Comprehensive Genome and Transcriptome Structural Analysis of a Breast Cancer Cell Line using PacBio Long Read Read Sequencing

Maria Nattestad

Schatz + McCombie + Hicks at Cold Spring Harbor Laboratory
McPherson + Beck at the Ontario Institute for Cancer Research
Pacific Biosciences
DNAnexus
SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)
PacBio long-read DNA sequencing

mean read length: 9 kb
max read length: 71 kb

72X coverage

Genome-wide coverage averages around 54X
Coverage per chromosome varies greatly as expected from previous karyotyping results
Structural variant detection

**Assembly-based**
- Assembly with Falcon on DNAnexus
  - Alignment with MUMmer
    - Call variants between consecutive alignments with ABVC
    - Call variants within alignments with ABVC

**Alignment-based**
- Alignment with BWA-MEM
  - Copy number analysis
  - SV-calling from split reads with Sniffles
    - Validations
    - SplitThreader
      - Detailed analysis of Her2 amplifications

~ 11,000 local variants
  50 bp < size < 10 kbp

350 long-range variants
  (>10kb distance)
Long-range structural variants found by Sniffles (> 10kb)

Threshold: 10 split reads broken within a 200 bp interval on both sides of the translocation
Her2

Chr 17: 83 Mb

8 Mb
SplitThreader:
Graphical threading to retrace complex history of rearrangements in cancer genomes
1. Healthy chromosome 17
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8
Transcriptome analysis with IsoSeq

IsoSeq
long-read RNA-seq

Gene fusions

DNA + RNA evidence:
• 13 fusions in previous literature
• 4 novel fusions
  • CYTH1-MTBP
  • SAMD12-EXT1
  • PHF20-PR4-723E3.1
  • AMZ2-CASC8

RNA evidence only:
• 188 fusions

Novel isoforms

~ 45,000 novel isoforms (2+ reads)
~ 7,400 with 10+ reads

279 putative novel genes
• 10+ reads of the same isoform
• Not overlapping existing annotation
CYTH1-EIF3H gene fusion

Chr 17

CYTH1

30 IsoSeq reads

3.7 Mb

MTBP

8 kb

Chr 8

EIF3H

27

36
The genome informs the transcriptome

Data and additional results: http://schatzlab.cshl.edu/data/skbr3/
Acknowledgments

Sara Goodwin
Timour Baslan
Fritz Sedlazeck
Tyler Garvin
Han Fang
James Gurtowski
Elizabeth Hutton
Marley Alford
Melissa Kramer
Eric Antoniou
James Hicks
Michael Schatz
W. Richard McCombie

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

DNAnexus

Pacific Biosciences
Elizabeth Tseng
Jason Chin

Our software in development: Sniffles by Fritz, ABVC and SplitThreadder by Maria