LUMC
Leids Universitair Medisch Centrum

WES or WGS… statistics & power in practice

Gijs WE Santen, MD, PhD
Leiden University Medical Center
Leiden, The Netherlands

Case study: NGS for ID

- Suppose you plan to introduce NGS for intellectual disability diagnostics
- Some of the questions that need answering:
  1. Is WGS inherently better than WES?
  2. Should a targeted gene pael approach be considered?
  3. How to determine sequencing criteria?

(GENE PANEL), EXOME OR GENOME?

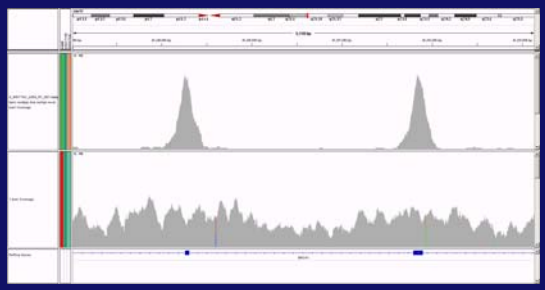
Genome or Exome?

Aspects	Genome	Exome	Even more targeted
Costs	Cheaper (in a few years)	Cheaper (now)	Cheapest (now)
Sequence depth	+/-	+	+++
Capture bias	-	+	+
Lab work	Easier (no capture)	One capture for all diseases	Different captures per disease
Bio-informatic load (storage, run times, etc)	high	low	Lowest
CNV calling	++	+/-	?

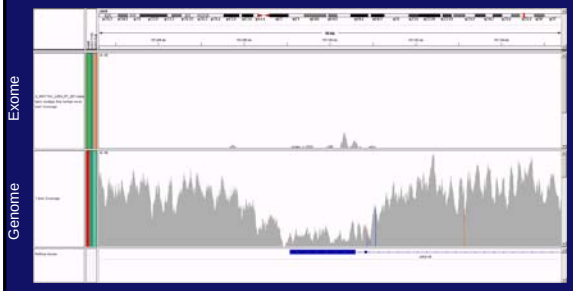
Reasons to sequence the exome rather than the genome

- ~~Costs~~ --> Sequencing prices will drop in coming years
- ~~Coverage~~ --> Sequencing prices will drop in coming years
- ~~Fewer variants/Easier interpretation~~ --> Can limit analysis to 'exome'
- ~~Storage of data~~ --> Storage prices will decrease + possibility to store exome only
- ~~Bio-informatic processing~~ --> Moore's law!

Exome versus genome



Exome versus genome: exon 1



How to compare genome versus exome in practice (1)

- Optimally: sequence large number of patients/trios using the state of the art exome AND genome technology
  - Very expensive
- Perform genome sequencing in exome-negative cases
  - Difficult to make fair comparison (confounders are e.g. sequencing depth)
- Sequence exome and genome for a limited number of patients and make thorough comparison
  - Difficult to make fair comparison (confounders are e.g. sequencing depth)

How to compare genome versus exome in practice (2)

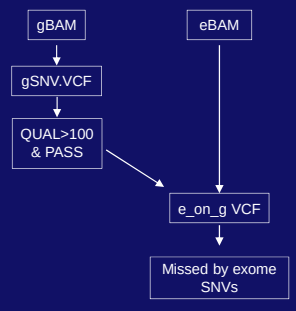
- Major issue: State of the art is continuously changing.
- Need to draw general conclusions!

Our approach

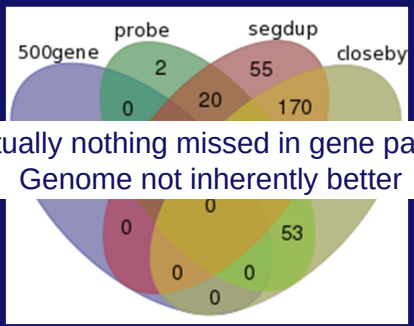
- Ask one simple question:
- Is the genome inherently better than the exome for simple variant calling?
  - Take DP/coverage out of the equation
  - Thus take capture kits/sequencing depth out of the equation

Methods exome versus genome

Yu Sun



A representative sample



Virtually nothing missed in gene panel: Genome not inherently better

Case study: NGS for ID

- What if your intention was to start diagnostic NGS for intellectual disability?
- Some of the questions that need answering:
  1. Is WGS inherently better than WES? → No
  2. Is a gene panel expected to perform better?
  3. How to determine sequencing criteria?

Is a gene panel expected to outperform WES?

- Depends on the panel, sequencing depth, etc
- How much room for improvement is there?
- Need to determine adequate coverage.
- But what is adequate coverage?

Some background: false negative rate versus DP

- Assume all qualities are perfect & p_{alt}=1/2
  - Reference read
  - Alternative read
- Sequencing results

●●●●●●●●●●

- Probability of observation if position is variant

½ * ½ * ½ * ½ * ½ * ½* ½ * ½ * ½ * ½ = 1/1024

$(½)^{30}$ = 1/1 billion

False negative rate versus DP

- Assume all qualities are perfect & p_{alt}=1/2
  - Reference read
  - Alternative read
- Sequencing results

●●●●●●●●●●

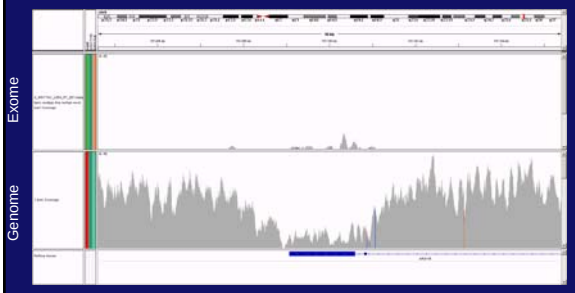
- Probability of observation if position is variant

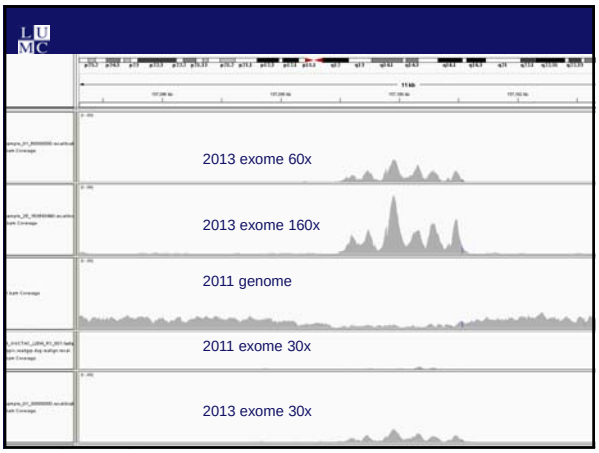
½ * ½ * ½ * ½ * ½ * ½* ½ * ½ * ½ * ½ * 10 = 1/102.4

How good is a 'state-of-the-art' exome?

- Sequence 13 exomes deeply (mean coverage 162x, range 145x-198x)
- Determine part of the exome with expected FN rate < 1/1000
- Mean coverage of the ID gene set:
  - 99.2 %, range 98.9-99.3 %.
- Thus very little room for improvement for a gene panel! (unless you are specifically interested in mosaics…)

Exome versus genome: exon 1





Case study: NGS for ID

- What if your intention was to start diagnostic NGS for intellectual disability?
- Some of the questions that need answering:
  1. Is WGS inherently better than WES? → No
  2. Is a gene panel expected to perform better? → no
  3. How to determine sequencing criteria?

Introduction of a new diagnostic test

- Main statistical criteria
  - False positive rate / FPR (1 – specificity)
  - False negative rate / FNR (1- sensitivity)
- HOWEVER
  - FPR is close to 0 if Sanger sequencing is performed
  - FNR might not be an appropriate measure for the case study
  - Perhaps 'yield' is more sensible?

Key outcome parameters:

- False positive rate
- False negative rate
- % not covered (Sanger seq?)
- Yield

All these relate to sequencing depth!

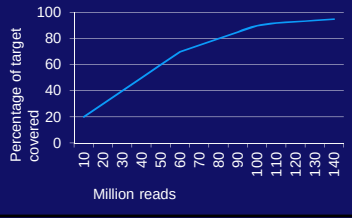
Yield *vs.* False-negatives

- No real core genes
- Happy with any catch
- E.g. NSID
- Current yield: low

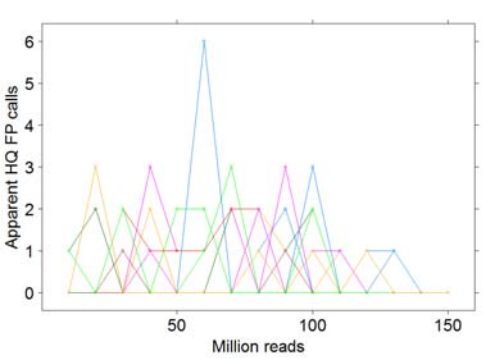
- Low number of core genes
- Do not want to miss anything!
- Breast cancer / classical CSS
- Acceptable FN rate ~ 0.1 %
- ~ 'Classical' diagnostics

Methods to investigate sequencing criteria

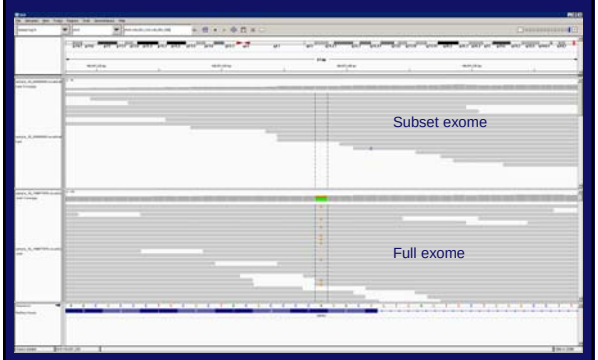
- Sequence 13 exomes to very high DP (~140X mean coverage)
- Assume these are 'as good as it gets'
- Investigate outcome parameters of subsets of the data versus the full sets



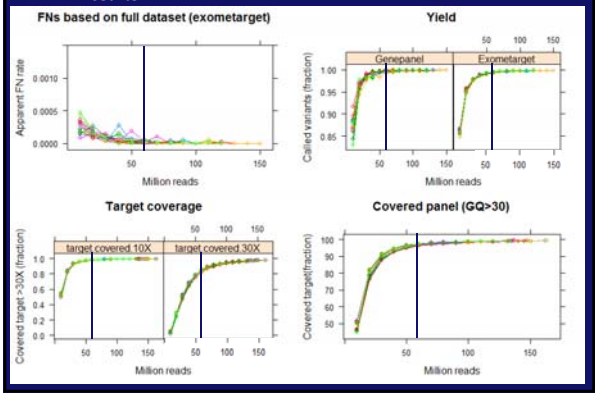
False positives: not problematic



3 variants missed in gene panel (out of 77k): bad luck



Results



Conclusions

Comparison of sequencing approaches needs to be fair: easier said than done!

Genome is not inherently better than the exome

State-of-the art exome does a good job of covering exonic regions (even GC-rich ones)

Subset-based approach can be useful to determine minimal sequencing thresholds

Acknowledgements

- Yu Sun
- Wibowo Arindrarto
- Martin van der Kerkhoff
- Emmelien Aten
- Leon Mei
- Jeroen Laros
- Marjolein Kriek
- Mariette Hoffer
- Claudia Ruivenkamp
- Bert Bakker
- Johan den Dunnen
- Christi van Asperen

Centre for Genome Diagnostics