Paired-End Read Length Lower Bounds for Genome Re-sequencing

Rayan Chikhi
ENS Cachan Brittany
PhD student in the Symbiose team, Irisa, France

Next-Generation Sequencing

Next-gen vs. traditional (Sanger) **DNA sequencers** :

- ⬆ Throughput
- ⬇ Cost
- ⬆ High quality reads (≈0.4% error rate per base on Illumina)
- ⬇ Shorter reads (starting at 25 bp)
- ⬆ Paired reads (both ends of a fragment of size 200-10k nt)

NEXT-GENERATION SEQUENCING

Next-gen vs. traditional (Sanger) **DNA sequencers** :

- ⬆ Throughput
- ⬇ Cost
- ⬆ High quality reads (≈0.4% error rate per base on Illumina)
- ⬇ Shorter reads (starting at 25 bp)
- ⬆ Paired reads (both ends of a fragment of size 200-10k nt)

Some bioinformatics applications :

- Genome re-sequencing
- De novo (w/out reference genome) and comparative assembly

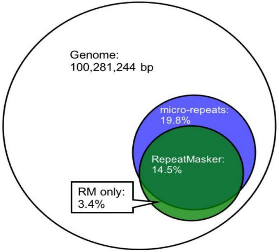
Topic of this talk : **Determine when NGS paired read length is too short for re-sequencing.**

Context : Genome re-sequencing

Re-sequencing : align reads to a reference sequence to improve it and detect SNPs/indels

Resequencing ambiguity :

```
map: AACGTATGCA
to: -AACGTTTGCA------------AACGTTTGCA--
```



From E. Mardis, Whole-genome sequencing and variant discovery in C. elegans, 2008

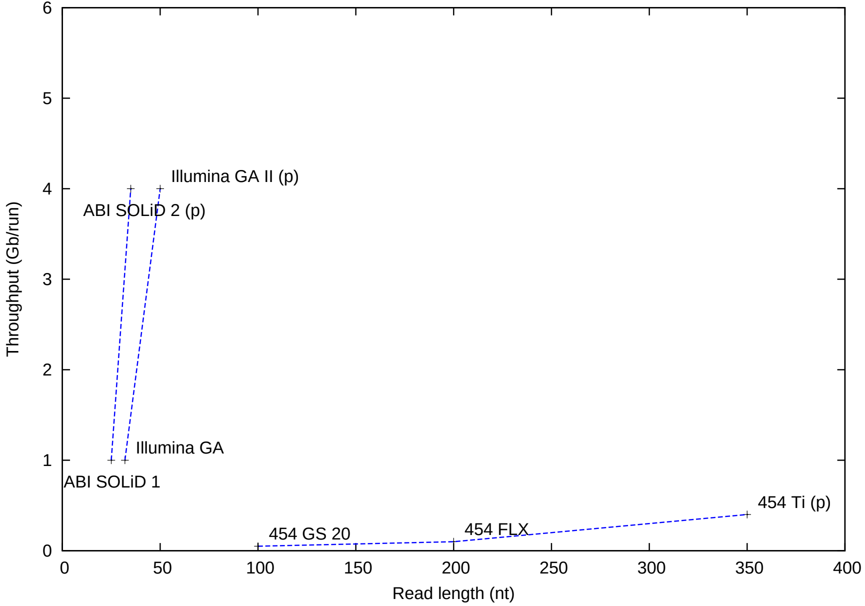
Micro-repeated (30-50 nt) regions are :
-**not re-sequencable** with short reads
-not fully predicted by simple models (such as BLAST's E-value) nor RepeatMasker

- **Solution** : analyze actual genomes

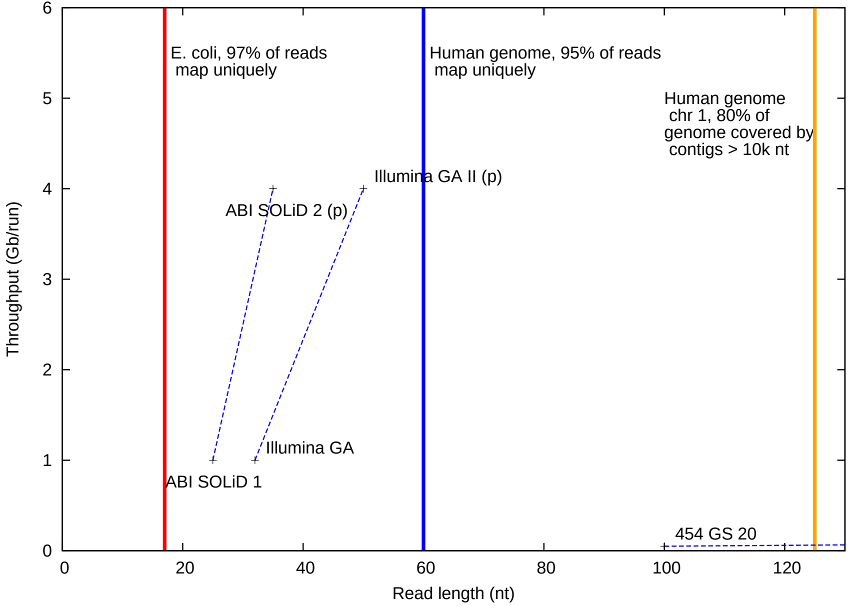
[Whiteford et al, 2005] : *perfect* uniqueness of single reads.

Genomes	Viral	Bacterial	Small eukaryote	Human
Read length for max uniqueness	12 nt (100%)	18 nt (97%)	20-50 nt (90-95%)	30-60 nt (85-95%)

SEQUENCERS CHART



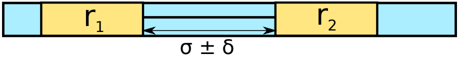
SEQUENCERS CHART



METHODS

We study the *perfect* uniqueness of mate-paired reads :

(σ, δ)-pair :



Example values for Illumina [Lee 09] : $\sigma \approx 150$, $\delta \approx 15$

a (σ, δ)-pair (r_1, r_2) is unique ⇔there is no other (r_1, r_2) pair distant of $\sigma \pm \delta$ in the genome

```
---AACGT---TTGCA------------AACGT-----TTGCA--
```

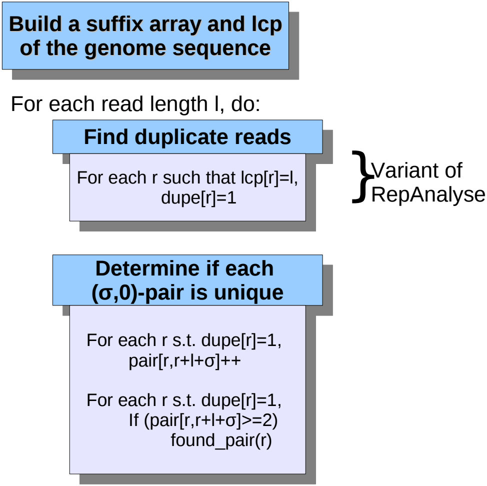
Here, the $(3, 2)$-pair AACGT−−−TTGCA is not unique.

$$U = \frac{\text{number of unique } (\sigma, \delta)\text{-pairs}}{\text{number of } (\sigma, \delta)\text{-pairs}}$$

METHODS, ALGORITHM

We developed a novel pairs-counting algorithm based on a suffix array. Here, $\delta = 0$ case is shown.

Complexity : $O(n + n\delta)$ time and memory



Build a suffix array and lcp of the genome sequence

For each read length l, do:

Find duplicate reads

```
For each r such that lcp[r]=l,
    dupe[r]=1
```

}Variant of RepAnalyse

Determine if each (σ,0)-pair is unique

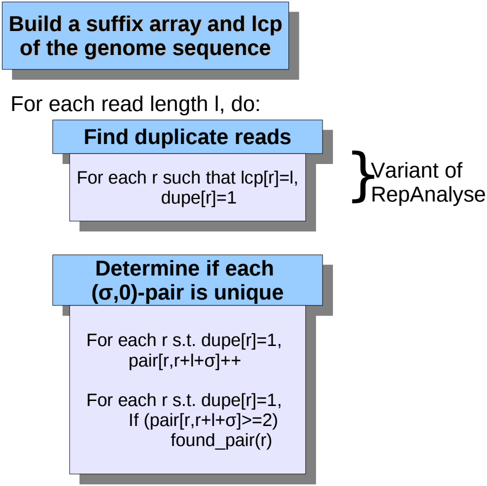
```
For each r s.t. dupe[r]=1,
    pair[r,r+l+σ]++

For each r s.t. dupe[r]=1,
    If (pair[r,r+l+σ]>=2)
        found_pair(r)
```

METHODS, ALGORITHM

Notes :

- One-time computation, but high memory usage.
- RepAnalyse on the human genome : **2 days**
- Our algorithm on Medaka genome, $\delta = 0$: **12 hrs**.



Build a suffix array and lcp of the genome sequence

For each read length l, do:

Find duplicate reads

```
For each r such that lcp[r]=l,
    dupe[r]=1
```

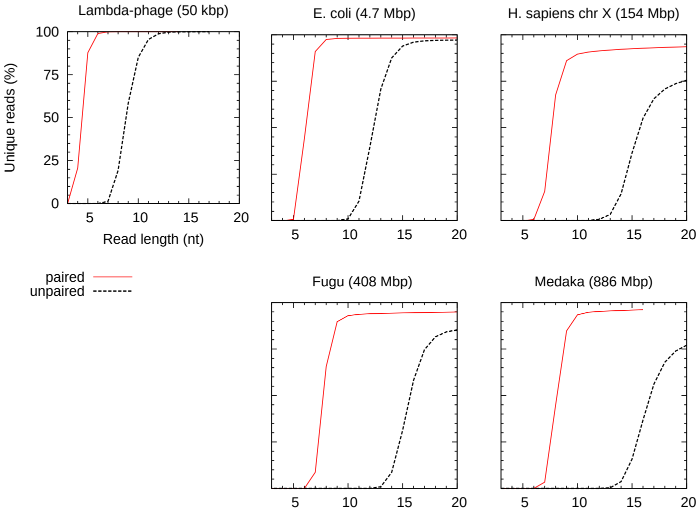
Variant of RepAnalyse

Determine if each (σ,0)-pair is unique

```
For each r s.t. dupe[r]=1,
    pair[r,r+l+σ]++

For each r s.t. dupe[r]=1,
    If (pair[r,r+l+σ]>=2)
        found_pair(r)
```

RESULTS : Comparison between paired and single uniqueness



Lambda-phage

E. coli

H. sapiens chr X

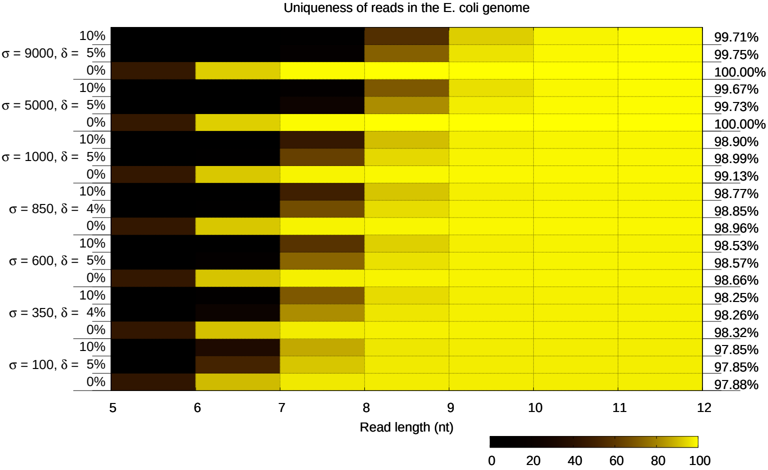
Fugu

Medaka

both strands are considered, and (σ, δ)=(300,0)

RESULTS :

Evaluation of mate-pair separation vs. uniqueness in the E. coli genome



CONCLUSION

Conclusion :

- Best recipe for paired-end sequencing :
  1. ⬆increase mate-pair separation
  2. ⬇keep variability of separation as low as possible
  3. ⬆use longer read lengths
- *Given perfect separation precision*, short (estimate : **15-20** nt) mate-paired reads should map uniquely to $\approx$ 95% of the human genome.

Perspectives :

- Use statistical models to obtain bounds for *approximate* uniqueness of reads.
- Find theoretical lower bounds for *de novo* assembly.
- Study the time complexity of paired-end *de novo* assembly (prelim results : NP-hardness of several models).

ACKNOWLEDGEMENTS

- Dominique Lavenier, PhD advisor
- Aurélien Roult @ Biogenouest genomics center
- Symbiose team and ENS Cachan Britanny CS dept

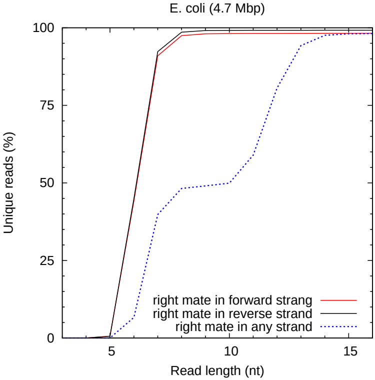
Any Question ?

SUPPL. MATERIAL : Is the right mate always in the same strand ?

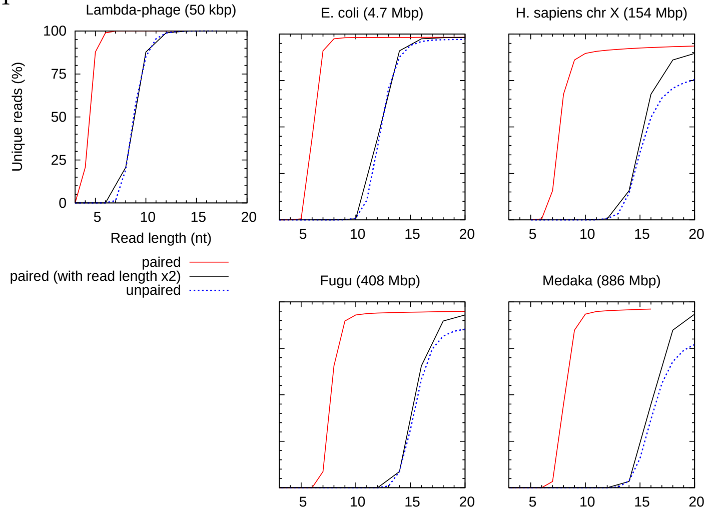
[H. Li et al, 2008] “Correct“ paired-end reads :

- SOLiD : Right mate always on the same strand
- Illumina GA : Right mate always on the other strand

If one wishes to perform structural variation detection, then the uniqueness of both correct and *discordant* reads matters.



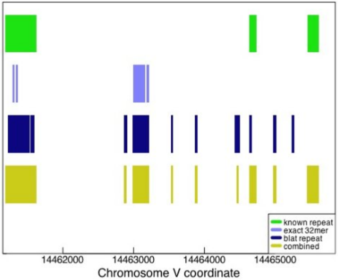
Suppl. Material : Comparison between paired, 2x paired and single uniqueness



both strands are considered, and (σ, δ)=(300,0)

SUPPL. MATERIAL : Near-perfect micro-repeats

- Counting only perfect micro-repeats gives a upper bound on unicity, hence a lower bound on read length.



known repeats (RepeatMasker)	14.5%
perfect (32bp) micro-repeats	12.7%
near-perfect (32bp) micro-repeats	23.9%
total non-resequenceable	26.2%

From E. mardis, Genome re-sequencing and variant detection using the Illumina 1G Genome Analyzer