Store, align and explore your genome outside the Cloud, at home, on your PC

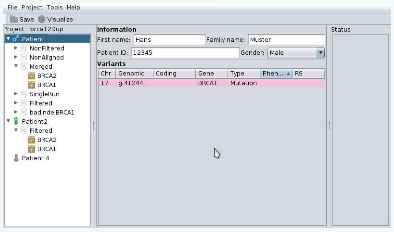
P. Kuonen, B. Wolf, University of Applied Sciences Western Switzerland
JT den Dunnen, Leiden University Medical Center
D.Atlan, Phenosystems SA
T. Dandekar, University of Würzburg

Description

Cost of whole genome sequencing is rapidly approaching 1000 euro, putting it within reach of amateur geneticists, scientifically curious consumers as well as teachers developing hands-on *omics training courses. Current direct-to-consumer offers do not give consumers the possibility to analyze the actual data, try various alignment algorithms and their parameters nor with the thresholds for calling variants. The raw sequence data remain at the provider and the algorithms used are a black box. Consumers like amateur geneticists and training course developers prefer to have control of the analysis pipeline for didactic or other purposes and thus need easy to use, flexible software that runs on standard PC's. We present GensearchNGS which we developed to analyse a whole genome on standard PC/Mac, from read alignment to variant detection and annotation. It includes an easy to use graphical user interface, integrating various public and proprietary alignment algorithms through plugins. Open file format standards offer the freedom to exchange data, with for example LOVDs gene variant database to view/query the database with its own associated functionality. We present the ability of GensearchNGS to perform a full genome analysis on a home computer.

Features

The goal in creating the software was to give the user all the tools needed to analyze NGS data, while keeping the software accessible. We created all the tools needed to go from raw sequencing data to a final variant report.
The software allows the import of all the standard NGS sequencing formats, FASTQ, and the Roche 454 SFF and FNA files. During the import the user can get an overview of the quality of the data, similar to the functionality of FASTQC[1]. After verifying the quality, the user can specify different filters, for example to split the data based on barcodes or to remove bad quality reads.
To help the user organize his data, a different approach to most other NGS analysis tools has been taken. After creating a project, the user can add Patients to the project, allowing to group the NGS data by patient and also creating variant lists tied to individual patients.
After importing the raw sequencing data, the user can align the data. The download of the reference sequences is integrated into the software and allows easy downloading of the human reference genome from the UCSC[2] servers. To align the data, a custom aligner was developed, but popular alignment software like BWA[3], Bowtie[4], Bowtie2[5] and Stampy[6] can also be used. The custom aligner allows the user to align the data on multiple computers in the same network, without having to manually setup a grid environment.
The aligned data can then be scanned for variants and the coverage of the reference sequence can be calculated. The user can easily filter the variants to only those that are relevant to him, thanks to interactive filters on properties like frequency, coverage or the consequence, which is based on the Variant Effect Predictor[7]. At any time, the individual variants or the complete alignment can be visualized in a fast and interactive way.
After identifying relevant Variants, they can be saved to be viewed at a later moment, or exported into a Variant report or VCF file.



Project : brca12Dup

- Patient
  - NonFiltered
  - NonAligned
  - Merged
    - BRCA2
    - BRCA1
  - SingleRun
  - Filtered
  - badIndelBRCA1
- Patient2
  - Filtered
    - BRCA2
    - BRCA1
- Patient 4

Information

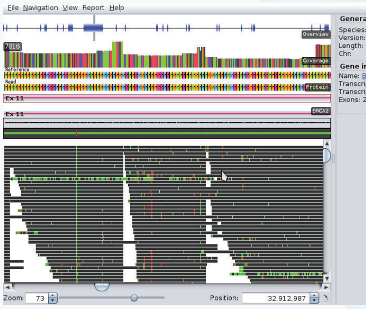
First name: Hans Family name: Muster
Patient ID: 12345 Gender: Male

Variants

Chr	Genomic	Coding	Gene	Type	Phen... ▲	RS
17	g.41244...		BRCA1	Mutation		

Status

Patients overview

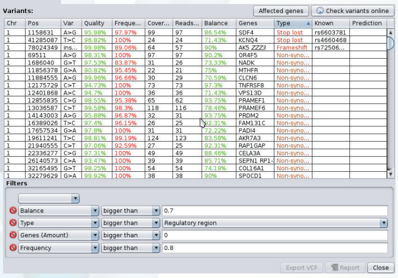


Zoom: 73
Position: 32,912,987

General info
Species: Human
Version: hg19
Length: 90,195
Chr: 13

Gene info
Name: BRCA2
Transcript: 001
Transcripts: 6
Exons: 28

Sequence alingment viewer



Variants:

Chr	Pos	Var	Quality	Freque...	Cover...	Reads...	Balance	Genes	Type ▲	Known	Prediction
1	1158631	A>G	95.98%	97.97%	99	97	86.54%	SDF4	Stop lost	rs6603781	
1	41285087	T>C	96.82%	100%	24	24	71.43%	KCNQ4	Stop lost	rs4660468	
1	78024349	ins...	99.98%	89.06%	64	57	90%	AK5 ZZZ3	Frameshift	rs72506...	
1	69511	A>G	98.31%	100%	97	97	90.2%	OR4F5	Non-syno...		
1	1686040	G>T	97.53%	83.87%	31	26	73.33%	NADK	Non-syno...		
1	11856378	G>A	90.82%	95.45%	22	21	75%	MTHFR	Non-syno...		
1	11884555	A>G	99.96%	96.66%	30	29	70.59%	CLCN6	Non-syno...		
1	12175729	C>T	94.73%	100%	73	73	97.3%	TNFRSF8	Non-syno...		
1	12401868	A>C	94.7%	100%	36	36	71.43%	VPS13D	Non-syno...		
1	12855835	C>G	98.55%	95.38%	65	62	93.75%	PRAMEF1	Non-syno...		
1	13036587	C>T	99.58%	98.3%	118	116	78.46%	PRAMEF6	Non-syno...		
1	14143003	A>G	95.88%	96.87%	32	31	93.75%	PRDM2	Non-syno...		
1	16389026	T>C	97.4%	96.15%	26	25	92.31%	FAM131C	Non-syno...		
1	17657534	G>A	97.8%	100%	31	31	72.22%	PADI4	Non-syno...		
1	19611241	T>C	98.81%	99.19%	124	123	83.58%	AKR7A3	Non-syno...		
1	21940555	C>T	97.06%	92.59%	27	25	92.31%	RAP1GAP	Non-syno...		
1	22336277	C>G	97.31%	100%	49	49	88.46%	CELA3A	Non-syno...		
1	26140573	C>A	93.47%	100%	39	39	85.71%	SEPN1 RP1-	Non-syno...		
1	32165495	G>T	98.25%	100%	54	54	74.19%	COL16A1	Non-syno...		
1	32279629	G>A	99.92%	100%	38	38	90%	SPOCD1	Non-syno...		

Filters

Balance | bigger than | 0.7
Type | bigger than | Regulatory region
Genes (Amount) | bigger than | 0
Frequency | bigger than | 0.8

Live variant filtering

Conclusion

The ability to analyze whole genome data is possible and doable on normal home computers. With a basic understanding about NGS data-analysis one can use GensearchNGS and perform the most common tasks in NGS data-analysis, like creating variant report ans coverage analysis. The algorithms improved enough so that they are usable in a reasonable time on standard computers, without the need to offload the analysis to the cloud or a dedicated grid. The problem of storing the data is one that is not yet resolved, but in the context of a single patient analysis, which what most hobbyist would do, it poses no problem. Future work will go into exploration on how to store the NGS data effectively on a personal computer, allowing the effective storage of more than one patient dataset, but also on how to take advantage of cloud environments and GPUs to accelerate the data analysis.

Alignment setup

Full genome analysis

While GensearchNGS started as a software package that was used to analyze single genes, with the arrival of full genome sequencing, new challenges had to be solved. The amount of data to analyze is not only a problem in term of storage space, but primarily processing power.
To test the ability to perform the complete analysis on a home computer, a full genome dataset was used. It consisted of 995 million sequences to give a 30x coverage. Of thise reads, 94% where successfully aligned against the human reference genome (hg19). Two standard home computers where connected trough GensearchNGS to perform this task in about **40 hours**.
GensearchNGS can filter the variants easily, thus reducing the amount of variants that need to be manually inspected. The initial scan, which selected variants with a frequency higher than 40%, a coverage of over 20 bases and which are close to a known gene, detected **1'714'314 Variants**. By limiting the variants to those with a frequency over 80%, having a balanced coverage (supported by both forward and backward reads), and a minimum quality requirement, only 500'858 variants were left. This list could be further reduced by only selecting Variants that were worse than a "Synonymous SNP", as defined by VEP[7]. This reduced the amount down to **3262 Variants** considered "interesting". From those, 86 are not known in dbSNP, also easily verified trough the software.
Additionally, the aligned data was compared to the Illumina 660K SNP chip data. Out of 657'367 variants, 656'983 where successfully converted from hg18 to hg19 coordinates. Of those, **122'419 (18.36%)** showed **homozygous** and **174'773** (26.6%) **heterozygous mutations**. 19'857 (3%) positions did have a coverage lower than 12, making realiable variant calling impossible.

References

[1] FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] UCSC Genome Bioinformatics http://genome.ucsc.edu/
[3] Li, H., & Durbin, R. (2009). Fast and accurate shortFast and accurate long-read alignment with Burrows-Wheeler Transform read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–60. doi:10.1093/bioinformatics/btp324
[4] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology, 10(3), R25. doi:10.1186/gb-2009-10-3-r25
[5] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357–9. doi:10.1038/nmeth.1923
[6] Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. GenomeRes 2011;21:936–9.
[7] Ensembl Variant Effect Predictor http://www.ensembl.org/info/docs/variation/vep/index.html

Contact: Beat Wolf, beat.wolf@hefr.ch

2013