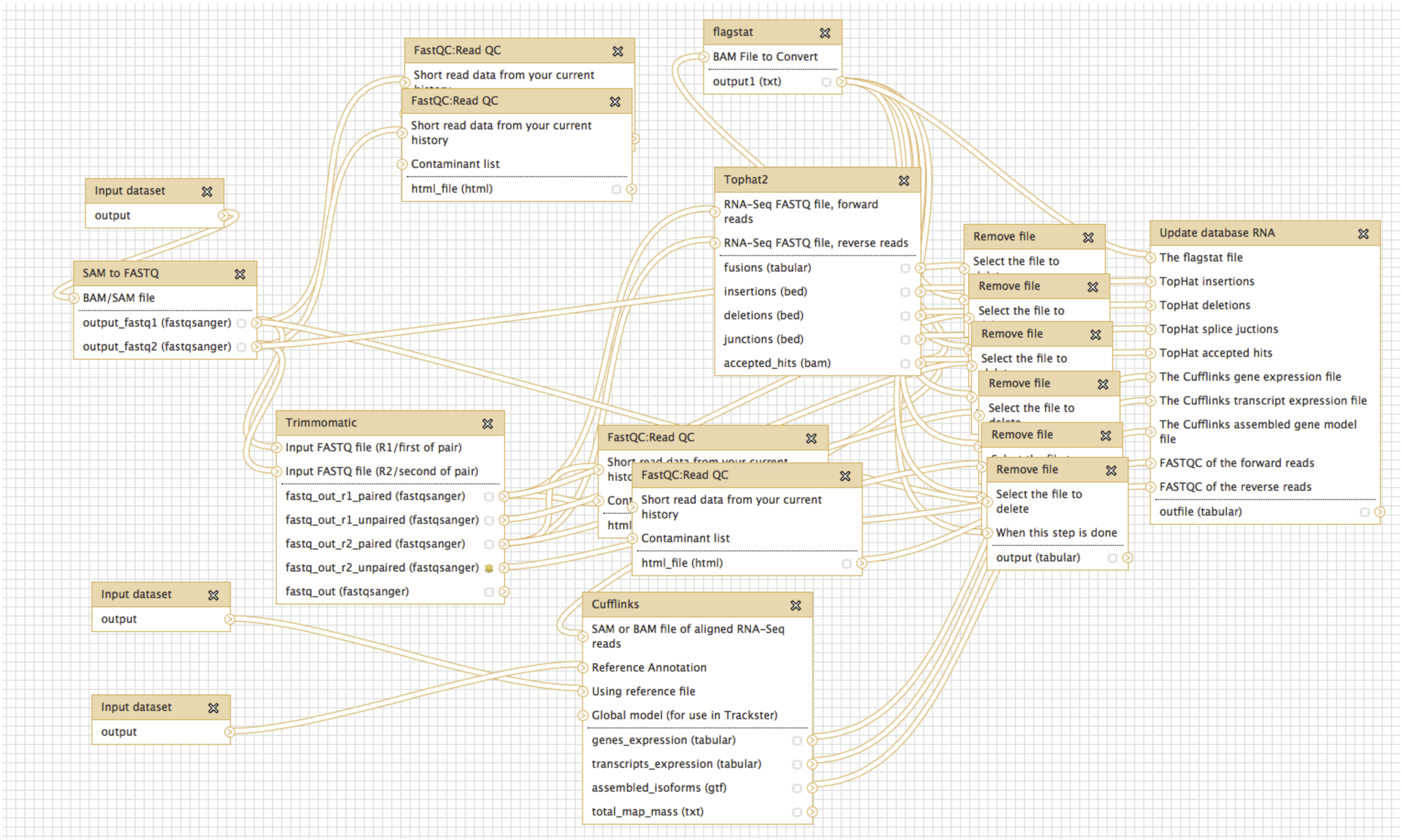


Galaxy / WEST

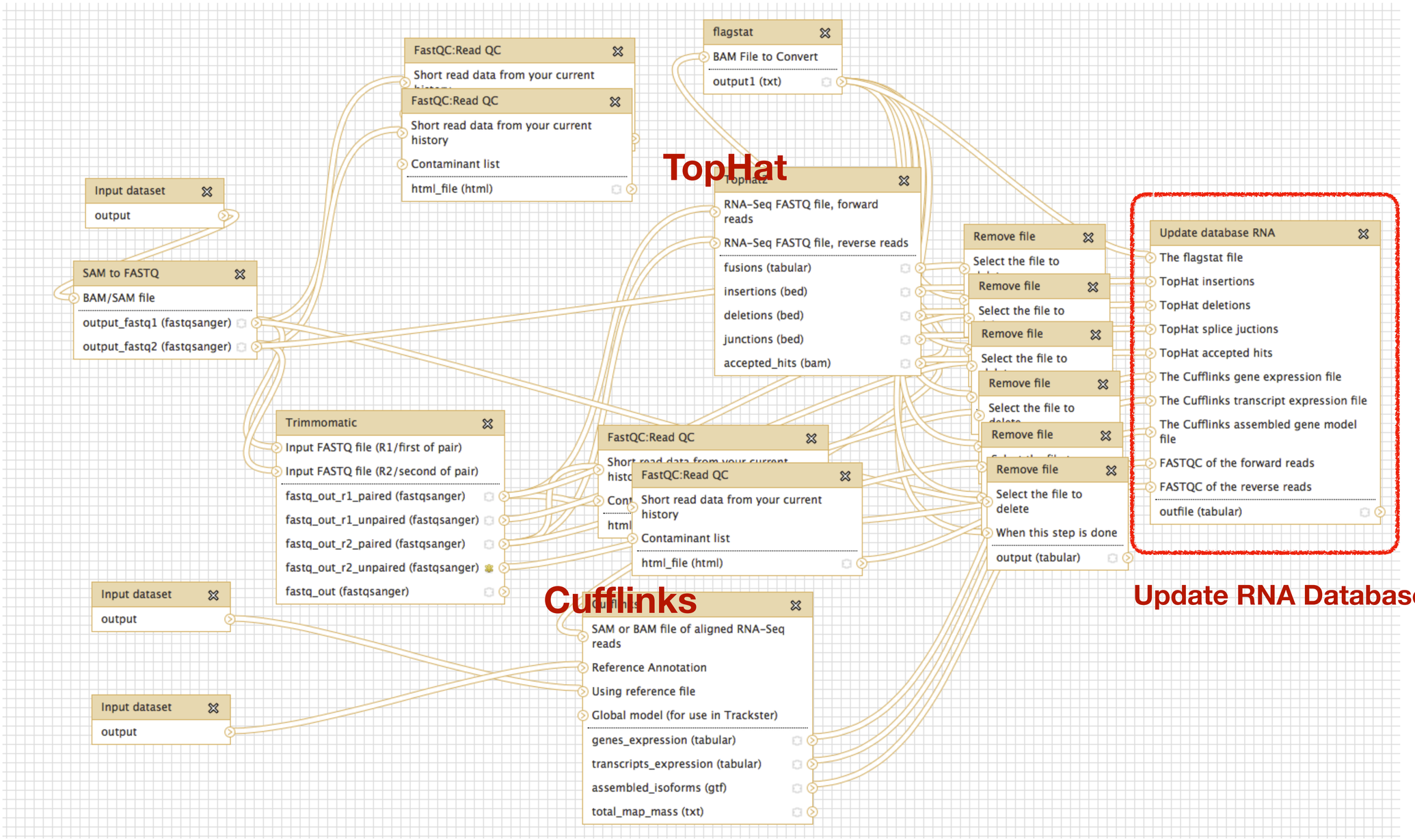
- Single workflow needed to handle all the samples.
  - Gene expression

Genome BritishColumbia
Leading ▸ Investing ▸ Connecting

Transcriptome Sequencing



Transcriptome Sequencing

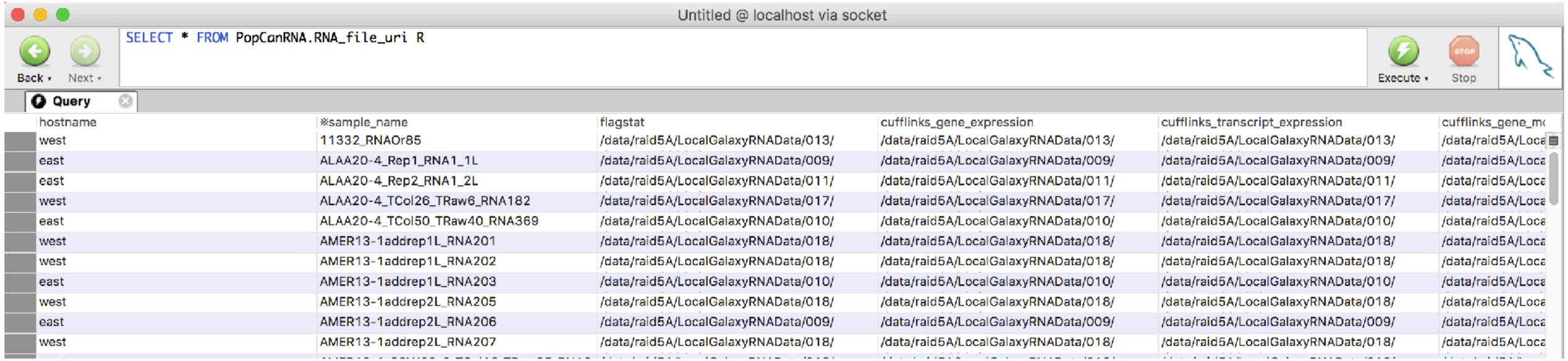


Transcriptome Sequencing

```
CREATE TABLE  `PopCanRNA`.`RNA_file_uri` (
`hostname` text,
`sample_name` text NOT NULL,
`flagstat` text,
`cufflinks_gene_expression` text,
`cufflinks_transcript_expression` text,
`cufflinks_gene_models` text,
`fastqc_fwd` text,
`fastqc_rev` text,
`tophat_insertions` text,
`tophat_deletions` text,
`tophat_splice_junctions` text,
`tophat_bam` text,
`tophat_bam_md5sum` text,
PRIMARY KEY (`sample_name`(30)) USING BTREE
) ENGINE=InnoDB DEFAULT CHARSET=utf8
```

Transcriptome Sequencing

- Independent MySQL database
- Allows for programmatic querying of results



hostname	※sample_name	flagstat	cufflinks_gene_expression	cufflinks_transcript_expression	cufflinks_gene_m
west	11332_RNAOr85	/data/raid5A/LocalGalaxyRNAData/013/	/data/raid5A/LocalGalaxyRNAData/013/	/data/raid5A/LocalGalaxyRNAData/013/	/data/raid5A/Loca
east	ALAA20-4_Rep1_RNA1_1L	/data/raid5A/LocalGalaxyRNAData/009/	/data/raid5A/LocalGalaxyRNAData/009/	/data/raid5A/LocalGalaxyRNAData/009/	/data/raid5A/Loca
east	ALAA20-4_Rep2_RNA1_2L	/data/raid5A/LocalGalaxyRNAData/011/	/data/raid5A/LocalGalaxyRNAData/011/	/data/raid5A/LocalGalaxyRNAData/011/	/data/raid5A/Loca
west	ALAA20-4_TCol26_TRaw6_RNA182	/data/raid5A/LocalGalaxyRNAData/017/	/data/raid5A/LocalGalaxyRNAData/017/	/data/raid5A/LocalGalaxyRNAData/017/	/data/raid5A/Loca
east	ALAA20-4_TCol50_TRaw40_RNA369	/data/raid5A/LocalGalaxyRNAData/010/	/data/raid5A/LocalGalaxyRNAData/010/	/data/raid5A/LocalGalaxyRNAData/010/	/data/raid5A/Loca
west	AMER13-1addrep1L_RNA201	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/Loca
west	AMER13-1addrep1L_RNA202	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/Loca
east	AMER13-1addrep1L_RNA203	/data/raid5A/LocalGalaxyRNAData/010/	/data/raid5A/LocalGalaxyRNAData/010/	/data/raid5A/LocalGalaxyRNAData/010/	/data/raid5A/Loca
west	AMER13-1addrep2L_RNA205	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/Loca
east	AMER13-1addrep2L_RNA206	/data/raid5A/LocalGalaxyRNAData/009/	/data/raid5A/LocalGalaxyRNAData/009/	/data/raid5A/LocalGalaxyRNAData/009/	/data/raid5A/Loca
west	AMER13-1addrep2L_RNA207	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/LocalGalaxyRNAData/018/	/data/raid5A/Loca

Transcriptome Sequencing

- Dynamic Building of Gene Expression Results table

Untitled @ localhost via socket

```
SELECT * FROM PopCanRNA.popcan_leaf_gene_expression p
```

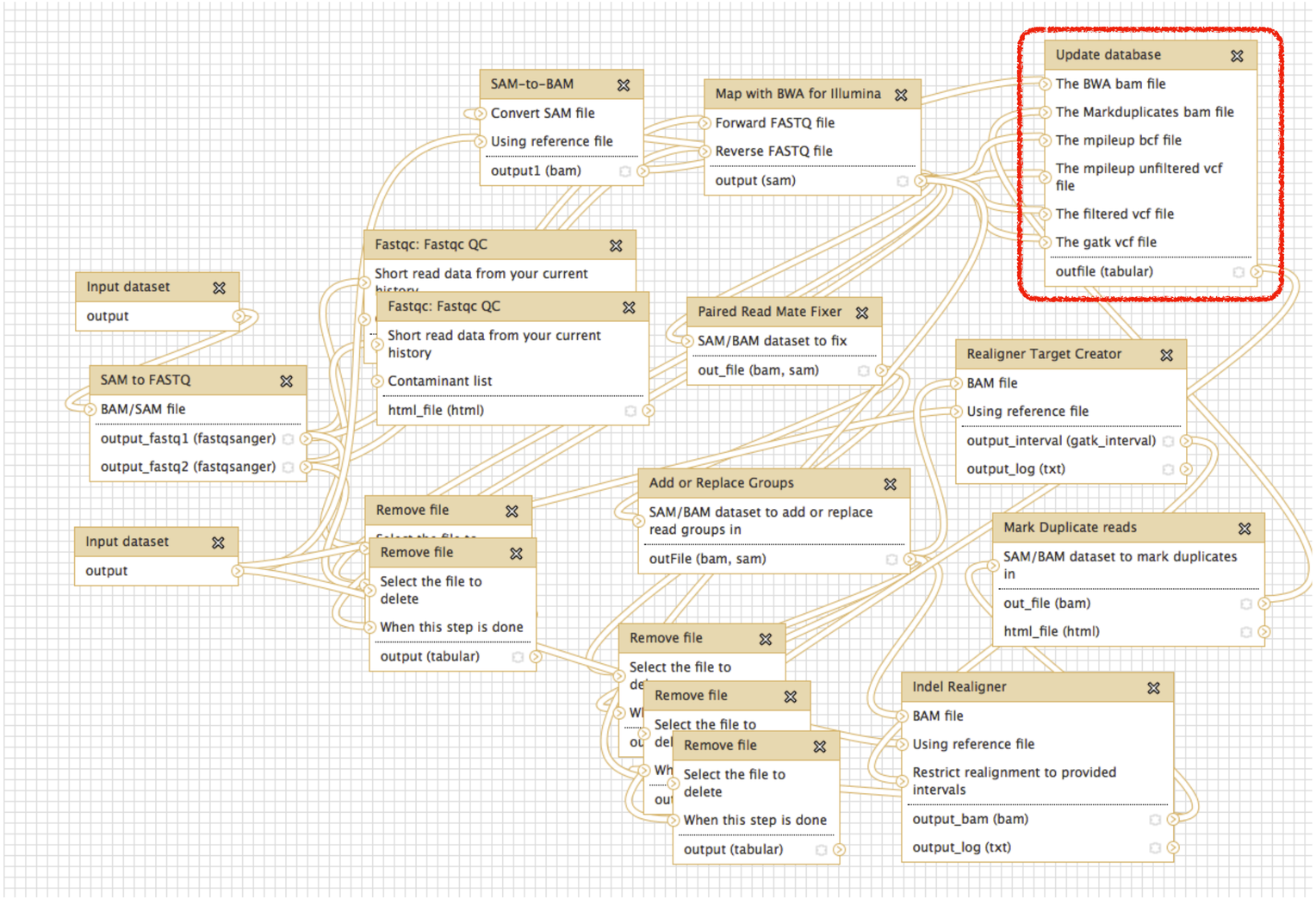
Back ▾ Next ▾ Execute ▾ Stop

Query

gene_number	gene_id	tissue	ALAA20-3_AMER13-1addrep1L_RNA201	ALAA20-3_AMER13-1addrep1L_RNA202	ALAA20-3_AMER13-1addrep1L_RNA203	ALAA20-3_AMER13-1
4	Potri.001G000400	first uncurled leaf	41.0282	29.491	28.6132	32.8809
5	Potri.001G000500	first uncurled leaf	0	0	0	0
6	Potri.001G000600	first uncurled leaf	0.0251797	0	0	0
7	Potri.001G000700	first uncurled leaf	33.0592	22.9251	26.793	25.8353
8	Potri.001G000800	first uncurled leaf	0.325482	0.220098	0.155444	0.175966
9	Potri.001G000900	first uncurled leaf	86.5151	84.2777	88.5511	82.5777
10	Potri.001G001000	first uncurled leaf	6.20827	1.92059	4.45097	2.79066
11	Potri.001G001100	first uncurled leaf	1.30059	1.72912	1.65504	1.70034
12	Potri.001G001200	first uncurled leaf	0.130111	0.0590433	0.144995	0.0438737
13	Potri.001G001300	first uncurled leaf	40.5672	24.2564	30.4796	30.8496
14	Potri.001G001400	first uncurled leaf	2.67977	2.76517	3.07866	2.72547
15	Potri.001G001500	first uncurled leaf	9.02086	10.1361	9.35012	9.2488
16	Potri.001G001600	first uncurled leaf	81.5252	61.6698	82.7354	73.2591
17	Potri.001G001700	first uncurled leaf	7.5533	5.45722	6.34066	7.18672
18	Potri.001G001800	first uncurled leaf	3.93523	1.58916	2.33021	2.78344

Combining Data

Genome Pipeline



- Genome BAM files
  - GATK pipeline
  - Different species
  - Selected individuals
- Fast, up to date sharing of the results

Results

The Plant Cell, Vol. 29: 2000–2015, August 2017, www.plantcell.org © 2017 ASPB.

ttp://orcid.org/0000-0001-7527-5461

Check for updates

Exploiting Natural Variation to Uncover an Alkene Biosynthetic Enzyme in Poplar OPEN

Eliana Gonzales-Vigil,[a] Charles A. Hefer,[b] Michelle E. von Loessl,[a] Jonathan La Mantia,[c] and Shawn D. Mansfie

[a] Depar
[b] Biotec
[c] U.S. D

ORCID

-554X

New Phytologist

Research

Comparative interrogation of the developing xylem transcriptomes of two wood-forming species: *Populus trichocarpa* and *Eucalyptus grandis*

Charles A. Hefer[1], Eshchar Mizrachi[2], Alexander A. Myburg[2], Carl J. Douglas[1] and Shawn D. Mansfield[3]

[1]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [2]Department of Genetics, Forestry and Agri

Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria 0028, South Africa; [3]Department of Wood Science, Faculty of

Sciences Centre, 4030-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada

MOLECULAR ECOLOGY

Molecular Ecology (2015) 24, 3243–3256 doi: 10.1111/mec.13126

FROM THE COVER

Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*)

A. GERALDES,*[1] C. A. HEFER,*[1] A. CAPRON,* N. KOLOSOVA,* F. MARTINEZ-NUÑEZ,* R. Y. SOOLANAYAKANAHALLY,† B. STANTON,‡ R. D. GUY,§ S. D. MANSFIELD,¶ C. J. DOUGLAS* and Q. C. B. CRONK*

*Department of Botany, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada, †Agroforestry Development Centre, Agriculture and Agri-Food Canada, Indian Head, SK S0G 2K0, Canada, ‡Greenwood Resources, Portland, OR 97201, USA, §Department of Forest and Conservation Sciences, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada, ¶Department of Wood Science, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

New Phytologist

Introgression from *Populus balsamifera* underlies adaptively significant variation and range boundaries in *P. trichocarpa*

Adriana Suarez-Gonzalez[1], Charles A. Hefer[1,2], Christian Lexer[3], Carl J. Douglas[1†] and Quentin C. B. Cronk[1]

[1]Department of Botany, University of British Columbia, Vancouver, BC Canada, V6T 1Z4; [2]Biotechnology Platform, Agricultural Research Council, Private Bag X05, Onderstepoort 0110,

tany and Biodiversity Research, Univers

Altering carbon allocation in hybrid poplar (*Populus alba* × *grandidentata*) impacts cell wall growth and development

Faride Unda[1], Hoon Kim[2,3], Charles Hefer[4], John Ralph[2,3], Shawn D. Mansfield[1,3]

GENOMICS OF HYBRIDIZATION

Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood)

ADRIANA SUAREZ-GONZALEZ,* CHARLES A. HEFER,*[1] CAMILLE CHRISTE,† OLIVER COREA,‡ CHRISTIAN LEXER,†§ QUENTIN C. B. CRONK* and CARL J. DOUGLAS*

AGIP/POPCAN

AGIP/POPCAN team

Carl Douglas
Shawn Mansfield
Quentin Cronk
Jürgen Ehlting
Yousry El-Kassaby
Rob Guy
Malcolm Campbell
Peter Constabel
Richard Hamelin
Marty Luckert
Tom Maness

Shofiul Azam
Hua Bao
Michael Friedmann
Miki Fujita
Armando Geraldes
Jan Hannemann
Catalin Ritea
Charles Hefer
Peter Kalynyak
Jaroslav Klapste
Jon LaMantia
Eryang Li
Athena McKown
Ilga Porth
Alex Skyba

Natalia Kolosova
Nima Farzaneh
Shucai Wang
Daisie Huang
Steffi Fritsche
Arnaud Capron
Mike Barker
Matt King
Heather Ramsay
Chris Grassa

Genome Sciences Centre

Inanc Birol
Reza Falsafi
Steve Jones
Marco Marra
Johnson Pang
Nina Thiessen
Yongjun Zhao

Collaborators

ORNL/BESC – Jerry Tuskan, Steve DiFazio, Gancho Slavov, Lee Gunter et al.
UPSC/POPLARENERGY – Rishi Bhalerao, Stefan Jansson
USFS Forest Products Lab - Dan Cullen
BC Ministry of Forests and Range - Alvin Yanchuk
Greenwood Resources - Brian Stanton
Kruger Inc. – Dan Carson
University of Pretoria – Zander Myburg

GenomeBritishColumbia

Applied Genomics Innovation Program

GenomeCanada

Large Scale Applied Research Program