LEIDEN UNIVERSITY MEDICAL CENTER

Effect prediction of phased variants

Martijn Vermaat, Michiel van Galen and Jeroen F.J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Next generation sequencing

Figure 1 : HiSeq 2000.

Introduction

Next generation sequencing

Figure 1 : HiSeq 2000.

Characteristics:

- Manufacturer: Illumina, Inc.
- Commercially available since 2010.
- Per cycle, one base is read.
- Reads up to 150×2 base pairs.
- Takes about 8 days.
- Produces about 150 Giga bases per run.

Introduction

Paired end data

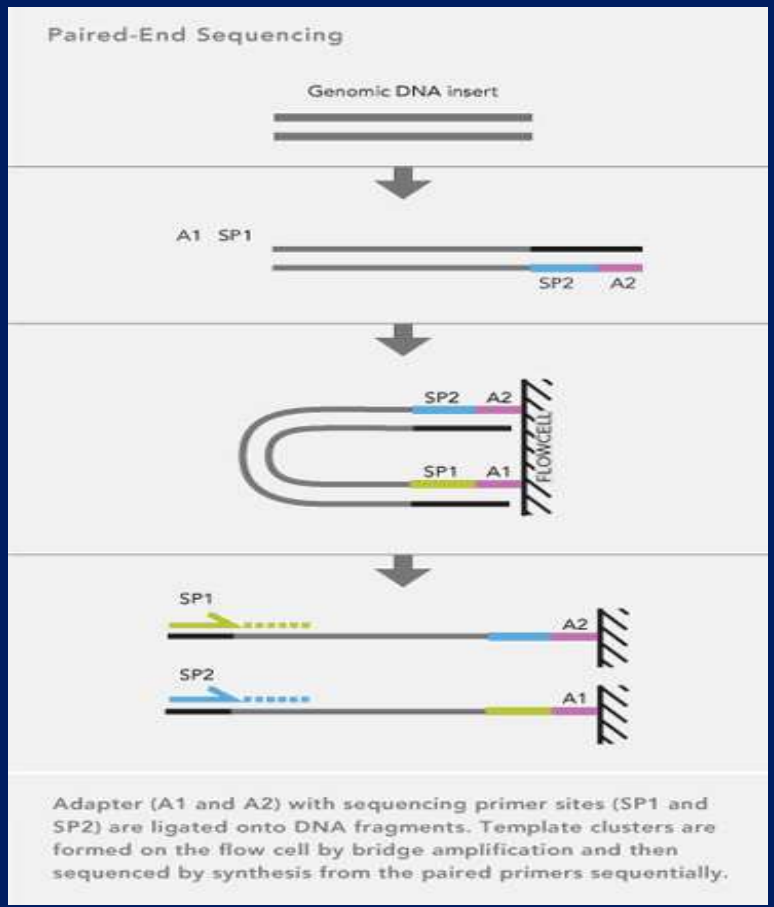

Figure 2 : PE sequencing.

Introduction

Paired end data

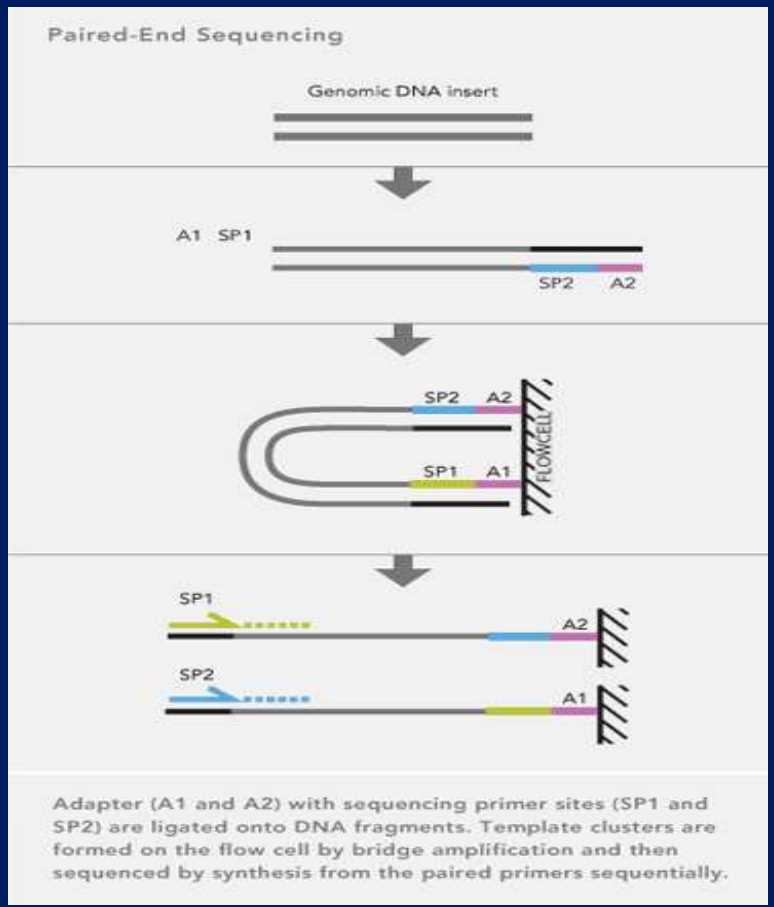

Figure 2 : PE sequencing.

Overview:

- Ligate adapters.
- Bridge amplification.
- Sequence first end.
- Cluster regeneration.
- Sequence second end.

Introduction

Variant calling

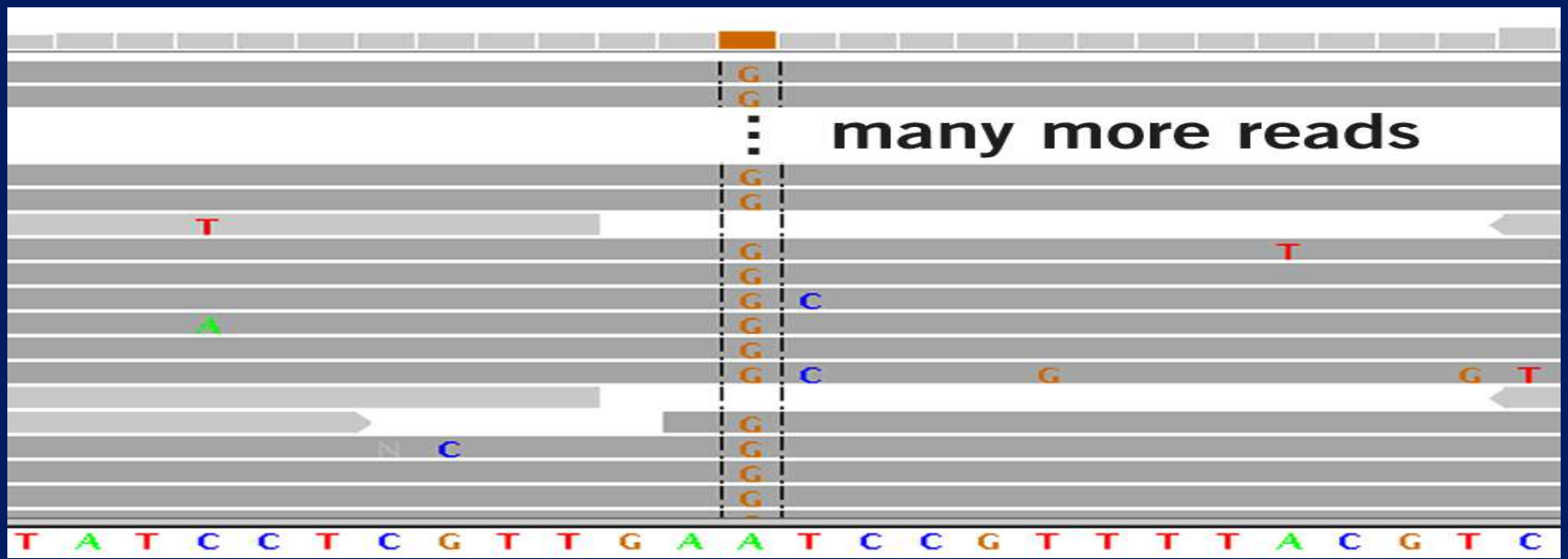

Figure 3 : Variant calling.

General idea:

- Align the reads to the reference genome.
- If we see a mismatch multiple times, it might be a variant.

Introduction

Effect prediction

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?

Introduction

Effect prediction

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
    - Is it in an intron?
        - Does it hit a splice site?

Introduction

Effect prediction

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - …

Introduction

Effect prediction

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - …
  - Is it in the 5'/3' UTR of a gene?
  - …

Introduction

Effect prediction

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - …
  - Is it in the 5'/3' UTR of a gene?
  - …
- Is it in a regulatory region?
- …

Limitations

Unphased variants

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Reference protein:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT LHGHWGLGQV VTDYVHGDAL
121  QKAAKAGLLA LSALTFAGLC YFNYHDVGIC KAVAMLWKL*
```

Protein predicted from variant coding sequence:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA RTIPWLQPSL FMVTGALDKL LLTMFMGMPC
121  RKLPRQGFWH FQL*
```

Figure 4 : Predicted frameshift.

NM_003002.2(SDHD_v001):c.272del

Limitations

Unphased variants

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Reference protein:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT LHGHWGLGQV VTDYVHGDAL
121  QKAAKAGLLA LSALTFAGLC YFNYHDVGIC KAVAMLWKL*
```

Protein predicted from variant coding sequence:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT SWSLGPWTSC Y*
```

Figure 5 : Predicted frameshift.

NM_003002.2(SDHD_v001):c.301_302del

Limitations

Phased variants

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Reference protein:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT LHGHWGLGQV VTDYVHGDAL
121  QKAAKAGLLA LSALTFAGLC YFNYHDVGIC KAVAMLWKL*
```

Protein predicted from variant coding sequence:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA RTIPWLQPSL HGHWGLGQVV TDYVHGDALQ
121  KAAKAGLLAL SALTFAGLCY FNYHDVGICK AVAMLWKL*
```

Figure 6 : Predicted indel.

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Project description

Phasing

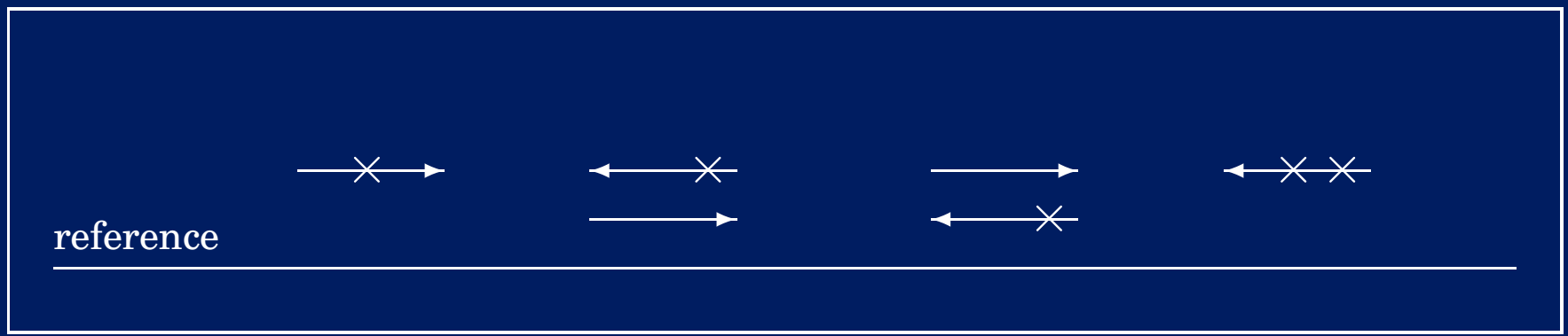

Figure 7 : Read backed phasing.

Direct inference of phased variants.

Project description

Layout

Proof of concept:

- Phasing variants.
- Conversion to HGVS.
- Lifting descriptions over to transcripts.
- Effect prediction using Mutalyzer.

Project description

Layout

Proof of concept:

- Phasing variants.
- Conversion to HGVS.
- Lifting descriptions over to transcripts.
- Effect prediction using Mutalyzer.

If there is time left:

- Phasing using imputation.

Questions?

Acknowledgements:

Michiel van Galen
Martijn Vermaat
Michel Villerius
Johan den Dunnen