Genome sequencing data analysis

—

Genetic causes of Parkinson’s disease and hypertrophic cardiomyopathy

Introduction and work plan

- Parkinson's disease (PD)
  - disorder of the central **nervous system**
  - affects the motor system[1]
  - both **genetic** and **environmental** factors
- Hypertrophic cardiomyopathy (HCM)
  - a portion of the **myocardium is thickened**[2, 3, 4, 5, 6]
  - cause of sudden cardiac death
  - **inherited** as an autosomal dominant trait, or come from a ***de novo*** **mutation**
- Study
  - 30 patients (with PD or HCM)
  - DNA sequencing with Illumina HiSeq machines
  - **preprocessing** (filtering, trimming)
  - **mapping** with a reference genome
  - variant **discovery**
  - **annotation** and analysis

I- Preprocessing

Formatıng of the reads before mappıng

1. Quality check

   Qualıty dıstrıbutıon, k-mer content…

2. Filtering

   “Bad reads” removal

3. Trimming

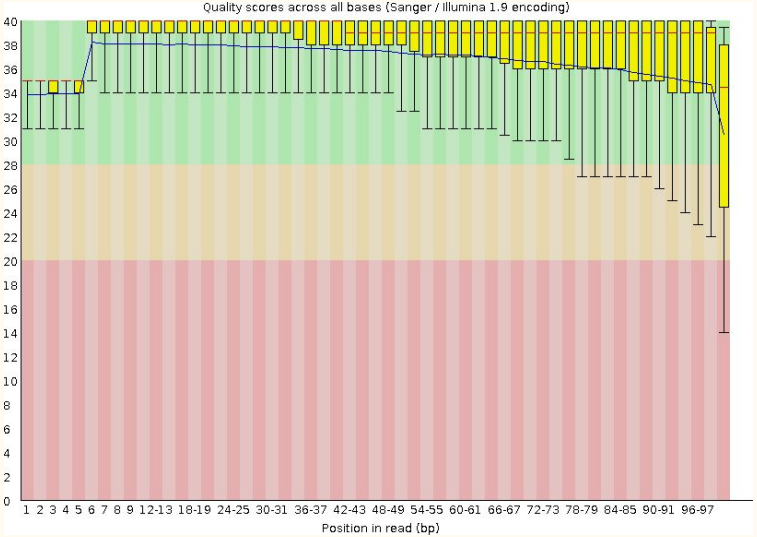
   “Bad bases” removal

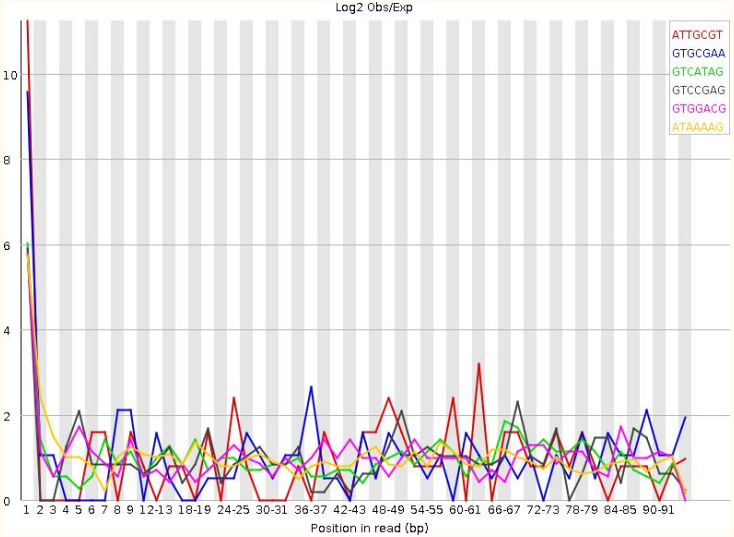
4. Pairing

   Partıtıon paıred-end / orphans

I. 1 - Quality check[7]

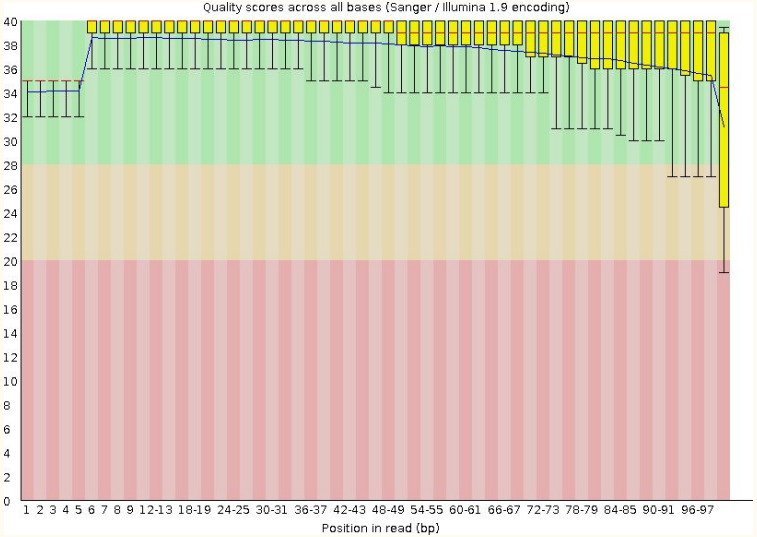
```
fastqc <*.fastq>
```

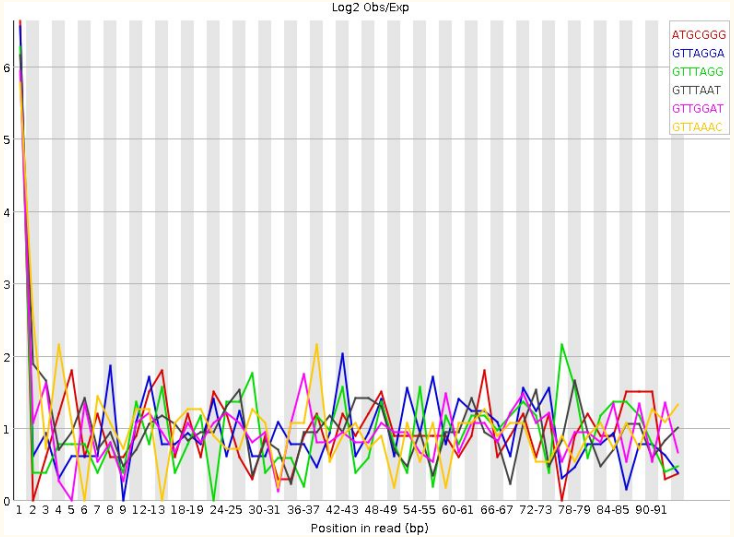




I. 2 - Filtering

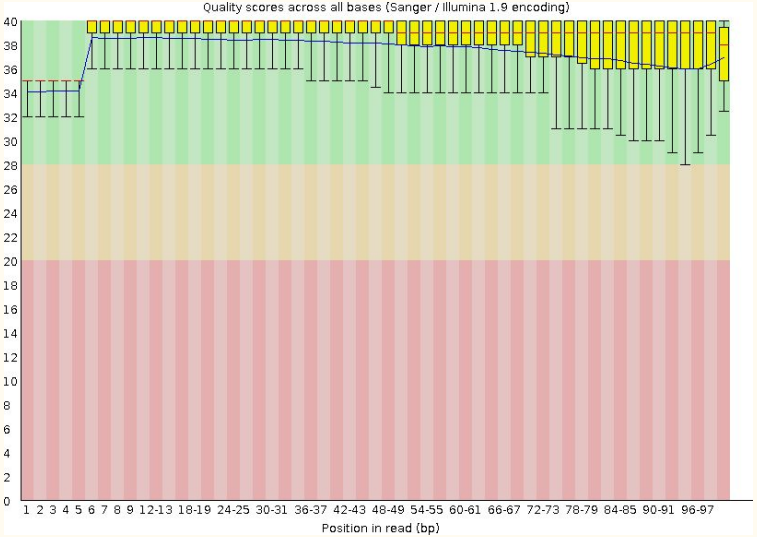
```
fastq_quality_filter -q 20 -p 80 -i <i.fastq> -o <o.fastq> -Q33
```

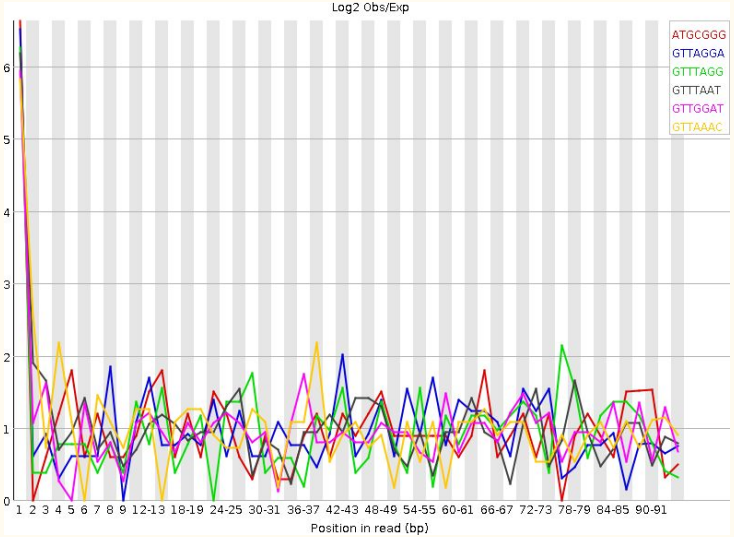




I. 3 - Trimming

```
fastq_quality_trimmer -t 30 -l 20 -i <i.fastq> -o <o.fastq> -Q33
```

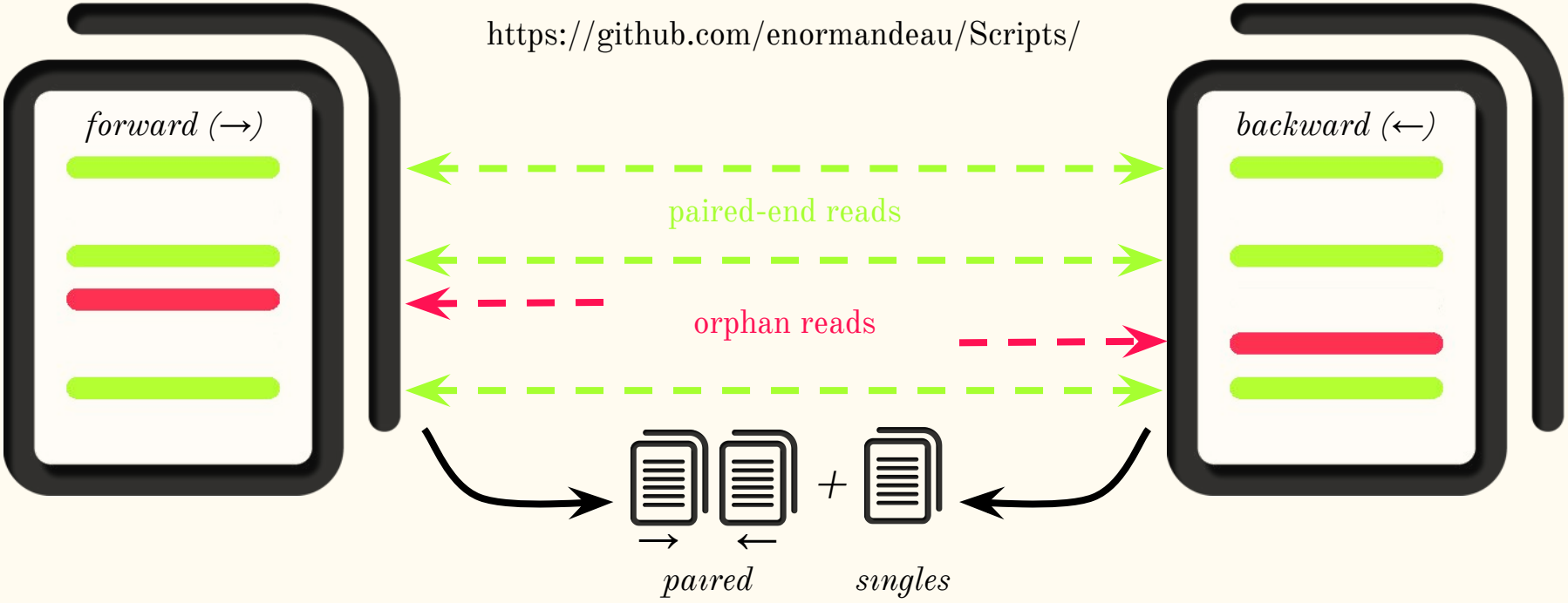




I. 4 - Pairing

Script on GitHub[8]

https://github.com/enormandeau/Scripts/



II- Mapping

Read-alignments against a reference genome

1. Indexing

   Genome preprocessing for faster queries

2. Alignment

   Approximate matching against reference

3. Visualization

   First approach with IGV

4. Benchmarking

   Softs comparison and mapping

II. 1 - Indexing

- Burrows-Wheeler Aligner (BWA)[9]

```
bwa index hg38.fasta
```

➔ **hg38.fasta.pac**
➔ **hg38.fasta.bwt**
➔ **hg38.fasta.sa**
➔ **hg38.fasta.amb**
➔ **hg38.fasta.ann**

- Bowtie2[10]

```
bowtie2-build hg38.fasta hg38
```

➔ **hg38.1.bt2**
➔ **hg38.2.bt2**
➔ **hg38.3.bt2**
➔ **hg38.4.bt2**
➔ **hg38.rev.1.bt2**
➔ **hg38.rev.2.bt2**

II. 2 - Alignment

- Burrows-Wheeler Aligner (BWA)
  - Paired reads : forward (→) read **and** backward (←) read

```
bwa mem <ref file> <r1.fastq> <r2.fastq> > <o.sam>
```

  - Single reads : forward (→) read **or** backward (←) read

```
bwa mem <ref file> <rs.fastq> > <out.sam>
```

- Bowtie2
  - Paired reads : forward (→) read **and** backward (←) read

```
bowtie2 --phred33 -a -x <ref prefix> -1 <r1.fastq> -2 <r2.fastq> -S <o.sam>
```

  - Single reads : forward (→) read **or** backward (←) read

```
bowtie2 --phred33 -a -x <ref prefix> -U <rs.fastq> -S <o.sam>
```

II. 3 - Visualization

- Preprocessing[14]
  - Conversion

```
samtools view -b <i.sam> > <o.bam>
```

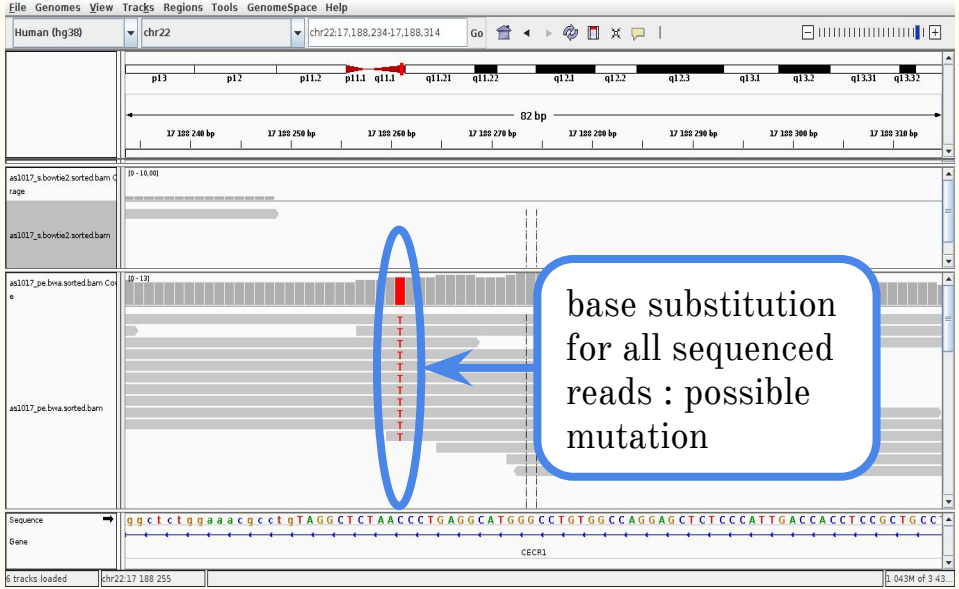
  - Sorting

```
samtools sort -o <o.bam> <i.bam>
```

  - Indexing

```
samtools index <o.bam>
```

- Integrative Genomics Viewer (IGV)[12, 13]



II. 4 - Benchmarking

Statistics of BWA and Bowtie2 for the first patient (ID : **as1017**)

Soft	% aligned	% proper pairs	% reversed	% supplementary or not primary	% alignment on different chromosome
BWA (paired-end)	**99.997**	**97.534**	50.045	8.300×10^{-2}	2.182
Bowtie2 (paired-end)	99.836	96.550	**50.005**	0[1]	**2.045**
BWA (singles)	**99.981**	---	51.007	8.587×10^{-2}	---
Bowtie2 (singles)	99.786	---	**50.960**	0[1]	---

[1] this result is due to Bowtie2 default parametrization ; that is why it is difficult to say that Bowtie2 is “better” with regard to multimaps.

III- Variant discovery

Research of polymorphisms, annotation and analysis

1. Preprocessing

   Genome indexing, sequence maps conversion

2. Variant calling

   Mutation discovery with the GATK

3. Annotation

   Mutation identification with Annovar

4. KEGG mapping

   Final graphic results

III. 1 - Preprocessing

- Reference genome preprocessing
  - Sequence dictionary generation[18]

```
java -jar picard.jar
CreateSequenceDictionary R=hg38.fasta
O= hg38.dict
```

  - Genome indexing

```
samtools faidx hg38.fasta
```

- Sequence maps (**.sam**) conversion
  - Paired-end (**pe**) and singles (**s**) merging and conversion

```
samtools merge <o.sam> <pe.sam> <s.sam>
```

  - Conversion

```
samtools view -b <i.sam> > <o.bam>
```

  - Header addition, sorting, indexing

```
java -jar picard.jar...
```

III. 2 - Variant calling[15, 16, 17]

- Use of **HaplotypeCaller**

```
java -jar GenomeAnalysisTK.jar -nct 30 \
                               -T HaplotypeCaller \
                               -R hg38.fasta \
                               -I <i.bam> \
                               --genotyping_mode DISCOVERY \
                               -stand_emit_conf 10 \
                               -stand_call_conf 30 \
                               -O <o.vcf>
```

- **.vcf** merging

```
bgzip <i.vcf> > <i.vcf.gz>
tabix -p vcf <i.vcf.gz>
vcf-merge *.vcf.gz > <o.vcf>
```

III. 3 - Annotation[19]

- Basic stats

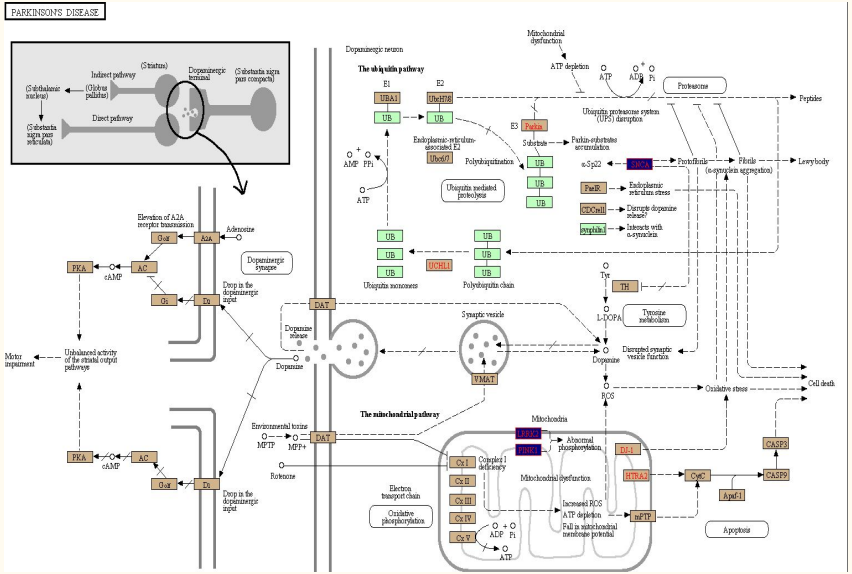
	PD (%)	HCM (%)
% exonic (& splicing) mutations	6.172	6.171
% non-synonymous mutations	3.169	3.184
# genes with non-synonymous mutations	40 885	40471

- Stats for “critical” genes

	“critical” genes (# non-synonymous mutations)	
PD	LRRK2 (8) VPS35 (1) ATP13A2 (5) PINK1 (4) PLA2G6 (1)	
HCM	MYBPC3 (5) MYL2 (1) MYH7 (4) ACTC1 (1)	TTR (1) CAV3 (1) GLA (1) LAMP2 (1)

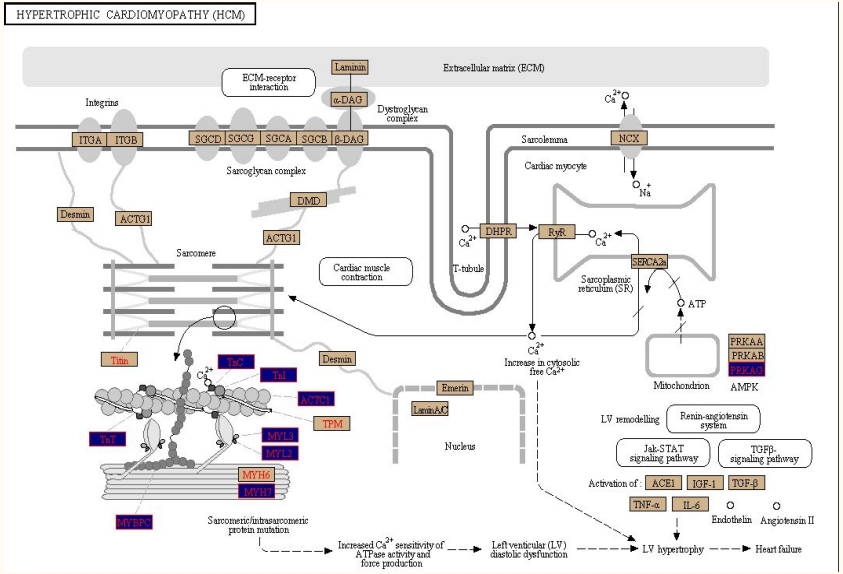
III. 4 - KEGG mapping[20]

- PD



(navy blue for "critical" mutations, tan for not synonymous mutations)

- HCM



(navy blue for "critical" mutations, tan for not synonymous mutations)

References

[1] NINDS, Parkinson's Disease Information Page, http://www.ninds.nih.gov/disorders/parkinsons_disease, [Online; accessed 30-June-2016; retrieved 10-July-2016].

[2] B. J. Maron, Hypertrophic cardiomyopathy: a systematic review, JAMA287 (10) (2002) 13081320.

[3] P. Richardson, W. McKenna, M. Bristow, B. Maisch, B. Mautner, J. O'Connell, E. Olsen, G. Thiene, J. Goodwin, I. Gyarfas, I. Martin, P. Nordet, Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the Definition and classfication of cardiomyopathies, Circulation 93 (5) (1996) 841-842.

[4] M. V. Sherrid, F. A. Chaudhry, D. G. Swistel, Obstructive hypertrophic cardiomyopathy: echocardiography, pathophysiology, and the continuing evolution of surgery for obstruction, Ann. Thorac. Surg. 75 (2) (2003) 620-632.

[5] E. D. Wigle, Z. Sasson, M. A. Henderson, T. D. Ruddy, J. Fulop, H. Rakowski, W. G. Williams, Hypertrophic cardiomyopathy. The importance of the site and the extent of hypertrophy. A review, Prog Cardiovasc Dis 28 (1) (1985) 1-83.

[6] E. D. Wigle, H. Rakowski, B. P. Kimball, W. G. Williams, Hypertrophic cardiomyopathy. Clinical spectrum and treatment, Circulation 92 (7) (1995) 1680-1692.

[7] H. Lab, FASTX-Tollkit, http://hannonlab.cshl.edu/fastx_toolkit/, [Online; accessed 16-August-2016].

[8] E. Normandeau, GitHub repository for the pairing script, https://github.com/enormandeau/Scripts, [Online; accessed 16-July-2016].

[9] H. Li, R. Durbin, Fast and accurate long-read alignment with burrows-wheeler transform (2010).

[10] B. Langmead, S. L. Salzberg, Fast gapped-read alignment with bowtie2, Nat Meth 9 (4) (2012) 357-359, brief Communication. URL http://dx.doi.org/10.1038/nmeth.1923

[11] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, Star: ultrafast universal rna-seq aligner, Bioinformatics 29 (1) (2013) 15-21, 23104886[pmid]. doi:10.1093/bioinformatics/bts635. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/

[12] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer, Nat Biotech 29 (1) (2011) 24-26. doi:10.1038/nbt.1754. URL http://dx.doi.org/10.1038/nbt.1754

References

[13] H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative genomics viewer (igv): high-performance genomics data visualization 30 and exploration, Briengs in Bioinformatics 14 (2) (2013) 178-192. arXiv:http://bib.oxfordjournals.org/content/14/2/178.full.pdf+html, doi:10.1093/bib/bbs017. URL http://bib.oxfordjournals.org/content/14/2/178.abstract

[14] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, . G. P. D. P. Subgroup, The sequence alignment/map format and samtools, Bioinformatics 25 (16) (2009) 2078-2079, btp352[PII]. doi:10.1093/bioinformatics/btp352. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/

[15] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation dna sequencing data, Nat Genet 43 (5) (2011) 491-498. doi:10.1038/ng.806. URL http://dx.doi.org/10.1038/ng.806

[16] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (9) (2010) 1297-1303.

[17] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, M. A. DePristo, From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, John Wiley & Sons, Inc., 2002. doi:10.1002/0471250953.bi1110s43. URL http://dx.doi.org/10.1002/0471250953.bi1110s43

[18] B. Institute, Picard, http://picard.sourceforge.net, [Online; accessed 15-August-2016].

[19] K. Wang, M. Li, H. Hakonarson, Annovar: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Research 38 (16) (2010) e164. arXiv:http://nar.oxfordjournals.org/content/38/16/e164.full.pdf+html, 31doi:10.1093/nar/gkq603. URL http://nar.oxfordjournals.org/content/38/16/e164.abstract

[20] M. Kanehisa, S. Goto, Kegg: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res 28 (1) (2000) 2730, gkd027[PII]. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/

Special thanks

Yuri, my superviser, **Nikolaï**, who helped by mapping the reads with STAR[11], **Dmitry**, without whom I wouldn't have been there, and **Denis**, for the early (and friendly) coffee cup !