De Novo Transcriptome Assembly

Dr. Jung Soh

657.001 Transcriptomics (WS 2018/2019)

What is transcriptomics?

- Transcriptomics is

  the study of the complete set of RNA transcripts produced by a given cell, organ, or living organism.

- The main focus of transcriptomics is

  to evaluate differential expression of genes between conditions. High-throughput technologies such as microarray and RNA-sequencing have become the standard for such experiments.

- Transcriptomics software tools are also used

  to study RNA modifications, RNA-protein interactions, non-coding RNA and RNA structure.

https://omictools.com/transcriptomics-category

What is RNA-seq?

- An experimental protocol

  that uses next-generation sequencing technologies

  to sequence the RNA molecules within a biological sample

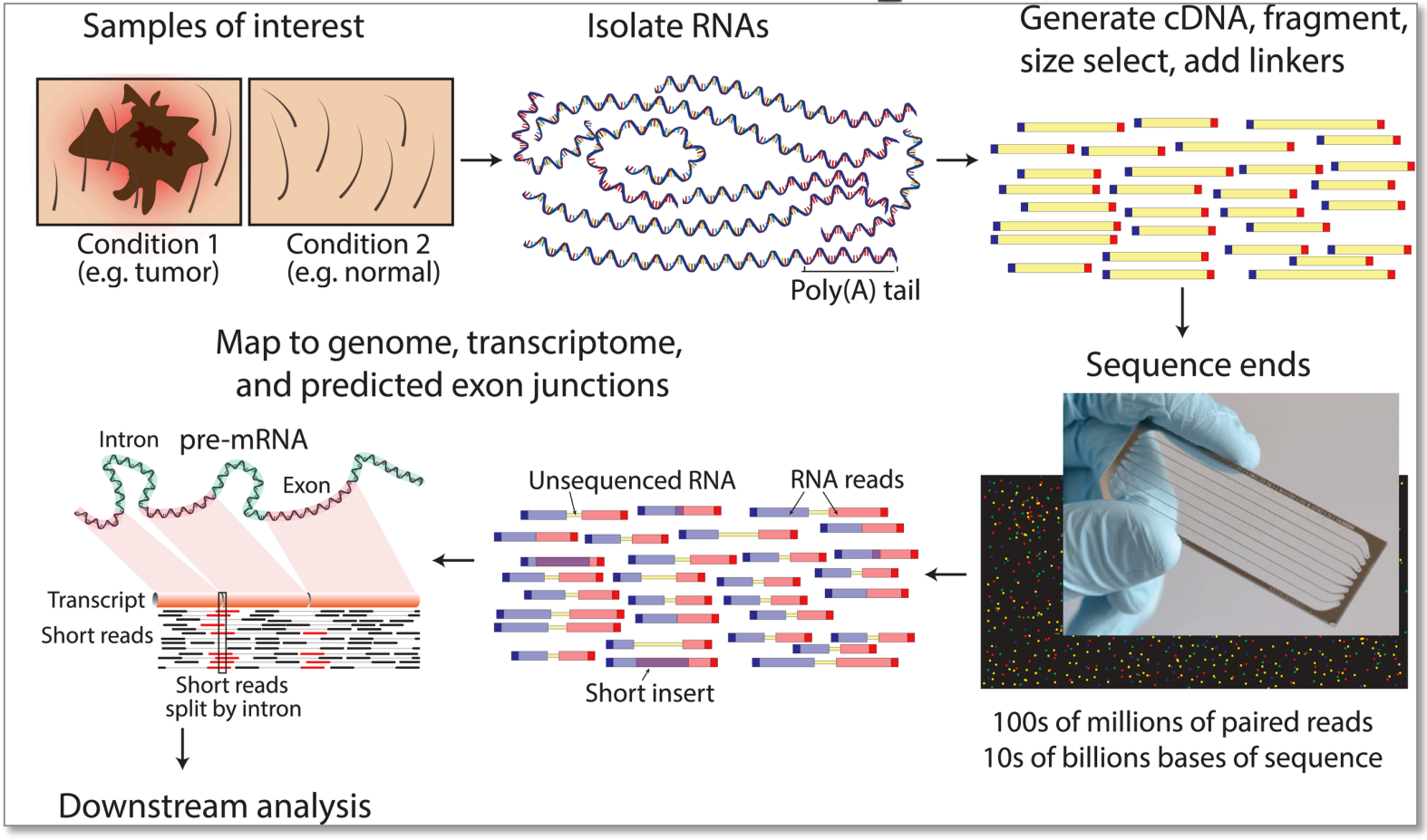
  in an effort to determine the primary sequence and relative abundance of each RNA

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet. 12(10):671-682

Why RNA-seq?

- To study functions based on gene expression changes
  - Drug treatment vs. no treatment
  - Patients vs. healthy people
  - Wild type vs. knock-out
  - Expressions in different tissues
- Some features available only at RNA level
  - Alternative isoforms, RNA editing, transcript fusion

Overview of RNA-seq

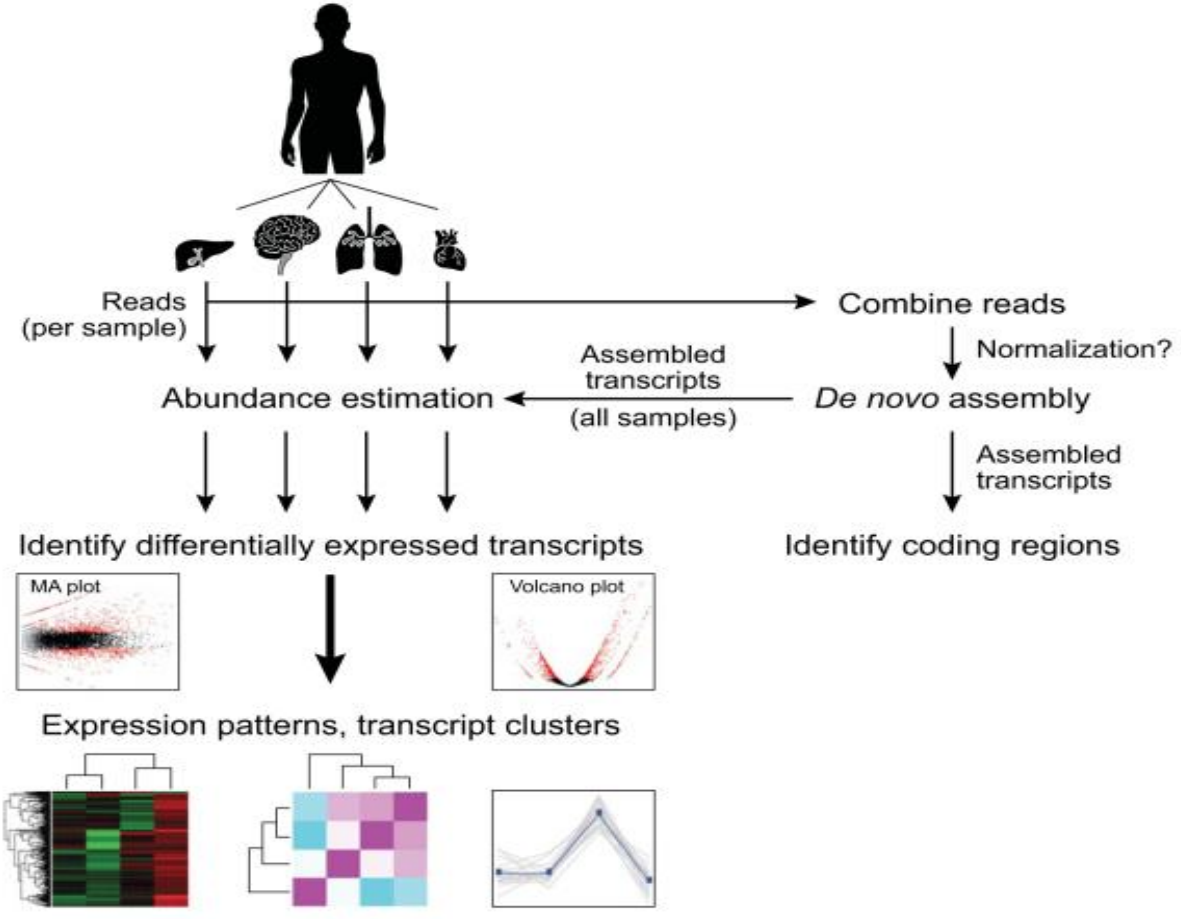


https://commons.wikimedia.org/wiki/File:Journal.pcbi.1004393.g002.png

RNA-seq challenges

- Difficulty with sampling
  - More fragile than DNA
  - Different sizes of RNA
- Relative abundance of RNAs hard to control
  - Can vary by orders of magnitude
  - Uneven coverage
    - Many reads from a small number of highly expressed genes
- Computational analysis challenges
  - Assembly, (spliced) alignment, quantification, normalization, visualization, …

De novo transcriptome assembly and analysis



Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols*, *8*(8)

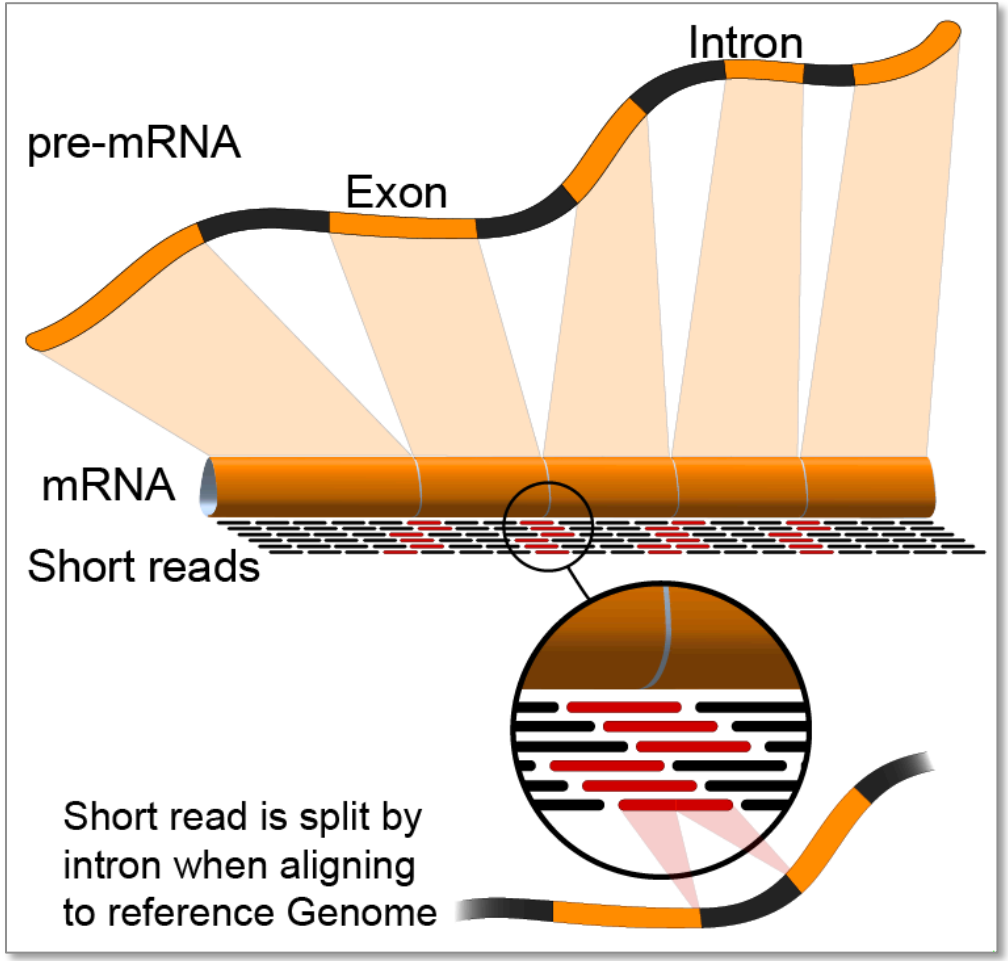
Major transcriptomic analysis tasks

- *De novo* transcriptome assembly
- Functional annotation of transcriptome
- Differential expression analysis
- Novel transcript discovery

Read alignment

- Essential method for
  - Transcript abundance estimation
  - De novo assembly quality assessment
- Can use reference genome or *de novo* assembled transcriptome
- Issues with aligning to reference genome
  - Reads spanning across exon junction
  - Alternative splicing
- Most common alignment results format
  - SAM: sequence alignment/map (or BAM: binary version)
- Many alignment tools
  - Bowtie 2, BWA, SOAP, SHRiMP, mrFAST, mrsFAST, ZOOM, SSAHA2, …

Aligning RNA-seq reads to genome



https://commons.wikimedia.org/wiki/File:RNA-Seq-alignment.png

SAM/BAM format

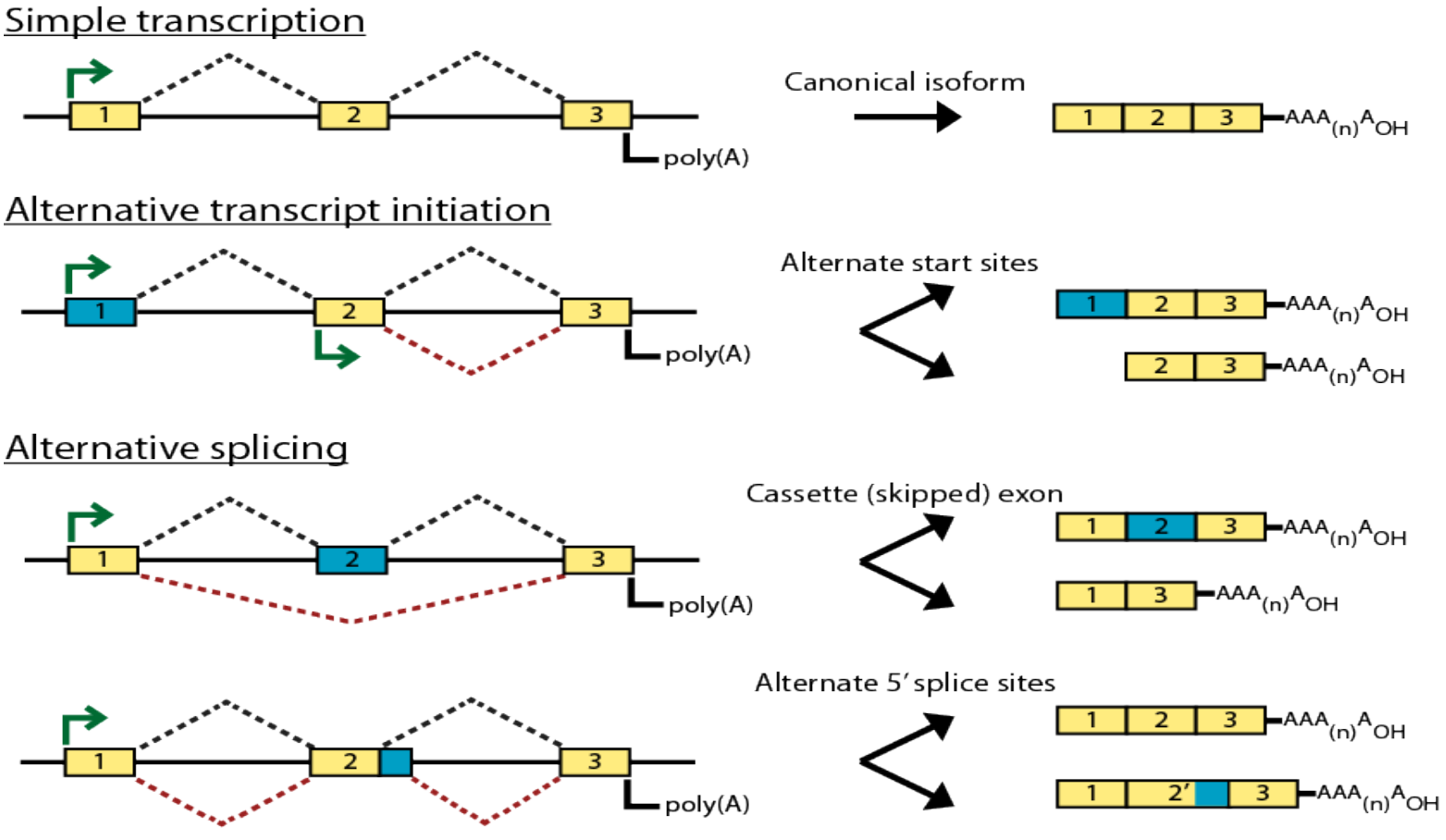
- Header section

```
mgriffit@linus270 ~> samtools view -H /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam  | grep -P "SN\:22|HD|RG|PG"
@HD     VN:1.4  SO:coordinate
@SQ     SN:22   LN:51304566     UR:ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa.gz AS:GRCh37-lite  M5:a718acaa6135fdca8357d5bfe94211dd  SP:Homo sapiens
@RG     ID:2888721359   PL:illumina     PU:D1BA4ACXX.3  LB:H_KA-452198-0817007-cDNA-3-lib1       PI:365  DS:paired end   DT:2012-10-03T19:00:00-0500     SM:H_KA-452198-0817007  CN:WUGSC
@PG     ID:2888721359   VN:2.0.8        CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0
@PG     ID:MarkDuplicates       PN:MarkDuplicates       PP:2888721359   VN:1.85(exported)       CL:net.sf.picard.sam.MarkDuplicates INPUT=[/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300.bam] OUTPUT=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300-post_dup.bam METRICS_FILE=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/staging-1iuJS/H_KA-452198-0817007-cDNA-3-lib1-2888360300.metrics REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=9500 TMP_DIR=[/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y] VALIDATION_STRINGENCY=SILENT MAX_RECORDS_IN_RAM=500000    PROGRAM_RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=MarkDuplicates MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 SORTING_COLLECTION_SIZE_RATIO=0.25 READ_NAME_REGEX=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MD5_FILE=false
mgriffit@linus270 ~>
```

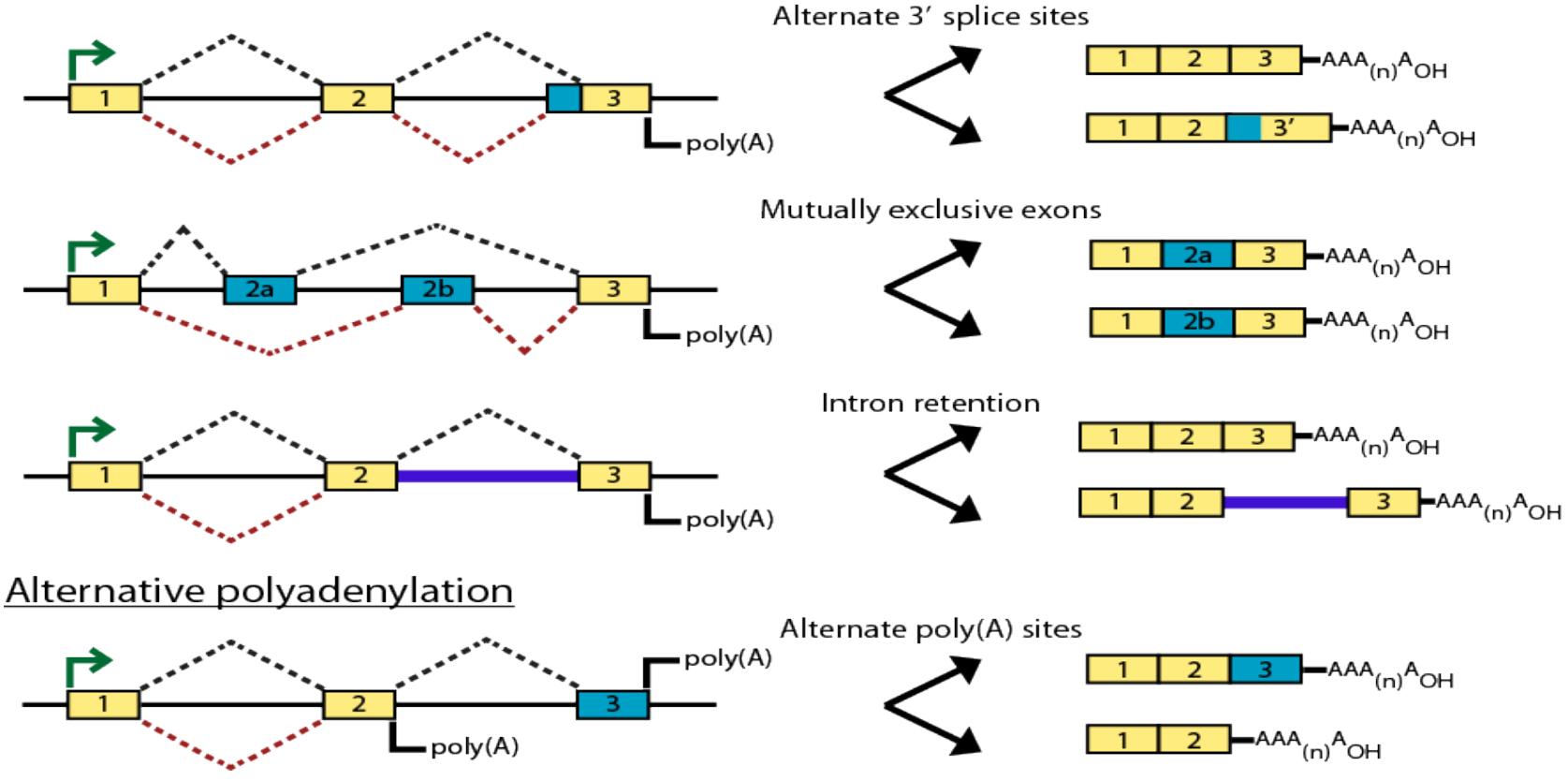
- Alignment section (10 alignments shown)

```
mgriffit@linus270 ~> samtools view -f 3 -F 1804 /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | head
HWI-ST495_129147882:3:2114:15769:38646  99      1       11306   3       100M    =       11508   302     ACTGCGGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGC  CCFFFFFHHHGHJJJJJJIJJJHGIJJIJJHIIJJJJJJHFDDDDDDDDDDDDCDDDDDDDDDDDDDD<CDDDDDDDDCCCC>@>CDDDDDDD?1<B      CC:Z:15 MD:Z:5A94       PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:1  XN:i:0  XO:i:0  CP:i:102519765  AS:i:-5 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2114:15769:38646  147     1       11508   3       100M    =       11306   -302    ACTCCTAAATATGGGATTCCTGGGTTTAAAAGTATAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAG5  ;5:CDCDCDECEFCD@9E=?7EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJIJGHGEIJJIJJJJJJJIHHCIEJJJHFHHGHFFEDFCCB      CC:Z:15 MD:Z:34A65      PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:1  XN:i:0  XO:i:0  CP:i:102519563  AS:i:-6 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   163     1       11810   3       100M    =       12055   345     CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGC  CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJIJGHHHGHFFFCDDDDDDCDCCCCCA;>@>@AA@:AA>AA      CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:0  XN:i:0  XO:i:0  CP:i:102519261  AS:i:0  XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   83      1       12055   3       100M    =       11810   -345    GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC  CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIIHHEHIIGJIJIJJIGHIIIJJJJJIIJJJJIIJJIJIJJHGHHHDFEFFCCC      CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:0  XN:i:0  XO:i:0  CP:i:102519016  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   163     1       12634   3       100M    =       12746   212     GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCACAGGTGAGAGGAGAG<  @@FFFFFDHHHH9FHGIIFGAFDHEGIJ>GHIIIIIIIIIIIIIIIIFHDDFFEEEECEECCCACCCCCC:AADCCBCC>CAC<CCCCCC:@CB@@BAB##      CC:Z:15 MD:Z:85G14       PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:1  XN:i:0  XO:i:0  CP:i:102518437  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   83      1       12746   3       100M    =       12634   -212    GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD  DCABDBDDDDDDDDDDDCDDDDBDB@BDDDB@;CCCCCDEFD@@;.?<HIGGEIGEHIGJJJIIGIGIIHEGFEHFJIIIIIGJJJJHHHHHFFFFFC@@      CC:Z:15 MD:Z:37G62       PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:1  XN:i:0  XO:i:0  CP:i:102518325  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   99      1       13503   3       100M    =       13779   376     CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC  CCFFFFFHHHHHJJJIJJJJJJJJJJJJGIIIIJJFHGGIJGIJJJEGIJIJJHHIHHGHFFEFDEEEECCCAACDDACDCDDDDDB?8?<B>A@CDC      CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:0  XN:i:0  XO:i:0  CP:i:114357414  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1309:15328:74082  99      1       13534   3       100M    =       13780   346     AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGG@  CCFFFADHHHHFIJJJJJIJIIHIJJIHJJIJJIIJJIJJJJJJBFHIIJJJJJIJHH=EEFFFFCEEECEDCDCDDDDDDDDDDDDDDDBDCCD      CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:0  XN:i:0  XO:i:0  CP:i:114357383  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1308:10126:19636  99      1       13779   3       100M    =       14027   348     CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGC  CCFFFFFHHGHHJJJJJJJJJIJJJJJJIJJJJJJJJJIJIJJJJJJJJJHJJJHGHHFFFFFFEEEEEDDDDCDCDCDFADCACCACDDCCCCCCD      CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:0  XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   147     1       13779   3       100M    =       13503   -376    CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTG#  ##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIIJHJJIJJIGCIIJJJJJIGHGJJIJIJJJJJIJIIIGGFGHHHHFFFFFCCC      CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:0  XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
mgriffit@linus270 ~>
```

Types of alternative expression



Types of alternative expression



https://bioinformatics.ca/rnaseq2016module-5-ppt

De novo transciptome assembly

- No reference genome, use only RNA-seq reads
- Required for most non-model organisms
- Similar to de novo genome assembly
  - RNA reads tend to be shorter
  - Uneven coverage depth
- Many assemblers
  - Trinity, rnaSPAdes, SOAPdenovo-trans,Velvet/Oases, …

Genome assembly metaphor

DNA clones

Reads

Reconstructed genome

The Call—Chronicle—Examiner

SAN FRANCISCO, THURSDAY, APRIL 19, 1906.

EARTHQUAKE AND FIRE: SAN FRANCISCO IN RUINS

NO HOPE LEFT FOR SAFETY OF ANY BUILDINGS

BLOW BUILDINGS UP TO CHECK FLAMES

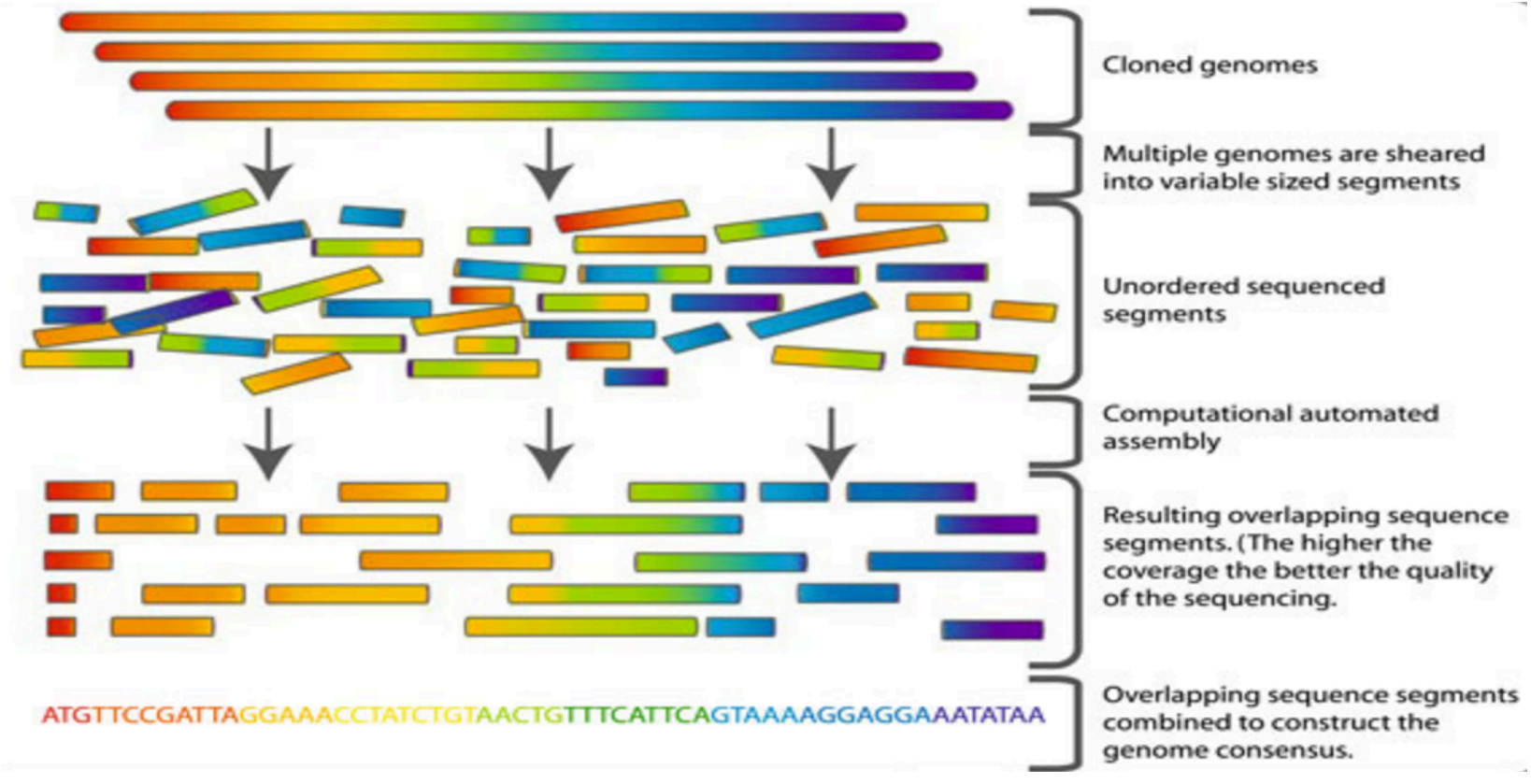
WHOLE CITY IS ABLAZE

CHURCH OF SAINT IGNATIUS IS DESTROYED

MAYOR CONFERS WITH MILITARY AND CITIZENS

http://www.vicbioinformatics.com/documents/Genome%20Assembly%20Strategies%20-%20Torsten%20Seemann%20-%20IMB%20-%205%20Jul%202010.pdf

Genome assembly: another view



Commins J, Toft C, Fares MA. (2009) Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biol Proced Online. 11:52-78.

Lab overview