Bioinformatic Evaluation of Next-Generation Sequencing Performance at Nancy Hospital

Alexis Praga

Laboratoire de génétique - CHRU de Nancy - Dr Céline Bonnet

July 8, 2021

Part 1: Manuscript

Introduction: Next-Generation Sequencing

Increased role in clinical practice:

- faster and cheaper (massively parallel)
- Uses:
  - Gene panel: set of targeted genes in week/months [intellectual disabilities]
  - Whole Exome Sequencing: genes not known, 1% genome
  - Whole Genome Sequencing: all genome

Introduction: Next-Generation Sequencing

Increased role in clinical practice:

- faster and cheaper (massively parallel)
- Uses:
  - Gene panel: set of targeted genes in week/months [intellectual disabilities]
  - Whole Exome Sequencing: genes not known, 1% genome
  - Whole Genome Sequencing: all genome

Goal of the internship: validate sequencing results

- assess performances
- evaluate errors
- quantify improvements

Introduction: issues with validation

1. Which reference ?
   - sequencing error
   - patient mutation
2. Ambiguous file format for variants !

Example: Variant `TCCG` → `CC`

1. TCCG → CCG then CCG → CC
2. TCCCG → CCG then CCG → CC

3. No consensus for metrics

	Truth = *Reference/Variant1*
Reference/Variant1	TP
Reference/Variant2	FP ? FN ?
Variant2/Variant1	FP ? FN ?

Introduction: issues with validation

1. Which reference ?
   - sequencing error
   - patient mutation
2. Ambiguous file format for variants !

Example: Variant `TCCG` $\rightarrow$ `CC`

1. TCCG $\rightarrow$ CCG then CC**G** $\rightarrow$ CC
2. **TC**CCG $\rightarrow$ **C**CG then **CCG** $\rightarrow$ **C**C

3. No consensus for metrics

	Truth = *Reference/Variant1*
Reference/Variant1	TP
Reference/Variant2	FP ? FN ?
Variant2/Variant1	FP ? FN ?

Introduction: issues with validation

1. Which reference ?
   - sequencing error
   - patient mutation
2. Ambiguous file format for variants !

Example: Variant `TCCG` $\rightarrow$ `CC`

1. TCCG $\rightarrow$ CCG then CC**G** $\rightarrow$ CC
2. **TC**CCG $\rightarrow$ **C**CG then **CC**G $\rightarrow$ **C**C

3. No consensus for metrics

	Truth = *Reference/Variant1*
Reference/Variant1	TP
Reference/Variant2	FP ? FN ?
Variant2/Variant1	FP ? FN ?

Methods: An answer with Krusche et al. 2019

1. Reference dataset

- reference DNA on multiple technologies
- apply filters manually (Illumina) or with a trained model (GIAB)

$\Rightarrow$ "high-confidence" regions

2. Hap.py

- Python/C++ script by Illumina
- 2 variant comparison algorithms : custom, `vcfeval`

3. Defined metrics

	Truth = *Reference/Variant1*
Reference/Variant1	TP
Reference/Variant2	FP with allele mismatch
Variant2/Variant1	FP with genotype mismatch

Methods: summary

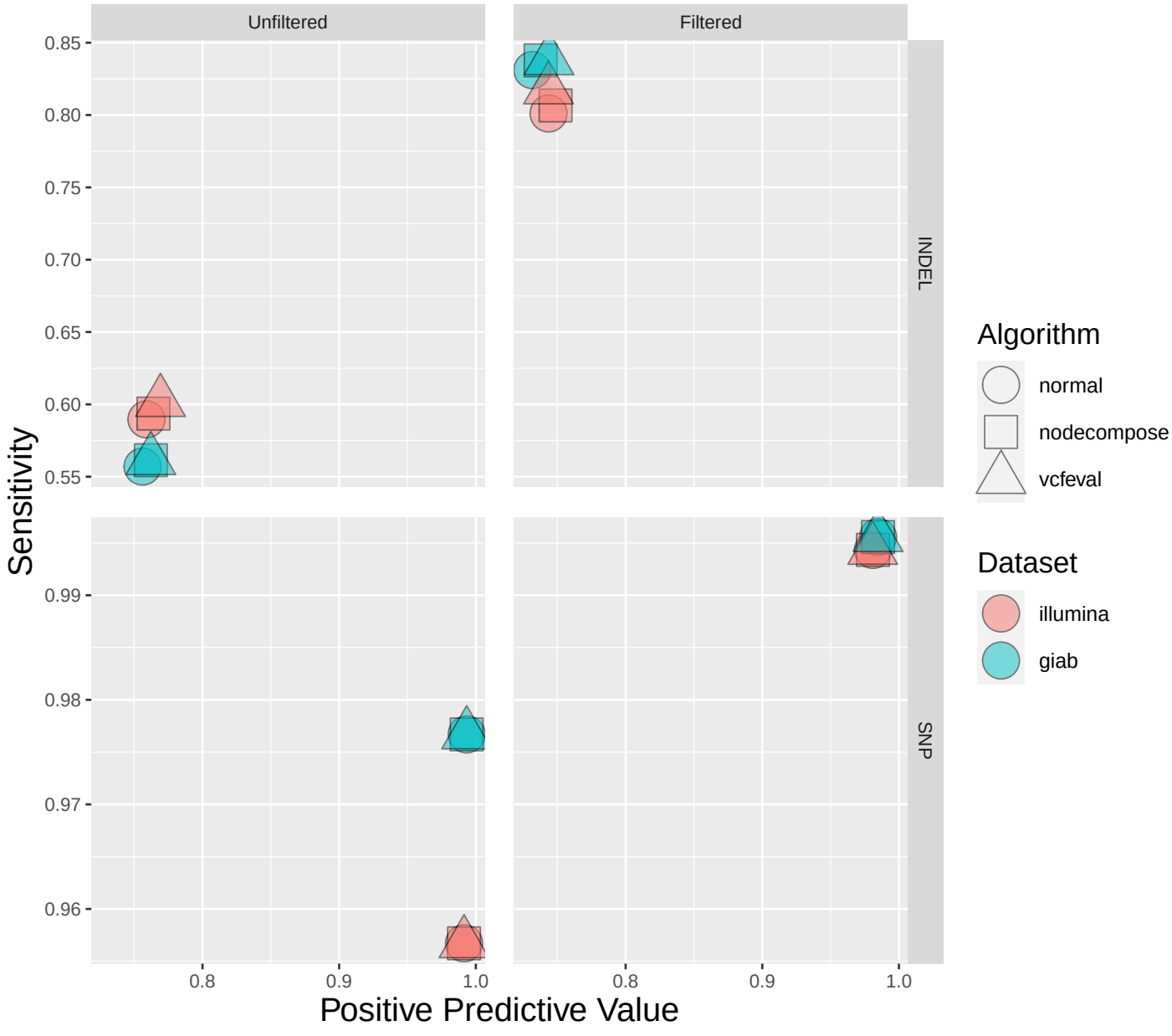
1. Sequence DNA of patient NA12878 in Nancy
2. Download 2 reference datasets (Illumina and GIAB)
3. Use `hap.py` to compare variants (2 algorithms) (VCF format)

Results: definition

$$\text{Positive Predictive Value} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (= \text{precision})$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (= \text{recall})$$

Results: Impact of datasets, algorithm and filters



Results: False Negative

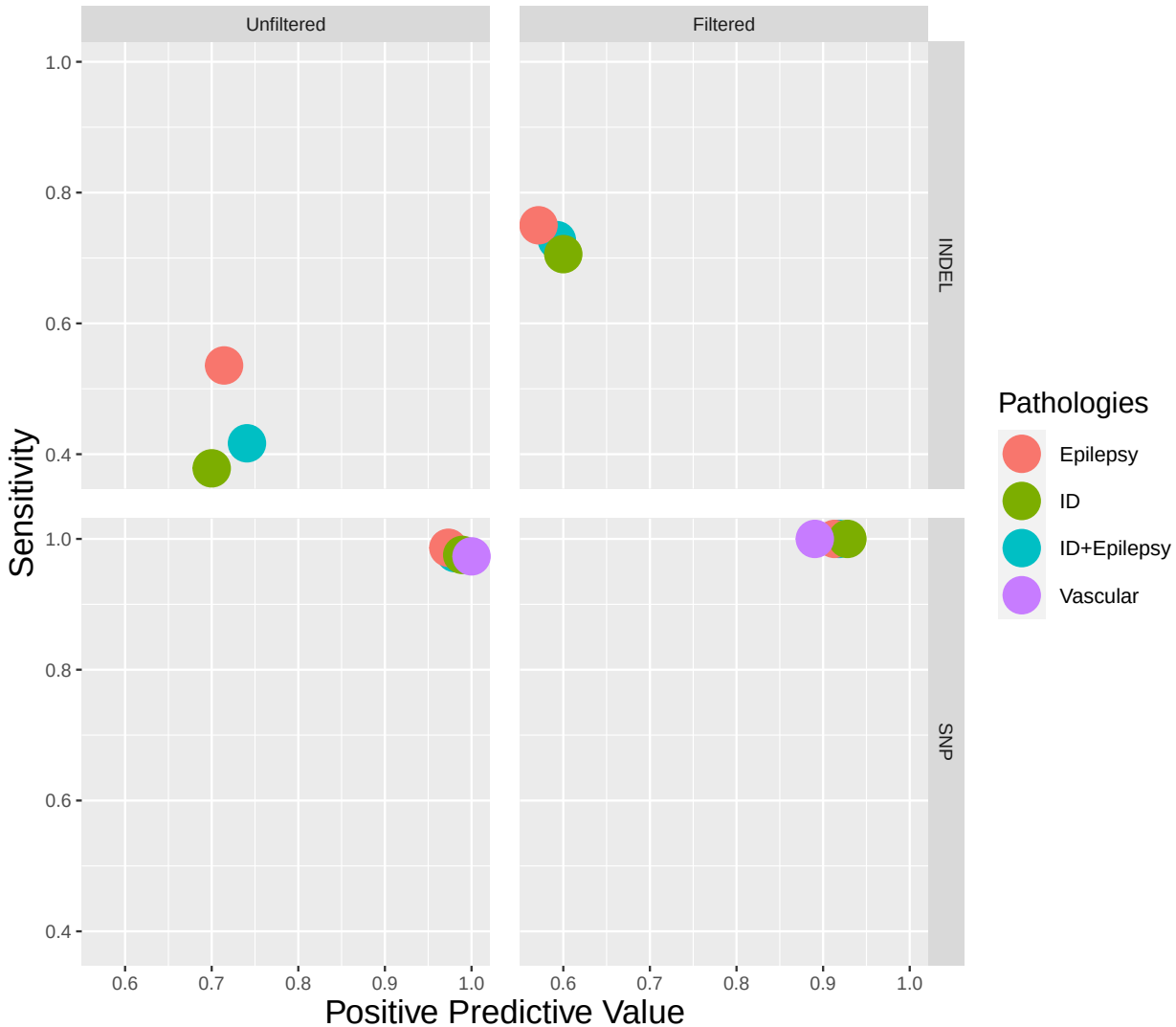
- Precision and recall are equivalent to whats expected for SNP but too low for indels

High ratio of false negative

- not due to dataset, nor algorithm
- not due to a specific chromosome
- in "difficult" regions, mostly homopolymers

Results: Clinical implications

- A set of genes instead of a patient
- Finding a gene transcript is difficult $\Rightarrow$ APPRIS database



Conclusion

- Standardized benchmarking method to the sequencing pipeline of Genetics Laboratory (Nancy)
- Reproductible (user manual in manuscript)
- Excellent performance for Single Nucleotide Polymorphism, sub-par for indels
- No major cause for rather high False Negative count
- Promising performances for a group of pathologies
- Need for further testing (other DNA, other pipelines)

Part 2: Tasks

Tasks

- Bibliography (Krusche et al. 2019 . . .)
- Lab visit

Tasks

- Bibliography (Krusche et al. 2019 . . .)
- Lab visit
- Analysis
  - File formats (BED, VCF)
  - Installing `hap.py` (Ubuntu, Archlinux)
  - Running benchmark (pre-processing, statistics)
  - FN analysis
- Group of pathologies: which gene transcript ?
  - MANE (20% missing): fusion of Refseq anad ENSEMBL gene set
  - APPRIS (10% missing): conservation and the characteristics of known proteins
  - comparision with BED from Illumina

Tasks

- Bibliography (Krusche et al. 2019 . . .)
- Lab visit
- Analysis
  - File formats (BED, VCF)
  - Installing `hap.py` (Ubuntu, Archlinux)
  - Running benchmark (pre-processing, statistics)
  - FN analysis
- Group of pathologies: which gene transcript ?
  - MANE (20% missing): fusion of Refseq anad ENSEMBL gene set
  - APPRIS (10% missing): conservation and the characteristics of known proteins
  - comparision with BED from Illumina
- Manuscript
- User manual for reproductibility

Appendix: further reading

P. Krusche and the Global Alliance for Genomics and Health Benchmarking Team

Best Practices for Benchmarking Germline Small-Variant Calls in Human Genomes

Nature Biotechnology

Appendix: contingency table

Table 1 | Contingency table describing the GA4GH definitions of TP, FP, FN, FP.AL, FP.GT, and unknown (UNK)

		Truth				
	Genotype	**Ref/ ref**	**Ref/var1**	**Var1/ var2**	**Var1/ var1**	**Outside bed**
Query	Ref/ref	–	FN	FN	FN	–
	Ref/var1	FP	TP	**FP.GT**	**FP.GT**	UNK
	Ref/var2	–	**FP.AL**	**FP.GT**	**FP.AL**	–
	Ref/var3	–	–	**FP.AL**	–	–
	Var1/var2	FP	**FP.GT**	TP	**FP.GT**	UNK
	Var1/var3	–	–	**FP.GT**	–	–
	Var2/var3	–	**FP.AL**	**FP.GT**	**FP.AL**	–
	Var3/var4	–	–	**FP.AL**	–	–
	Var1/var1	FP	**FP.GT**	**FP.GT**	TP	UNK
	Var2/var2	–	**FP.AL**	**FP.GT**	**FP.AL**	–
	Var3/var3	–	–	**FP.AL**	–	–

Table 2 | Examples of several combinations of truth and query SNV genotypes and how they are counted as TP, FP, FN, FP.GT, and FP.AL

REF	Truth	Query	Counted as
A	C/C	C/C	1 TP
A	A/A	C/C	1 FP
A	C/C	A/A	1 FN
A	C/C	A/C	1 FP, 1 FN, 1 FP.GT
A	C/C	G/G	1 FP, 1 FN, 1 FP.AL
A	C/G	C/C	1 FP, 1 FN, 1 FP.GT