1 - Introduction

Tuesday, 5 October 2021 09:24

MAIN IDEA:

DNA-SEQ IS GETTING SO FAST & CHEAP
THAT IT IS BEING USED ALL OVER
THE FIELDS OF LIFE-SCIENCES

- STUDYING DISEASES
- STUDYING CANCER
- KNOW MORE ABOUT OUR EVOLUTIONARY HISTORY
- MICROBES & ANTIMICROBIAL RESISTENCE
- METAGENOMICS
- HOW GENES WORK

SEQUENCING IS LIKE COMPUTING

NOW THAT THE TECH IS GOOD IT
IS USED EVERYWHERE

WHY COMPUTATIONAL GENOMICS?

WHAT IS POSSIBLE ?
WHAT IS PRACTICAL ?

EXAMPLE: "DE-NOVO SHOTGUN ASSEMBLY"

HARD PROBLEM BUT SOLVABLE W/ ENOUGH COMPUTING POWER

1ST-GEN SEQUENCING

SANGER-SEQ (CHAIN TERMINATION)

PRINCIPAL METHOD DURING HUMAN-GENOME PROJECT

2ND-GEN SEQUENCING (NGS)

2007 GENOME COST REVOLUTION

MASSIVELY-PARALLEL SEQUENCING

BILLIONS OF DNA MOLECULE IN PARALLEL

IMPROVEMENTS
- COST
- SPEED
- ACCURACY
- USABILITY

MOST LAB HAVE 2ND-GEN RIGHT NOW

GENOMES AS STRINGS

PROFOUND IMPLICATIONS ON HOW WE CAN ANALYZE IT

DNA-SEQUENCER ARE NOT VERY GOOD AT
READING LONG STRETCHES OF DNA

BEST: LOTS OF SHORT-STRETCHES

'SNIPPETS' ARE OFTEN CALLED 'READS'

USUALLY WE HAVE ENOUGH READS TO COVER
THE WHOLE GENOMES MANY TIMES OVER

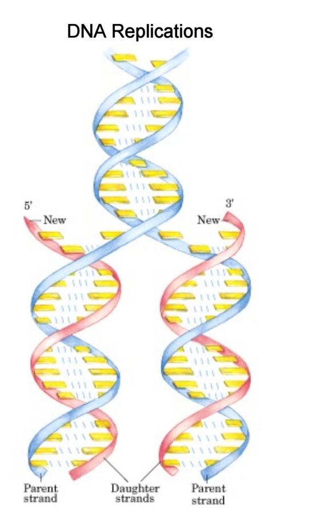
STRINGS

Σ	ALPHABET
$S(\Sigma) = S$	STRING TEXT
$\|S\|$	STRING TEXT
ε	EMPTY STRING ($\|\varepsilon\| = 0$)
$S[x]$	CHARACTER AT POSITION X (0-INDEXING)
$S + t$	CONCATENATION
$S[x:y]$	SUBSTRING $[x, y)$ Y NOT INCLUDED
$S[:x]$	PREFIX
$S[-y:]$	SUFFIX $S[x : \|S\|]$

SEQUENCING BY SYNTHESIS

HOW DNA GETS COPIED

DNA-POLYMERASE



2ND-GEN SEQUENCE DNA BY
EAVESDROPPING ON THIS PROCESS

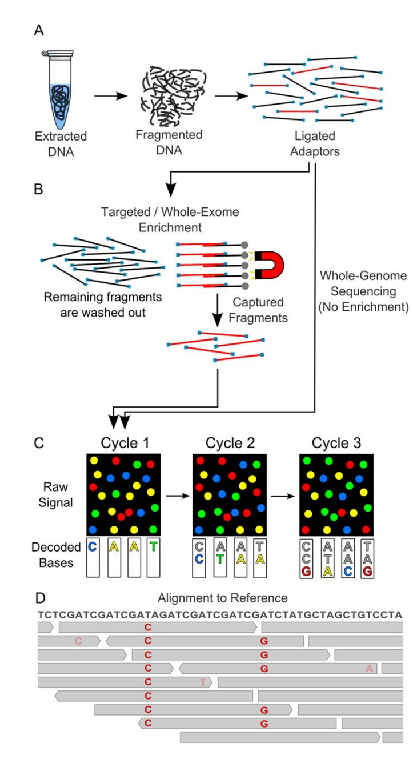
IF WE WATCH WHAT THE POLYMERASE
IS DOING AS IT BUILDS THE COMPLEMENTARY
STRAND, WE CAN INFER THE SEQUENCE
OF THE TEMPLATE MOLECULE

HOW 2-ND GEN SEQUENCING WORKS

1) INPUT DNA EXTRACTED
2) DNA CUT INTO SNIPPETS
SHORT SINGLE-STRANDED DNA TEMPLATES
3) TEMPLATES DEPOSITED ON SLIDE SURFACE
4) ADD DNA POLYMERASE + BASES
BASES ARE TERMINATED, POLI ADDS 1 AND STOPS
5) WE TAKE A PHOTO (FLUORESCENT TERMINATORS)

WE KNOW KNOW WHICH BASE WAS
ADDED TO THE TEMPLATE STRAND

6) REMOVE TERMINATORS AND REPEAT
ONE PHOTO PER SEQUENCING CYCLE



SEQUENCING ERRORS & BASE-QUALITY

WE FIST AMPLIFY TEMPLATE STRANDS
TO HAVE ENOUGH FLUORESCENCE

UNTERMINATED BASE ?
ONE OF THE TEMPLATE IS OUT OF SYNC

AS WE MOVE FROM 1 SEQ CYCLE TO THE NEXT ←
THE # OF OUT-OF-SYNC STRANDS WILL GROW

BASE-CALLER SOFTWARE

HAS TO DEAL WITH AMBIGUITY
WE WANT TO KNOW WHICH BASES ARE UNSURE

FOR EACH BASE THE BC REPORTS ITS 'QUALITY'

$$Q = -10 \log(P) \rightarrow P = 10^{-Q/10}$$

P = PROBABILITY THAT THE
BC IS INCORRECT

$Q = 10$	$1/10$
$Q = 20$	$1/100$
$Q = 30$	$1/1000$

HIGHER Q IS BETTER

SEQUENCING READS IN FASTQ

FASTA (2 LINES)

- HEADER
- SEQUENCE

FASTQ (4 LINES)

- HEADER
- SEQUENCE
- PLACEHOLDER
- QUALITY SCORES

QUALITY SCORES

EACH BASE QUALITY IS AN ASCII ENCODED
VERSION OF $Q = -10 \log P$

PHRED-33

ROUND Q TO NEAREST INT, ADD + 33
CONVERT INTO CORRESPONDING CHARACTER

PHRED33ToQ (x)
QToPHRED33 (x)

MOST # AT THE END
AS YOU STEP TOWARDS THE END YOUR
CONFIDENCE GETS LOWER

PHRED33ToQ (#) = 2 → 63 % ERROR
PHRED33ToQ (J) = 41 → $7{,}94 \cdot 10^{-5}$ % ERROR

GENOMIC PUZZLES

HOW IS MY GENOME DIFFERENT FROM YOURS ?
WHAT DOES IT TELL ME ABOUT MY PREDISPOSITION TO DISEASES ?

WE CANNOT ANSWER QUESTION LIKE THESE USING READS
WE NEED TO ASSEMBLE THE READS

ON AVERAGE 1 GENE IS 10^3 bp LONG
ONE READ CANNOT EVEN COVER A GENE

WE WANT TO INFER THE SEQUENCE INPUT

HUMANS ARE 99.8 % SIMILAR (JUST ANALYZING DNA)

WE CAN USE A REFERENCE GENOME AS A GUIDE

THE READ-ALLIGNMENT PROBLEM

WE ARE TAKING ADVANTAGE OF THE FACT THAT
DIFFERENT INDIVIDUALS OF THE SAME SPECIES
HAVE SIMILAR GENOME SEQUENCES

WE ARE LOOKING FOR THE PLACE IN THE REF
GENOME WHERE THE READ MATCHES MORE CLOSELY

WE ARE DOING IT FOR EVERY SEQ-READ

NAIVE EXACT MATCH PROBLEM (SIMPLIFIED VERSION)

AT WHAT OFFSET DOES PATTERN P
OCCUR WITHIN TEXT T ?

THIS IS A GREAT PLACE TO START
BUT WE NEED BETTER ALGORITHMS

FASTER & APPROXIMATE