




Updates on bioinformatics development

including Sarek pipeline

 Maxime U. Garcia

 maxulysse.github.io

 @MaxUlysse

 @gau

2019-02-08



Karolinska
Institutet



Barntumörbanken

SciLifeLab

The logo for the National Cancer Institute, featuring a stylized blue and white staircase design.
NATIONAL CANCER INSTITUTE
ATCAGENOMICSGT
INFRASTRUCTURE

NBDS

- Bioinformatics at BTB

- Bioinformatics at BTB
- Sarek pipeline

- Bioinformatics at BTB
- Sarek pipeline
- rnafusion pipeline

- Bioinformatics at BTB
- Sarek pipeline
- rnafusion pipeline
- munin server

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools
- Improve/automatize methods/tools

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools
- Improve/automatize methods/tools
- Use methods/tools

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools
- Improve/automatize methods/tools
- Use methods/tools
- Develop new methods/tools

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools
- Improve/automatize methods/tools
- Use methods/tools
- Develop new methods/tools
- Maintain current methods/tools

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools
- Improve/automatize methods/tools
- Use methods/tools
- Develop new methods/tools
- Maintain current methods/tools
- Methods/tools that are reproducible

Why do we need bioinformaticians?



Barntumörbanken

- Keep up to date current methods/tools
- Improve/automatize methods/tools
- Use methods/tools
- Develop new methods/tools
- Maintain current methods/tools
- Methods/tools that are reproducible
- Provide solutions

What is Sarek?



 <http://sarek.scilifelab.se/>

- Analysis germline and somatic workflow

What is Sarek?



 <http://sarek.scilifelab.se/>

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing

What is Sarek?



 <http://sarek.scilifelab.se/>

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing
- Developed with NGI and NBIS



What is Sarek?



 <http://sarek.scilifelab.se/>

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing
- Developed with NCI and NBIS
- Support from The Swedish Childhood Tumor Biobank



nextflow

 <https://www.nextflow.io/>



 <https://www.sylabs.io/singularity/>

nextflow

 <https://www.nextflow.io/>

Data-driven workflow language



 <https://www.sylabs.io/singularity/>

HPC specific container engine

BIOCONDA[®]

 <https://bioconda.github.io/>

- Virtual environment management system

Sarek exists in multiple flavors



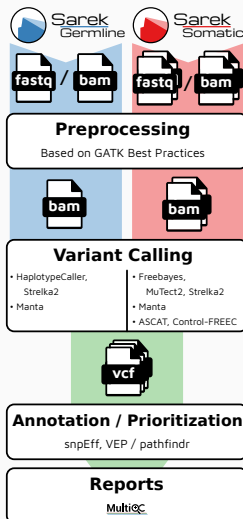
Sarek exists in multiple flavors



Sarek exists in multiple flavors



Data and files workflow




AWS iGenomes

 <https://ewels.github.io/AWS-iGenomes/>

- Human `GRCh37` from the GATK Resource Bundle
- Human `GRCh38` from the GATK Resource Bundle

AWS iGenomes

 <https://ewels.github.io/AWS-iGenomes/>

- Human GRCh37 from the GATK Resource Bundle
- Human GRCh38 from the GATK Resource Bundle
- Dog CanFam3.1 
- Mouse GRCm38 



 <https://software.broadinstitute.org/gatk/best-practices/>

Based on GATK Best Practices (GATK 4.0)



🌐 <https://software.broadinstitute.org/gatk/best-practices/>


Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`



 <https://software.broadinstitute.org/gatk/best-practices/>

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`
 - FASTQs or BAMs 



🌐 <https://software.broadinstitute.org/gatk/best-practices/>

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`
 - FASTQs or BAMs 🔧
- Duplicates marked with `picard MarkDuplicates`









🌐 <https://software.broadinstitute.org/gatk/best-practices/>








Based on GATK Best Practices (GATK 4.0)











- Reads mapped to reference genome with `bwa mem`
 - FASTQs or BAMs 🔧
- Duplicates marked with `picard MarkDuplicates`
- Recalibrate with `GATK BaseRecalibrator`

- SNVs and small indels:


- SNVs and small indels:
 - Freebayes 
 - HaplotypeCaller 
 - MuTect2 
 - Strelka2  

- SNVs and small indels:
 - Freebayes 
 - HaplotypeCaller 
 - MuTect2 
 - Strelka2  
- Structural variants:

- SNVs and small indels:
 - Freebayes 
 - HaplotypeCaller 
 - MuTect2 
 - Strelka2 /
- Structural variants:
 - Manta /
- Sample heterogeneity, ploidy and CNVs:

- SNVs and small indels:
 - Freebayes 
 - HaplotypeCaller 
 - MuTect2 
 - Strelka2 /
- Structural variants:
 - Manta /
- Sample heterogeneity, ploidy and CNVs:
 - ASCAT 
 - Control-FREEC  

- VEP and SnpEff
-  ClinVar, COSMIC, dbSNP, GENCODE, gnomAD, polyphen, sift, etc.

- VEP and SnpEff
-  ClinVar, COSMIC, dbSNP, GENCODE, gnomAD, polyphen, sift, etc.
- Possibility to use cache directories

What we need

- Adapt settings where necessary, and ensure they give good results

Markus Mayrhofer

What we need

- Adapt settings where necessary, and ensure they give good results
- Coherent overview to allow critical assessment of sequence and variant quality

Markus Mayrhofer

What we need

- Adapt settings where necessary, and ensure they give good results
- Coherent overview to allow critical assessment of sequence and variant quality
- Tables of variants with probable relevance for the disease

Markus Mayrhofer

Our solution

- Parse all results into R environment

Our solution

- Parse all results into R environment
- Rank variants based on evidence for being a driver mutation

Our solution

- Parse all results into R environment
- Rank variants based on evidence for being a driver mutation
- Visualize in portable html report for easy browsing

Scoring based on VEP + custom annotations

- Noncoding impact
 - TFBS
 - regulatory
 - conserved sequence

Scoring based on VEP + custom annotations

- Noncoding impact
 - TFBS
 - regulatory
 - conserved sequence
- Coding impact
 - SIFT/Polyphen
 - VEP impact

Scoring based on VEP + custom annotations

- Noncoding impact
 - TFBS
 - regulatory
 - conserved sequence
- Coding impact
 - SIFT/Polyphen
 - VEP impact
- COSMIC
 - Cancer census genes
 - High impact in TSG

Scoring based on VEP + custom annotations

- Noncoding impact
 - TFBS
 - regulatory
 - conserved sequence
- Coding impact
 - SIFT/Polyphen
 - VEP impact
- COSMIC
 - Cancer census genes
 - High impact in TSG
- MSK-impact
 - known hotspot
 - adjacent to known hotspot(s)

Scoring based on VEP + custom annotations

- Noncoding impact
 - TFBS
 - regulatory
 - conserved sequence
- Coding impact
 - SIFT/Polyphen
 - VEP impact
- COSMIC
 - Cancer census genes
 - High impact in TSG
- MSK-impact
 - known hotspot
 - adjacent to known hotspot(s)
- CLINVAR
 - pathogenic or risk

Final results

- Scores are generated for
 - snvs/indels
 - copy number alteration
 - structural alteration

Final results

- Scores are generated for
 - snvs/indels
 - copy number alteration
 - structural alteration
- All for both somatic and germline variants
- Displayed as ranked tables
 - All information from VCF files included

Final results

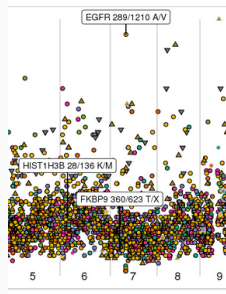
- Scores are generated for
 - snvs/indels
 - copy number alteration
 - structural alteration
- All for both somatic and germline variants
- Displayed as ranked tables
 - All information from VCF files included
- Additional annotation
 - Concomitant LOH marked
 - Swegen, TOPMED allele ratio

Final results

- Scores are generated for
 - snvs/indels
 - copy number alteration
 - structural alteration
- All for both somatic and germline variants
- Displayed as ranked tables
 - All information from VCF files included
- Additional annotation
 - Concomitant LOH marked
 - Swegen, TOPMED allele ratio
- Visualization of all data
 - Genome- and chromosome-wide figures
 - Self-contained HTML report files with all tables and figures

Tables and visualization

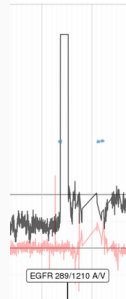
Gene	Mutation	Rank score	Rank Terms	Allele_ratio
EGFR	289/1210 A/V	10	T1_gene moderate_impact clinvar polyphen/SIFT hotspot cosmic_>50	0.94
TP53		9	T1_gene high_impact high+TSG clinvar cosmic_>50	0.86
HIST1H3B	28/136 K/M	7	T1_gene	0.23



SNVs, Indels

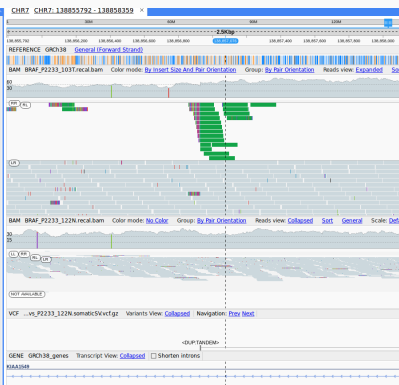
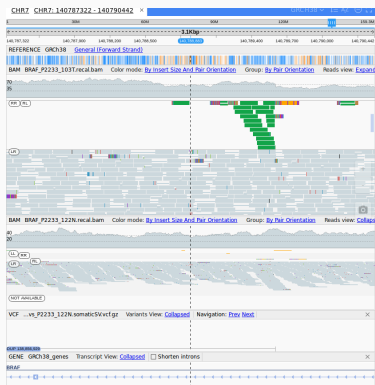
Tables and visualization

Gene	Mutation	Rank score	Rank Terms
EGFR	gain	5	T1_gene focal high_amp
ETNK1	gain	5	T1_gene focal high_amp
KRAS	gain	5	T1_gene focal high_amp

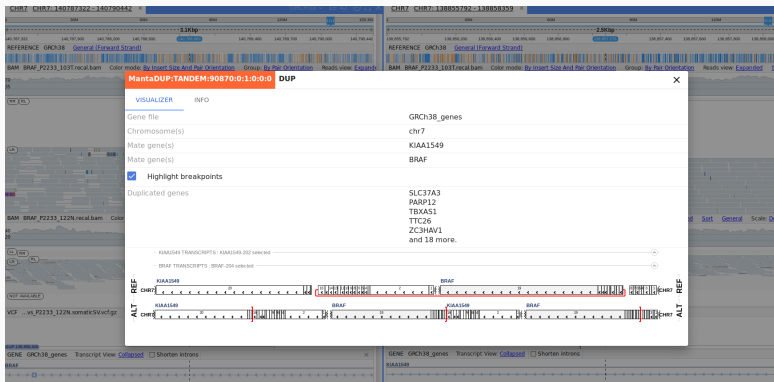


Copy number

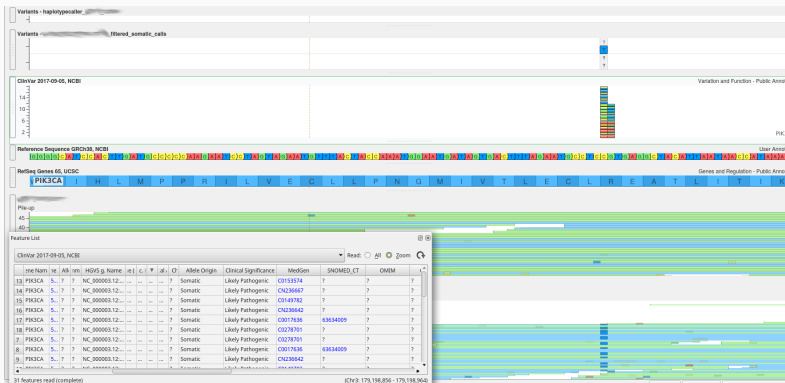
BRAF-KIAA1549



BRAF-KIAA1549



PIK3CA



- 50 tumor/normal pairs with GRCh37 reference

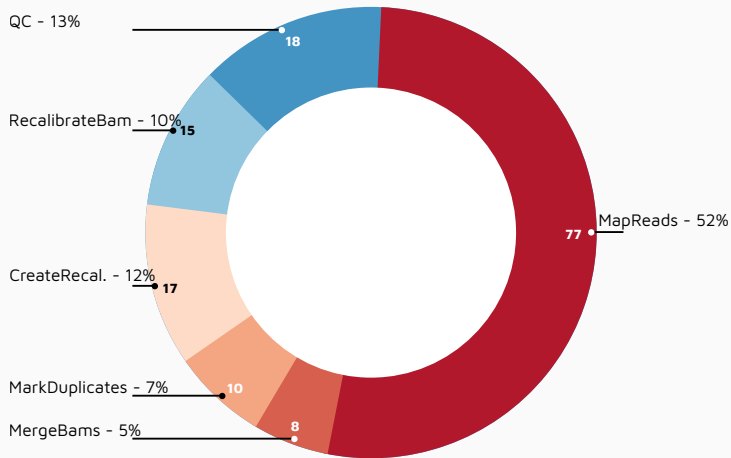
- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference

- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
 - 1 000 samples in germline settings

- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
 - 1 000 samples in germline settings
- Clinical samples with Genomic Medicine Sweden initiative

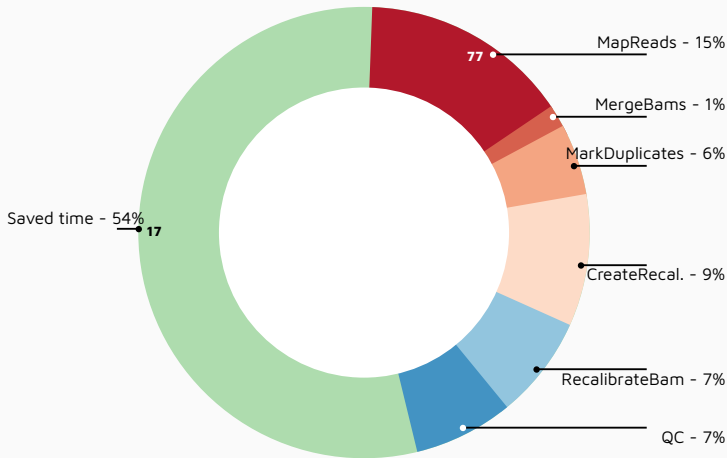
- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
 - 1 000 samples in germline settings
- Clinical samples with Genomic Medicine Sweden initiative
- Used at NGI
 - 200 samples
 - testing it in production
 - plans for validation

Preprocessing time



On a research settings
On UPPMAX secure cluster Bianca

Preprocessing time



On a clinical settings
On our own secure server munin



 <https://aws.amazon.com/>


Johannes Alneberg



 <https://aws.amazon.com/>

- Improved AWS usage

Johannes Alneberg



Loading report...

General Stats

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

Picard

Samtools

Percent Mapped



Alignment metrics

QualiMap

Coverage histogram

Cumulative genome coverage

Insert size histogram

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Contact Name Maxime Garcia

Contact E-mail max.u.garcia@gmail.com

Genome smallGRCh37

Loading report...

Report generated on 2018-06-29, 13:54 based on data in: /home/nax/workspace/github/Sarek/work/96/3fe7059b9b38097724fc981cea80cf

General Statistics

Copy table

Configure Columns

Plot

Showing 44/44 rows and 10/23 columns.

Sample Name	% Dups	% GC	M Seqs	% Dups	Error rate	M Non-Primary	M Reads Mapped	% Mapped
1234N				4.6%				
1234N.recal					1.02%	0.0	0.0	99.1%
1234N_0.md.real					1.02%	0.0	0.0	97.1%
9876T				4.8%				
9876T.recal					1.33%	0.0	0.0	98.6%

 <http://multiqc.info/>

nf-core/ rnafusion

 <https://github.com/nf-core/rnafusion/>

- RNA-fusions detection, filtering and visualization pipeline

nf-core/ rnafusion

 <https://github.com/nf-core/rnafusion/>

- RNA-fusions detection, filtering and visualization pipeline
- Developed with NGI



Martin Proks

nf-core/ rnafusion

 <https://github.com/nf-core/rnafusion/>

- RNA-fusions detection, filtering and visualization pipeline
- Developed with NGI
- Support from The Swedish Childhood Tumor Biobank



Martin Proks

- Pipeline
 - fusioncatcher
 - STAR-Fusion
 - Ericscript
 - Pizzly
 - Squid
- Visualization
 - Fusion Inspector
- Fusion database
 - FusionGDB

Results

[Dashboard](#)

SUMMARY

Dashboard fusion summary

List of all detected fusions

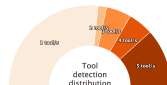
Made with by Martin Proke
at SCL/HiLaB

Summary

Generated 05/01/2019
reads

The pipeline ran successfully and was able to detect 30 fusions for sample reads.
We were able to find and describe 12 fusions from our local database.
Tools used during the analysis were Fusioncatcher, Star_fusion, Ericscript, Pizzly, Squid.

Dashboard fusion summary



List of all detected fusions

Show 25 entries

[Copy](#) [CSV](#) [PDF](#) [Print](#)Search:

Fusion gene	%	Found in DB	%	fusioncatcher	%	star_fusion	%	ericscript	%	pizzly	%	squid	%	Tools hits	%
AKAP9—BRAF		Found		Found		Found		Found		Found		Found		5	
BRD4—NTRK1		Found		Found		Found		Found		Found		Found		5	
ETV6—NTRK3		Found		Found		Found		Found		Found		Found		5	
ENGR1—ATF1		Found		Found		Found		Found		Found		Found		5	
FGFR3—TACC3		Found		Found		Found		Found		Found		Found		5	
HXK3—RET		Found		Found		Found		Found		Found		Found		5	

rnafusion results page



The CMM server room



munin

Acknowledgments



Barntumörbanken	Elisa Basmaci Szilveszter Juhas Gustaf Ljungman Monica Nistér Gabriela Prochazka Johanna Sandgren Teresita Díaz De Ståhl Katarzyna Zielinska-Chomej	NGI	Johannes Alneberg Anandashankar Anil Franziska Bonath Orlando Contreras-López Phil Ewels Sofia Haglund Max Käller Anna Konrad Pär Lundin Remi-Andre Olsen Senthilkumar Panneerselvam Fanny Taborsak Chuan Wang	NBIS	Sebastian DiLorenzo Malin Larsson Marcel Martin Markus Mayrhofer Björn Nystedt Markus Ringné Pall I Olason Jonas Söderberg
Grupp Nistér	Saad Alqahtani Min Guo Daniel Hägerstrand Anna Hedrén Martin Proks Rong Yu Jian Zhao	Clinical Genetics	Jesper Eisfeldt	Clinical Genomics	Kenny Billiau Hassan Foroughi Asl Valtteri Wirta
				Nextflow folks	Paolo Di Tommaso Sven Fillingier Alexander Peltzer



Any questions?

 <https://github.com/SciLifeLab/Sarek/>

 <https://gitter.im/SciLifeLab/Sarek/>

 <http://sarek.scilifelab.se/>

 <https://github.com/nf-core/rnafusion/>

 <https://nfcore.slack.com/>

 <https://nf-co.re/>

