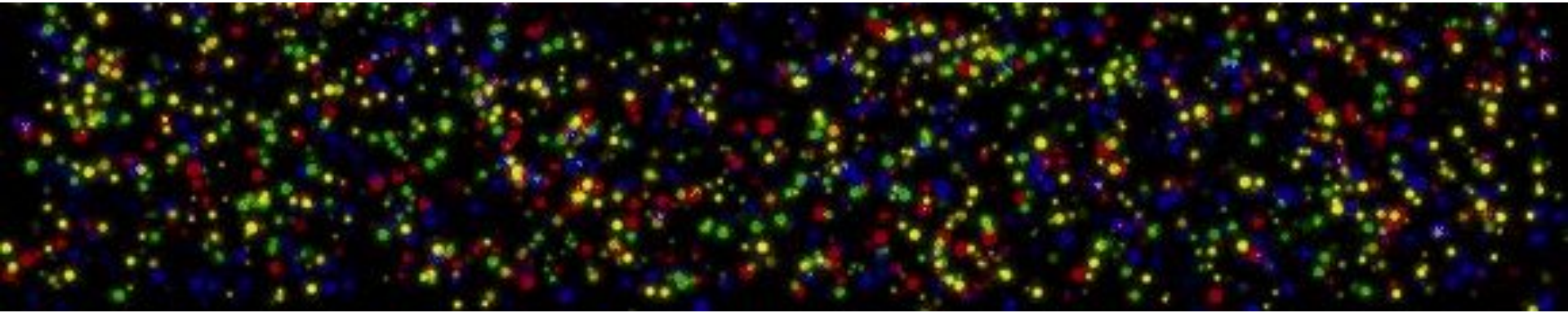
RNA-seq for Plant Science



Jędrzej Jakub Szymański

@NAMlab www.szymanskilab.com

IPK LEIBNIZ INSTITUTE

CEPLAS Cluster of Excellence on Plant Sciences

JÜLICH Forschungszentrum

The birthplace of high-throughput sequencing (HTS)

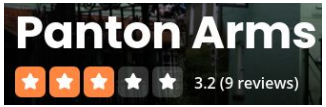
The birthplace of high-throughput sequencing (HTS)

David Klenerman

Shankar Balasubramanian

The birthplace of high-throughput sequencing (HTS)

David Klenerman



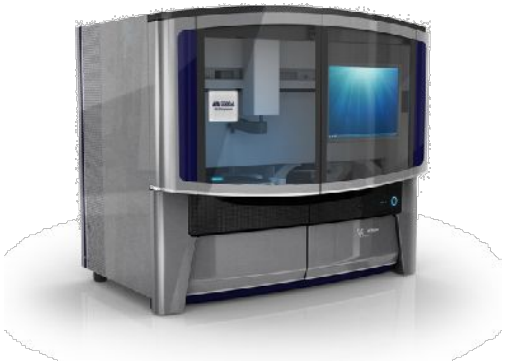
Shankar Balasubramanian

Sequencing circa 2012

Roche 454 ‘pyrosequencing’

Life Technologies
Ion Torrent

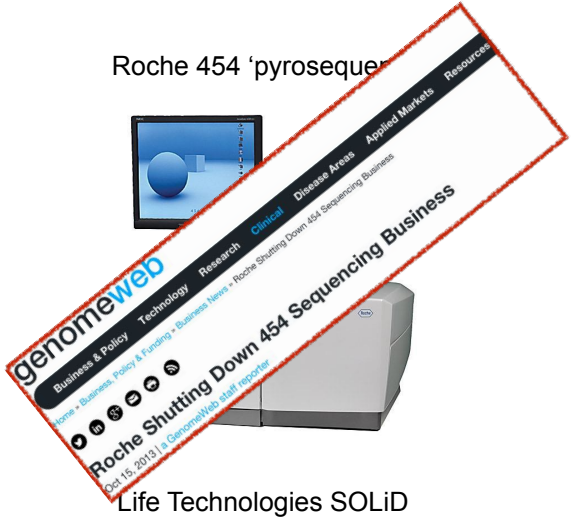
Life Technologies SOLiD



Illumina

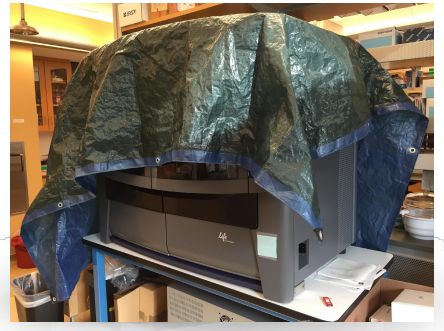
Sequencing circa 2012

Roche 454 ‘pyrosequer



Life Technologies
Ion Torrent

Life Technologies SOLiD



Illumina

The year of sequencing

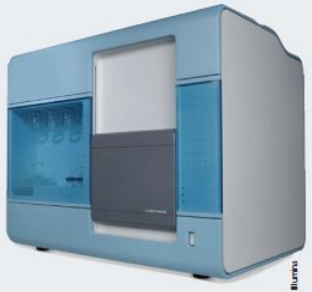
In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

CREATING THE GENOME ANALYZER

When John West started as CEO of Solexa Ltd. in August of 2004, the longest stretch of DNA that the company could sequence was only six bases long. "That was a little bit intimidating," recalls West, now the vice president and general manager of Illumina's DNA sequencing business unit following the acquisition of Solexa Ltd. by Illumina. "The problem was we never had a commercial platform to be able to sequence it on."

The next-generation sequencing platform they developed and commercialized by June 2006, West says, has been a huge success. Since the commercial release of the platform they have sold 100 instruments and increased the scale of what they have tackled using the technology—from a 5,300-base-pair viral genome to a 150-million-base-pair human X chromosome. But the machine was a challenge to develop. The developers had to bring together key elements of chemistry, people and technology to make it work.

By the time Solexa Ltd. announced its plans to merge with Lynx Therapeutics in August 2004, a lot of the core sequencing-by-synthesis chemistry and molecular biology had already been done, West says. The early chemistry work, done by Solexa Ltd. in the United Kingdom, led to the creation of a reversible terminator nucleotide and a polymerase that would incorporate it. Solexa Ltd. and Lynx Therapeutics had bought cluster technology for solid phase amplification together from the Swiss company Manteia SA, and did some instrumentation design.



Illumina's (Solexa) Genome Analyzer.

Still, the company was in a state of flux in late 2004 and needed to figure out how to combine the chemistry and technology into a complete system. The researchers needed to meet and brainstorm, operating on an eight-hour time difference between the two branches, one in the UK and one in the US. Ligation chemistry developed by Lynx Therapeutics was an option. There were differing opinions about which chemistry, ligation or polymerase would work best, but ultimately they took a gamble on the Solexa polymerase because it could complete reactions faster. "It was risky at the time because it wasn't all proven," West remembers. "We basically put all our eggs in one basket."

And it worked. It became easier, with the cluster technology and polymerase chemistry, to increase the sequence lengths. By February 2005, they increased the read length to 25 bases. They sequenced the 5,300-base-pair virus PHIX174 genome, the same genome that Sanger first sequenced. In October 2005, they sequenced an 180,000-base-pair bacterial artificial chromosome.

Illumina acquired Solexa, Inc. in January of last year, and commercial sales increased, West says. In 2007, the company completed sequencing the human X chromosome and has since moved on to sequencing the human genome using paired-end sequencing. "I think [2007] has been a great year for us," West says.

KRC

The year of sequencing

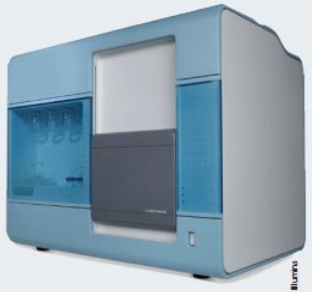
In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

CREATING THE GENOME ANALYZER

When John West started as CEO of Solexa Ltd. in August of 2004, the longest stretch of DNA that the company could sequence was only six bases long. "That was a little bit intimidating," recalls West, now the vice president and general manager of Illumina's DNA sequencing business unit following the acquisition of Solexa Ltd. by Illumina. "The problem was we never had a commercial platform to be able to sequence it on."

The next-generation sequencing platform they developed and commercialized by June 2006, West says, has been a huge success. Since the commercial release of the platform they have sold 100 instruments and increased the scale of what they have tackled using the technology—from a 5,300-base-pair viral genome to a 150-million-base-pair human X chromosome. But the machine was a challenge to develop. The developers had to bring together key elements of chemistry, people and technology to make it work.

By the time Solexa Ltd. announced its plans to merge with Lynx Therapeutics in August 2004, a lot of the core sequencing-by-synthesis chemistry and molecular biology had already been done, West says. The early chemistry work, done by Solexa Ltd. in the United Kingdom, led to the creation of a reversible terminator nucleotide and a polymerase that would incorporate it. Solexa Ltd. and Lynx Therapeutics had bought cluster technology for solid phase amplification together from the Swiss company Manteia SA, and did some instrumentation design.



Illumina's (Solexa) Genome Analyzer.

Still, the company was in a state of flux in late 2004 and needed to figure out how to combine the chemistry and technology into a complete system. The researchers needed to meet and brainstorm, operating on an eight-hour time difference between the two branches, one in the UK and one in the US. Ligation chemistry developed by Lynx Therapeutics was an option. There were differing opinions about which chemistry, ligation or polymerase would work best, but ultimately they took a gamble on the Solexa polymerase because it could complete reactions faster. "It was risky at the time because it wasn't all proven," West remembers. "We basically put all our eggs in one basket."

And it worked. It became easier, with the cluster technology and polymerase chemistry, to increase the sequence lengths. By February 2005, they increased the read length to 25 bases. They sequenced the 5,300-base-pair virus PHIX174 genome, the same genome that Sanger first sequenced. In October 2005, they sequenced an 180,000-base-pair bacterial artificial chromosome.

Illumina acquired Solexa, Inc. in January of last year, and commercial sales increased, West says. In 2007, the company completed sequencing the human X chromosome and has since moved on to sequencing the human genome using paired-end sequencing. "I think [2007] has been a great year for us," West says. *KRC*

The year of sequencing

In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

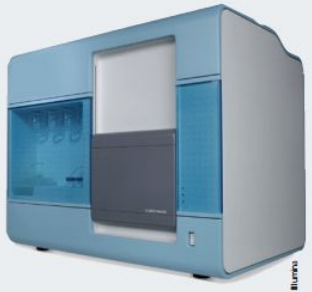
CREATING THE GENOME ANALYZER

When John West started as CEO of Solexa Ltd. in August of 2004, the longest stretch of DNA that the company could sequence was only six bases long. "That was a little bit intimidating," recalls West, now the vice president and general manager of Illumina's DNA sequencing business unit following the acquisition of Solexa Ltd. by Illumina. "The problem was we never had a commercial platform to be able to sequence it on."

The next-generation sequencing platform they developed and commercialized by June 2006, West says, has been a huge success. Since the commercial release of the platform they have sold 100 instruments and increased the scale of what they have tackled using the technology—from a 5,300-base-pair viral genome to a 150-million-base-pair human X chromosome. But the machine was a challenge to develop. The developers had to bring together key elements of chemistry, people and technology to make it work.

By the time Solexa Ltd. announced its plans to merge with Lynx Therapeutics in August 2004, a lot of the core sequencing-by-synthesis chemistry and molecular biology had already been done, West says. The early chemistry work, done by Solexa Ltd. in the United Kingdom, led to the creation of a reversible terminator nucleotide and a polymerase that would incorporate it. Solexa Ltd. and Lynx Therapeutics had bought cluster technology for solid phase amplification together from the Swiss company Manteia SA, and did some instrumentation design.

Still, the company was in a state of flux in late 2004 and needed to figure out how to combine the chemistry and technology into a complete system. The researchers needed to meet and brainstorm, operating on an eight-hour time difference between the two branches, one in the UK and one in the US. Ligation chemistry developed by Lynx Therapeutics was an option. There were differing opinions about which chemistry, ligation or polymerase would work best, but ultimately they took a gamble on the Solexa polymerase because it could complete reactions faster. "It was risky at the time because it wasn't all proven," West remembers. "We basically put all our eggs in one basket."



Illumina's (Solexa) Genome Analyzer.

And it worked. It became easier, with the cluster technology and polymerase chemistry, to increase the sequence lengths. By February 2005, they increased the read length to 25 bases. They sequenced the 5,300-base-pair virus PHIX174 genome, the same genome that Sanger first sequenced. In October 2005, they sequenced an 180,000-base-pair bacterial artificial chromosome.

Illumina acquired Solexa, Inc. in January of last year, and commercial sales increased, West says. In 2007, the company completed sequencing the human X chromosome and has since moved on to sequencing the human genome using paired-end sequencing. "I think [2007] has been a great year for us," West says. *KRC*

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi,[1]* Zhong Wang,[1]* Karl Waern,[1] Chong Shou,[2] Debasish Raha,[1] Mark Gerstein,[2,3] Michael Snyder[1,2,3]†

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3′-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

6 JUNE 2008 VOL 320 **SCIENCE**

Illumina sequencing

Illumina sequencing

MiniSeq System | MiSeq Series | NextSeq Series | HiSeq Series | HiSeq X Series | NovaSeq Series

Data output

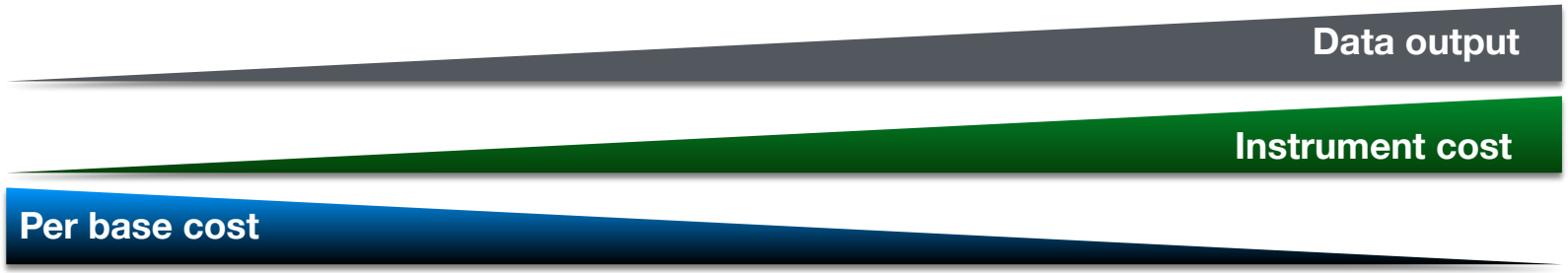
Illumina sequencing

MiniSeq System | MiSeq Series | NextSeq Series | HiSeq Series | HiSeq X Series | NovaSeq Series

Data output

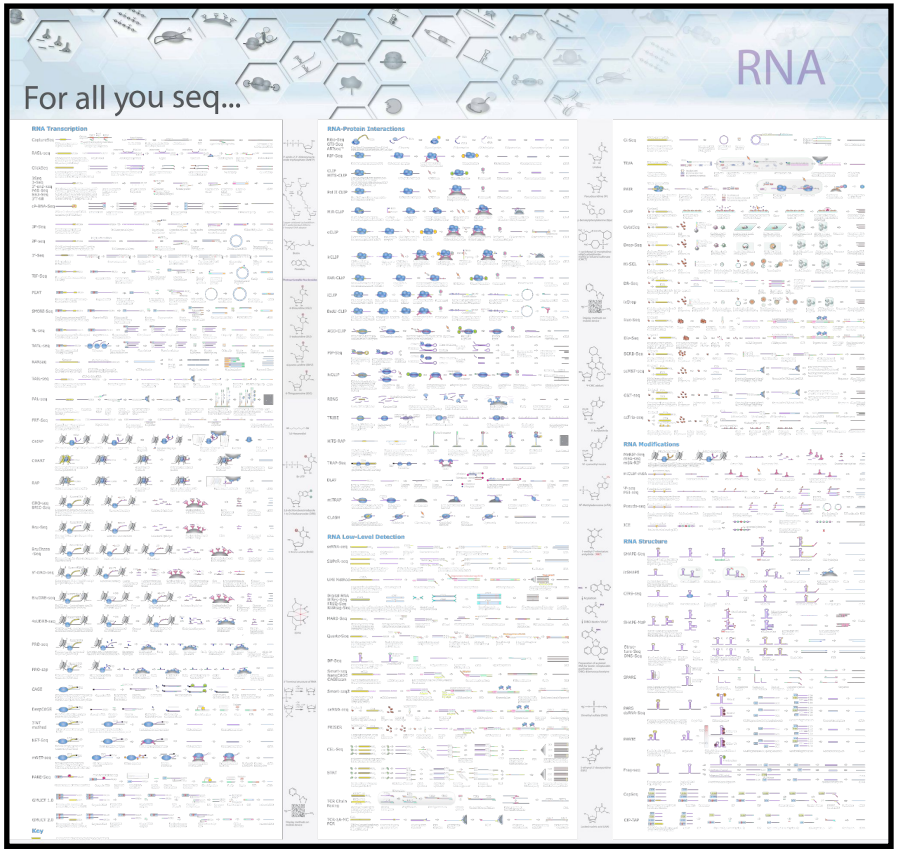
Instrument cost

Illumina sequencing

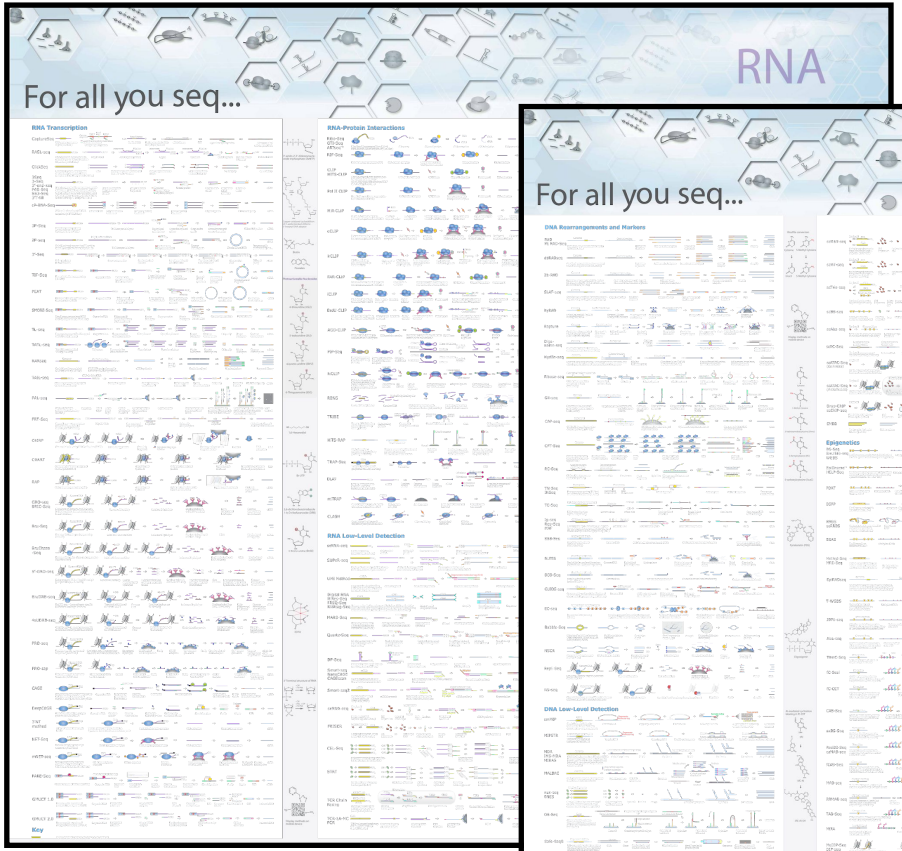


Democratization of sequencing

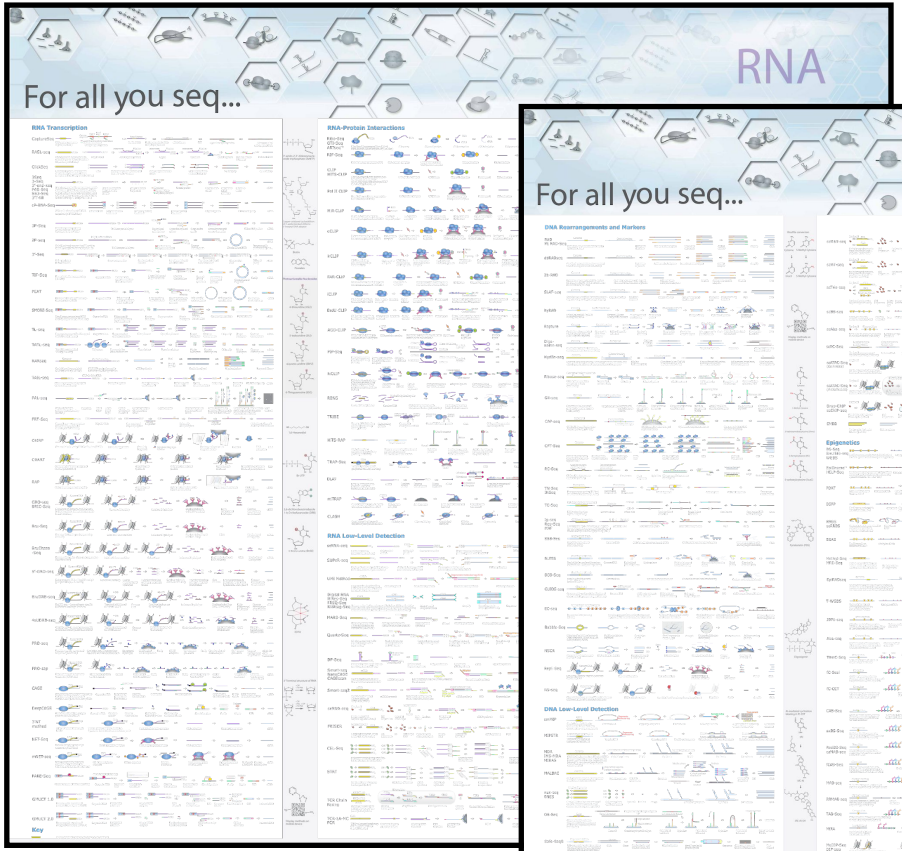
Democratization of sequencing

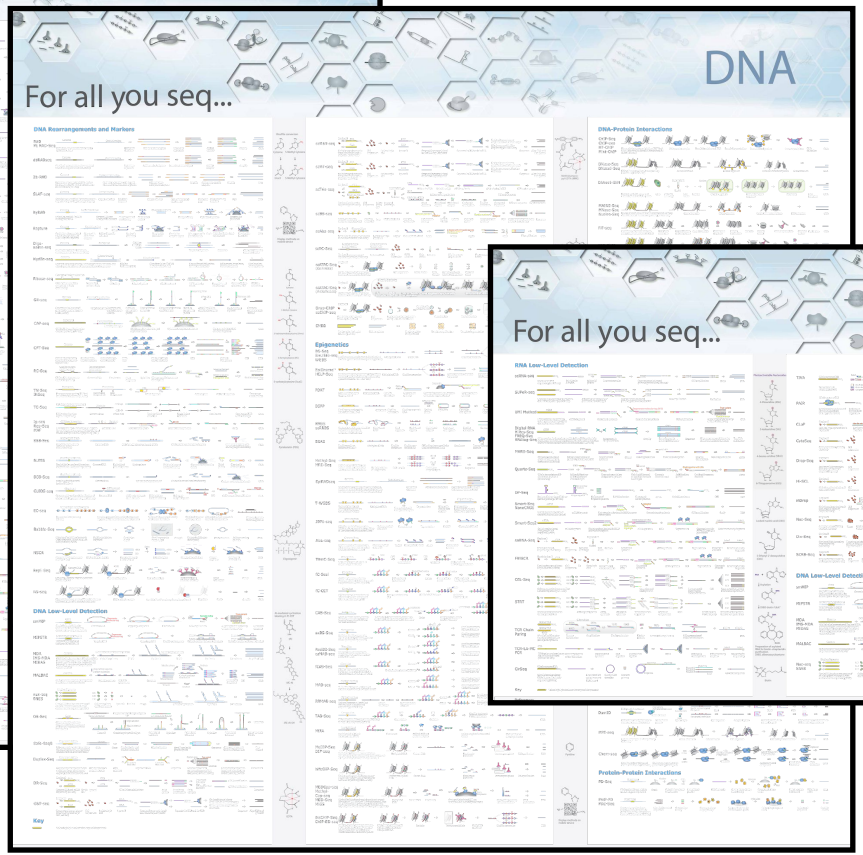


Democratization of sequencing

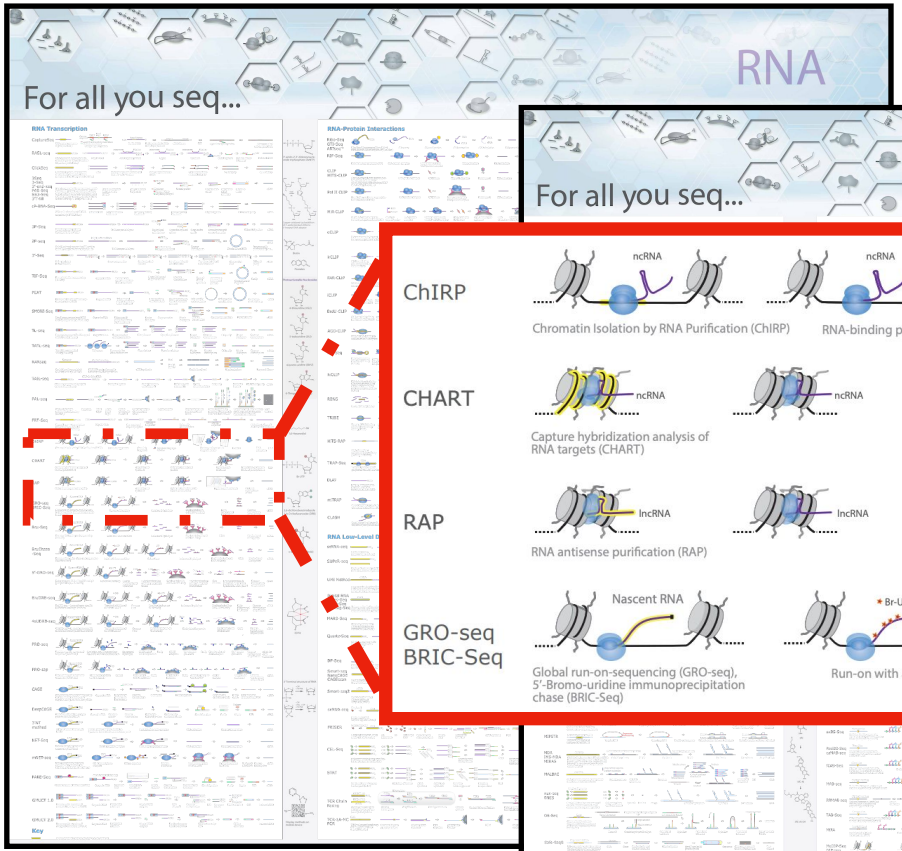


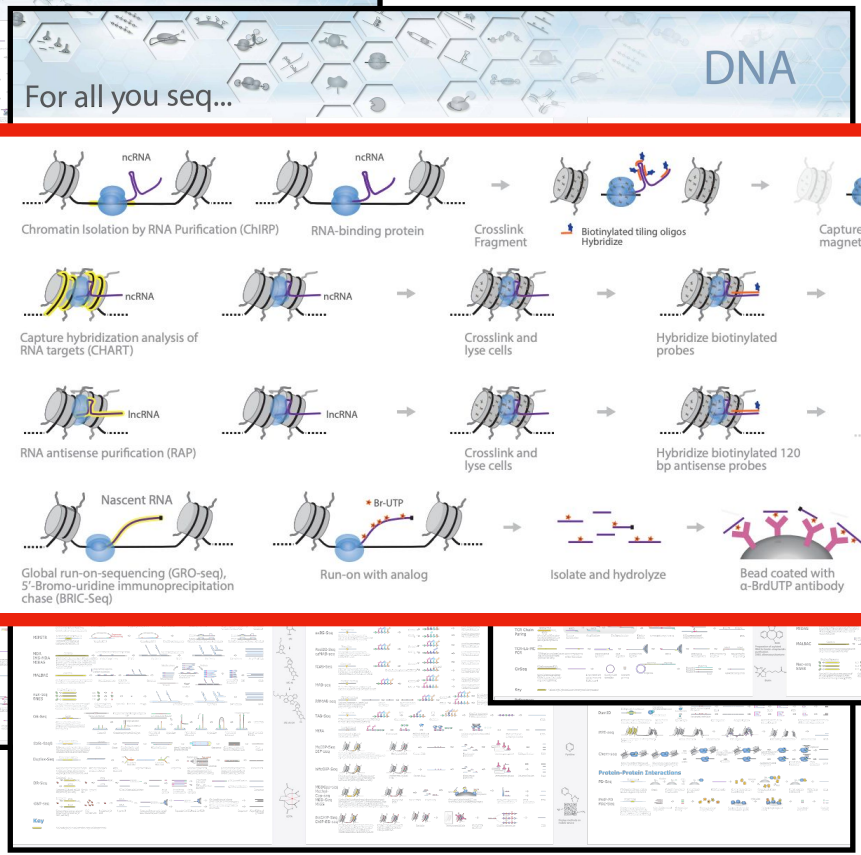
Democratization of sequencing



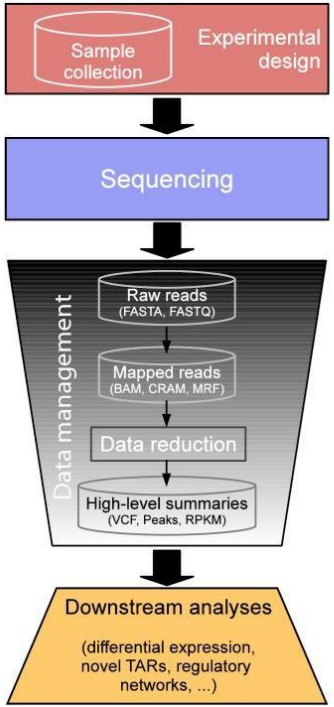


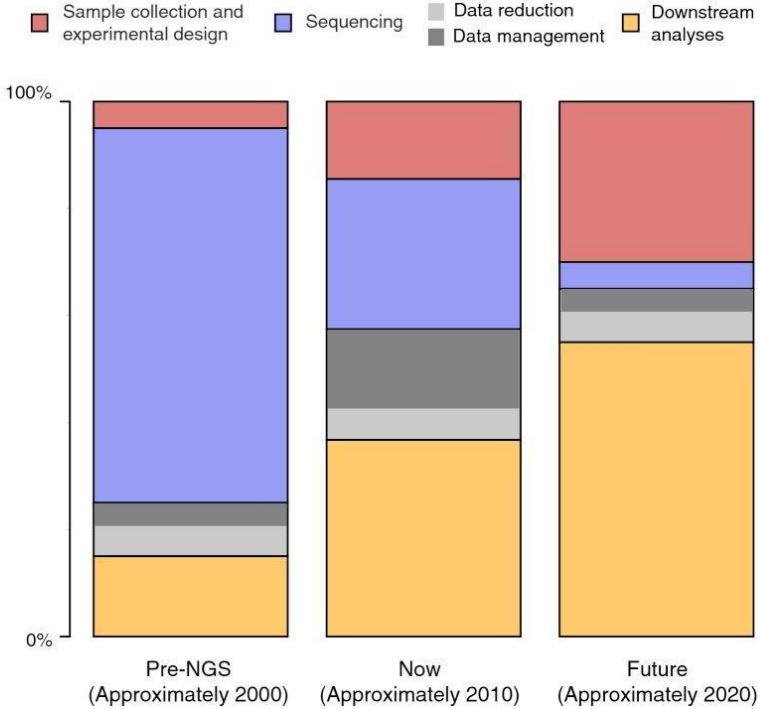
Democratization of sequencing





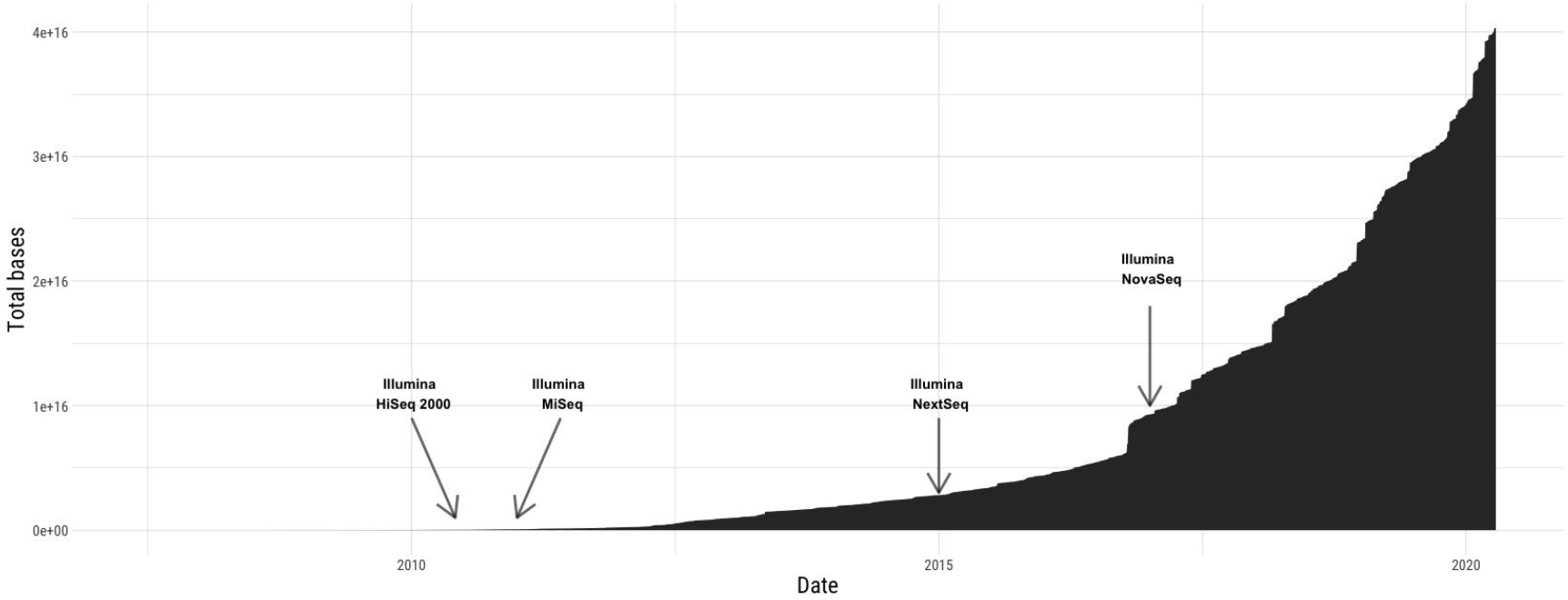
A shift in how time and money are spent on genomics experiments





Sboner A., Genome Biology, 2011

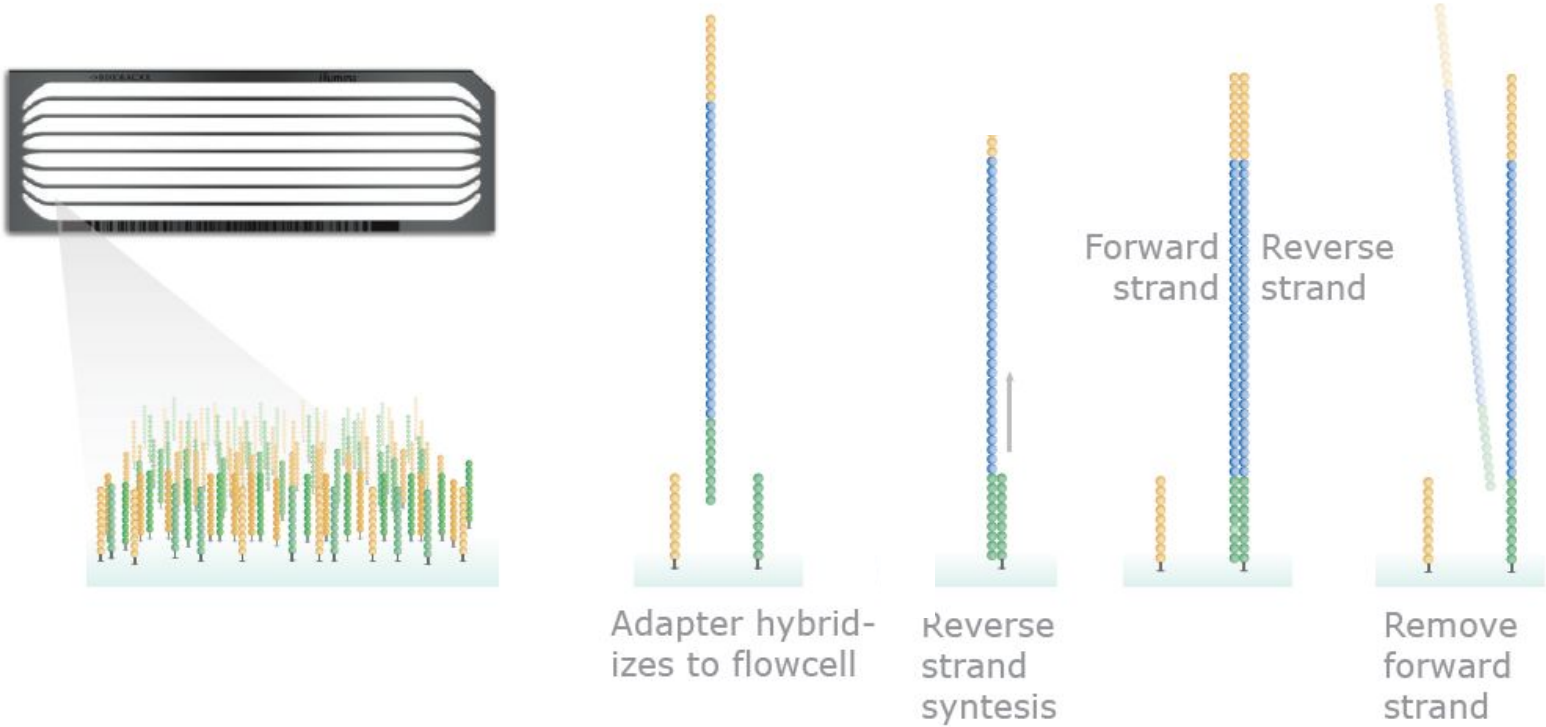
NCBI Sequence Read Archive (SRA)



Illumina sequencing

‘Sequencing By Synthesis’ (SBS) technology

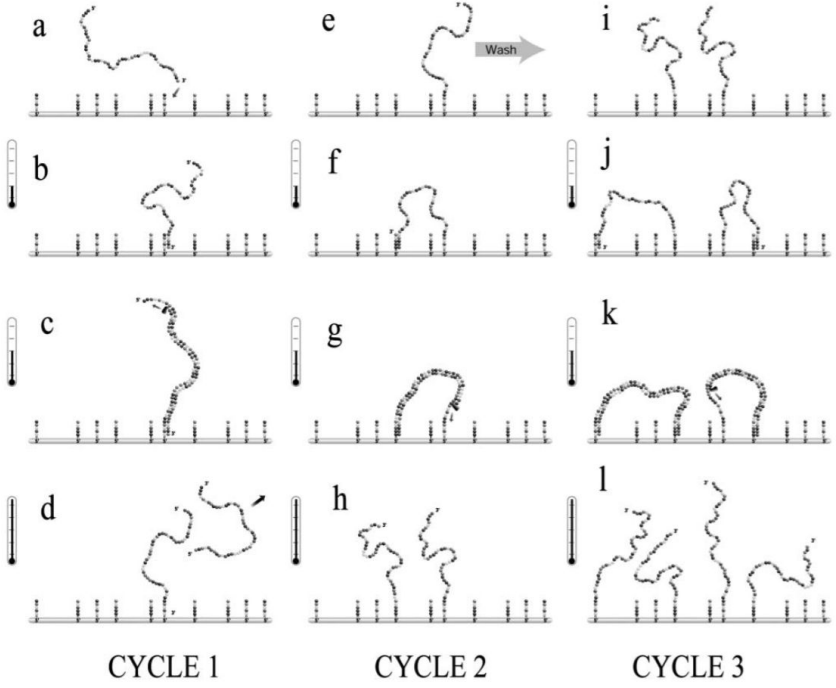
Sequencing by Synthesis



Biophysical Journal Volume 85 October 2003 2075–2086

Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model

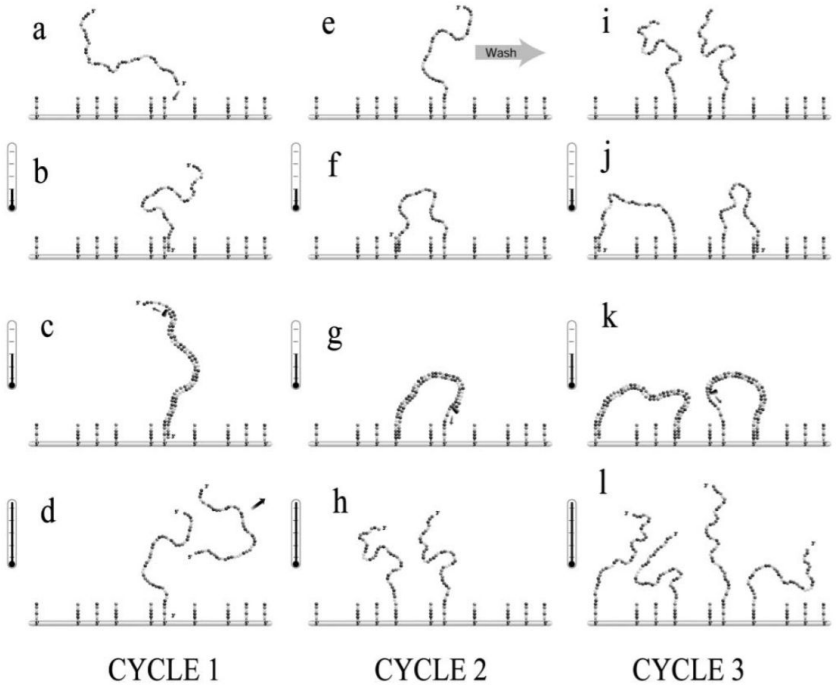
Jean-Francois Mercier,* Gary W. Slater,* and Pascal Mayer†
*Department of Physics, University of Ottawa, Ottawa, Ontario, Canada; and †Manteia Predictive Medicine S.A., Coinsins, Switzerland



Biophysical Journal Volume 85 October 2003 2075–2086 2075

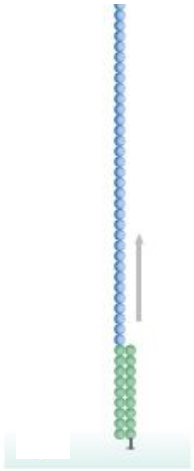
Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model

Jean-Francois Mercier,* Gary W. Slater,* and Pascal Mayer†
*Department of Physics, University of Ottawa, Ottawa, Ontario, Canada; and †Manteia Predictive Medicine S.A., Coinsins, Switzerland

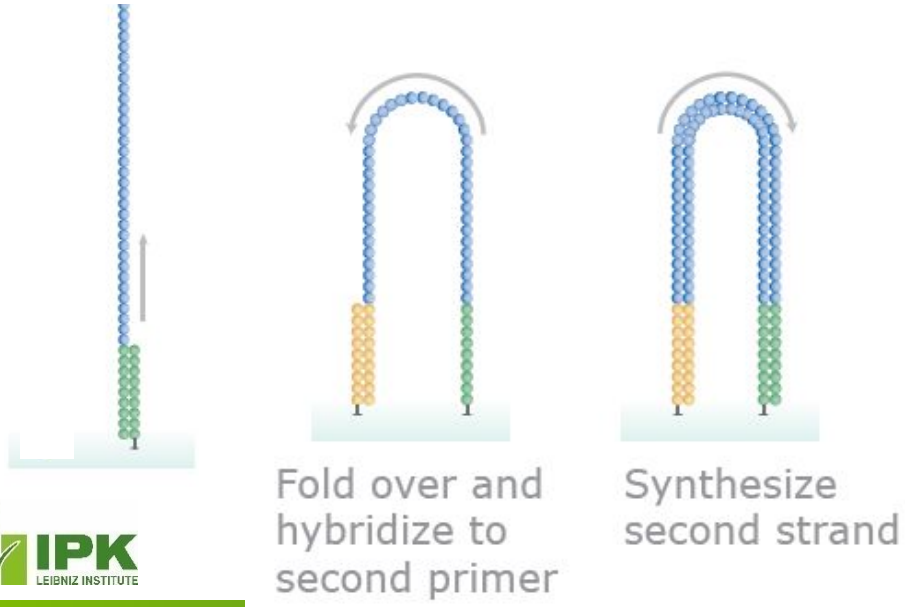


"**solid phase DNA amplification leads to the growth of a colony of molecules attached to the surface and located in the same region**. This characteristic could easily be exploited in the design of DNA microarrays."

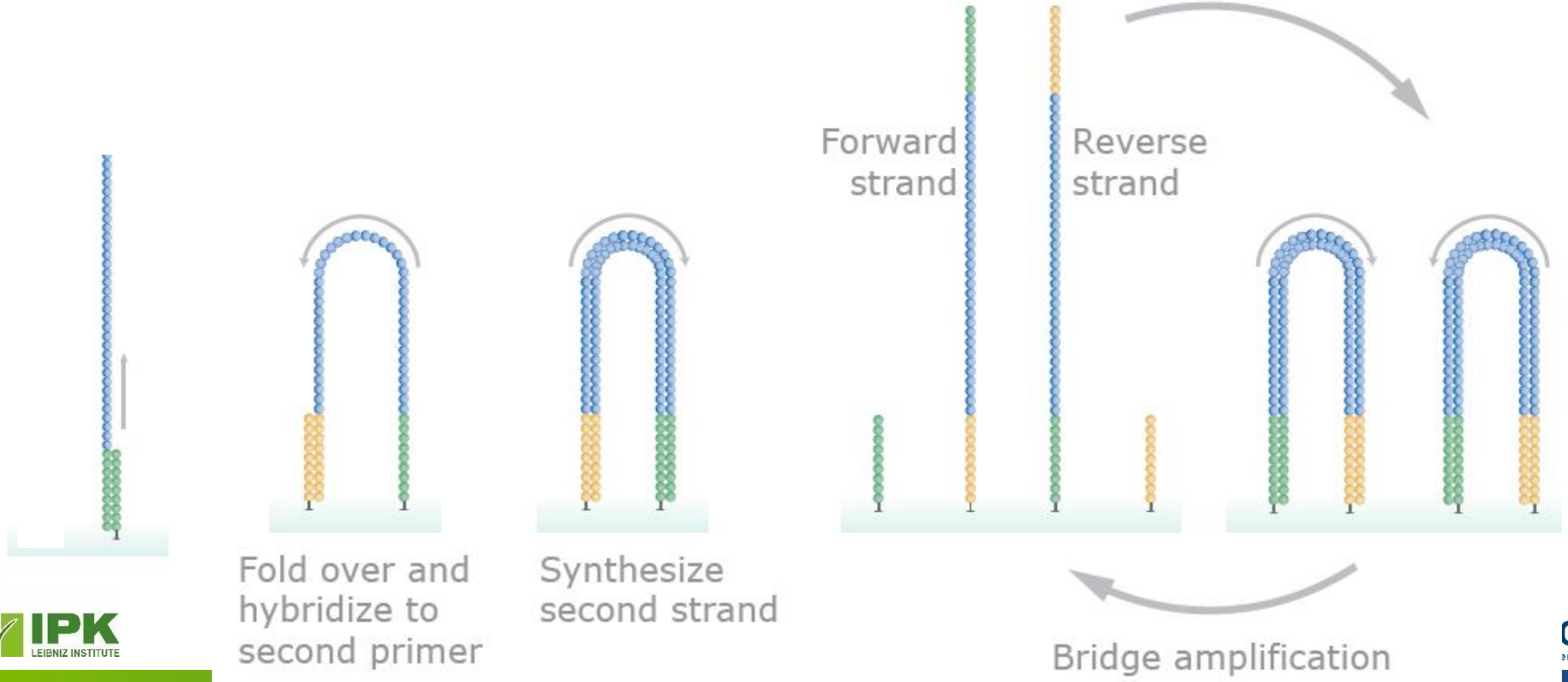
Sequencing by Synthesis (SBS)



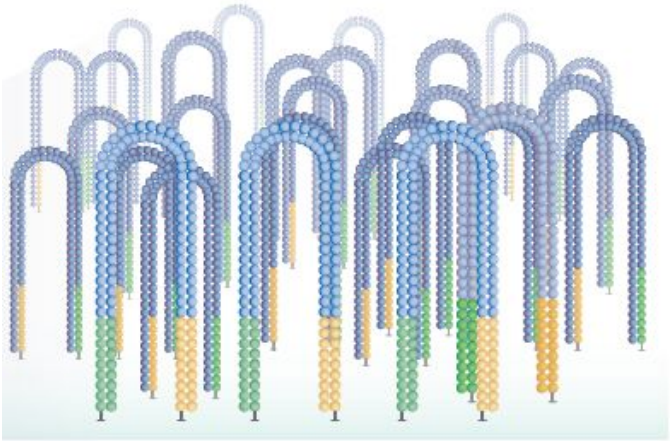
Sequencing by Synthesis (SBS)



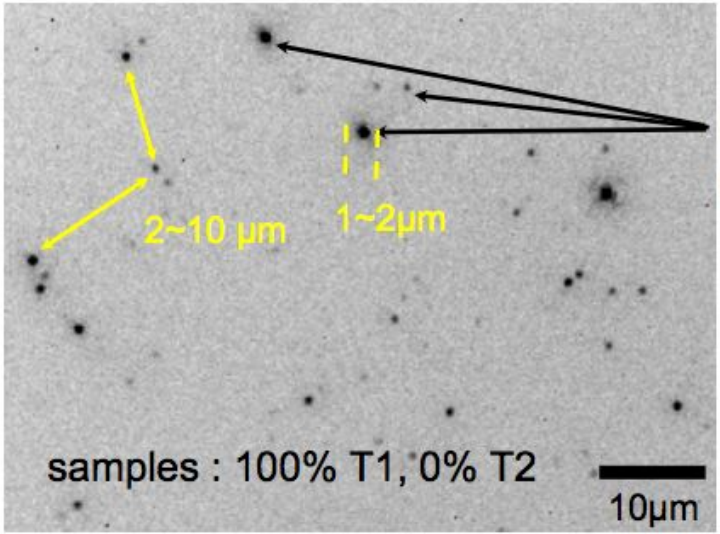
Sequencing by Synthesis (SBS)



PCR colony (aka, 'polony') size and distribution can be precisely controlled

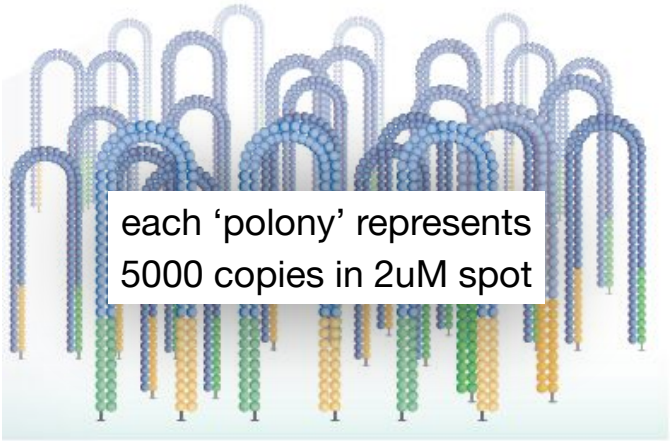


Thousands of molecules are amplified in parallel

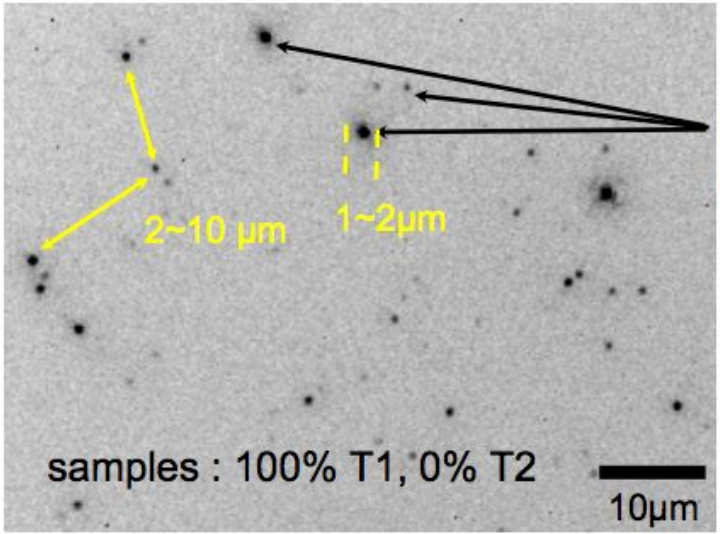


Mayer et al., Presentation 1998

PCR colony (aka, 'polony') size and distribution can be precisely controlled

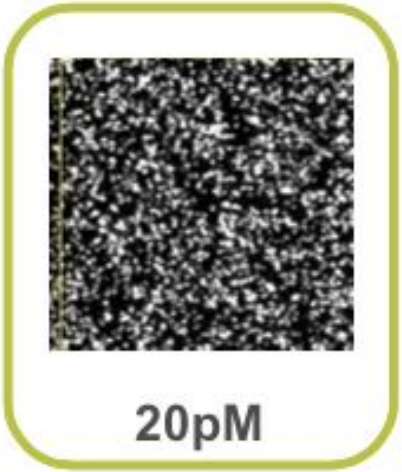


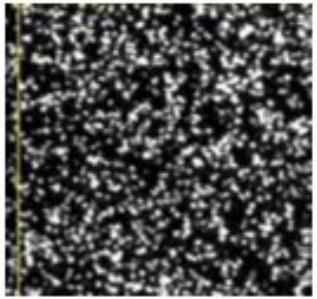
Thousands of molecules are amplified in parallel

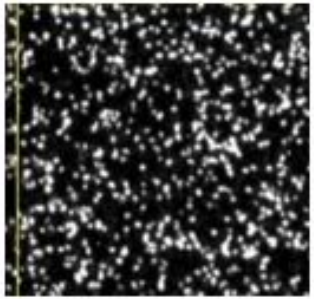


Mayer et al., Presentation 1998

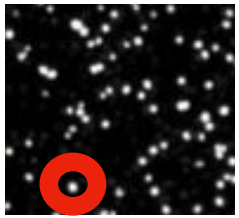
Clustering density is key to data output and quality

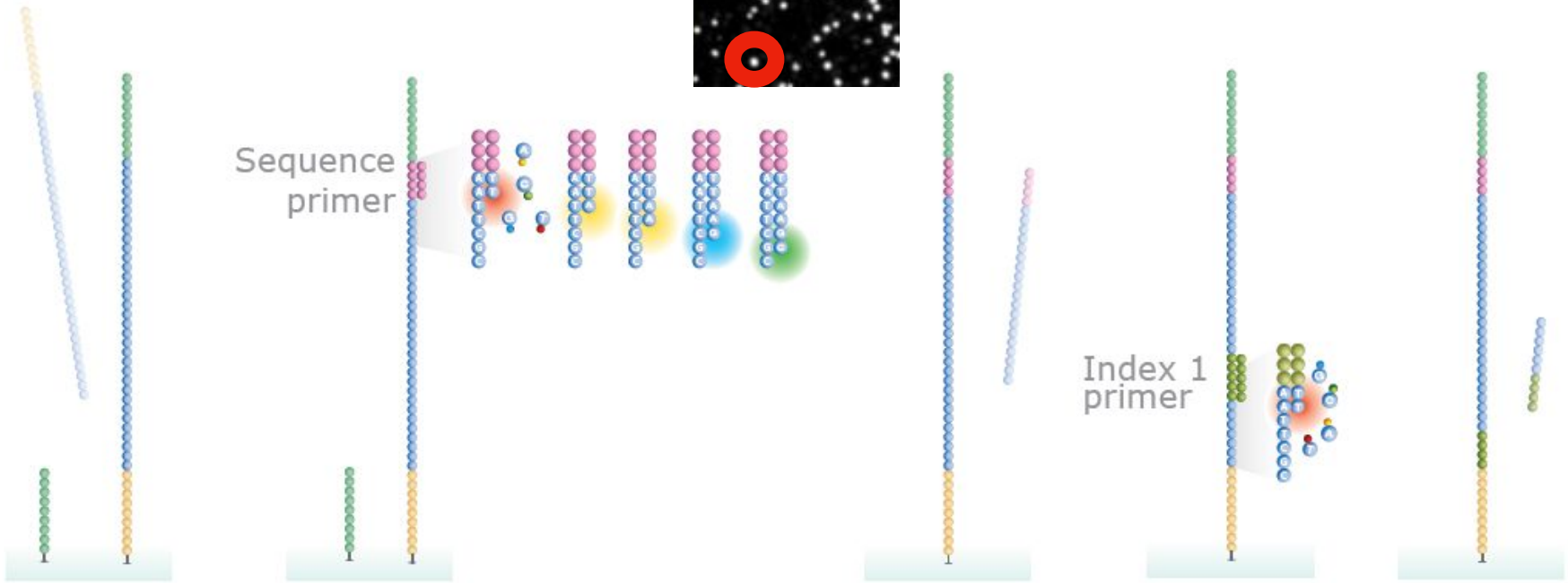






HiSeq: 950-1050K clusters/mm^2
NextSeq: 170-220K clusters/mm^2





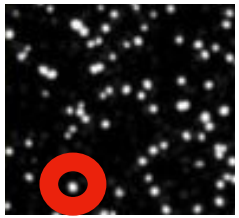
The reverse strand is cleaved and washed away

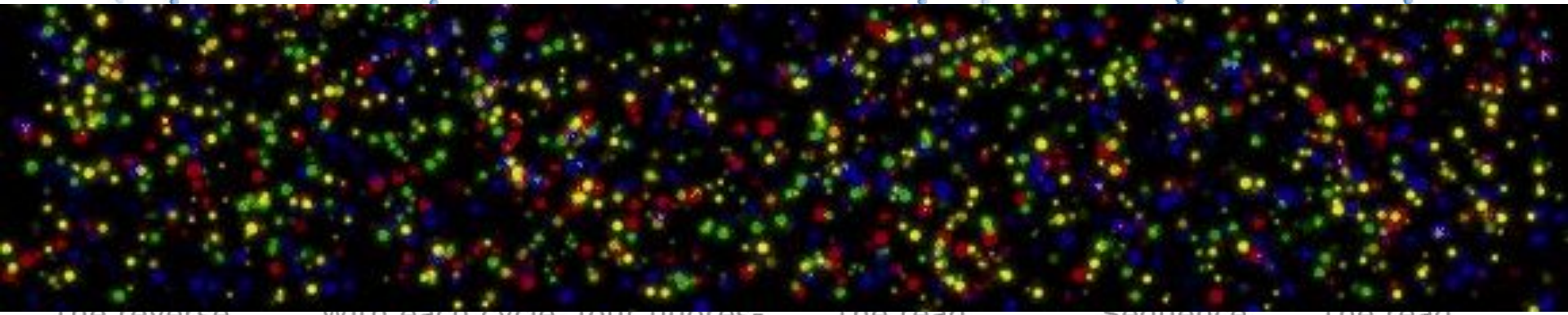
With each cycle, four fluores-cently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the sequence of the template.

The read product is washed away

Sequence Index1

The read product is washed away





The reverse strand is cleaved and washed away

With each cycle, four fluores-cently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the sequence of the template.

The read product is washed away

Sequence Index1

The read product is washed away

bit.ly/Illumina_SBS

High-throughput sequencing data is stored in .fastq format

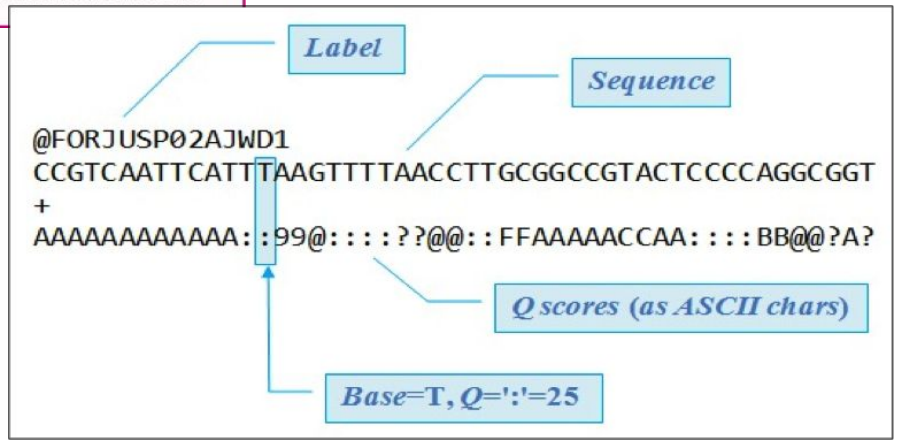
Fasta

Header
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

These are just text files

Fastq



Costs associated with bulk* sequencing

Library Preparation

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,067
stranded mRNA LT	RS-122-2101	$2,430

all kits process 48 samples

Sequencing

mid output
130M clusters

high output
400M clusters

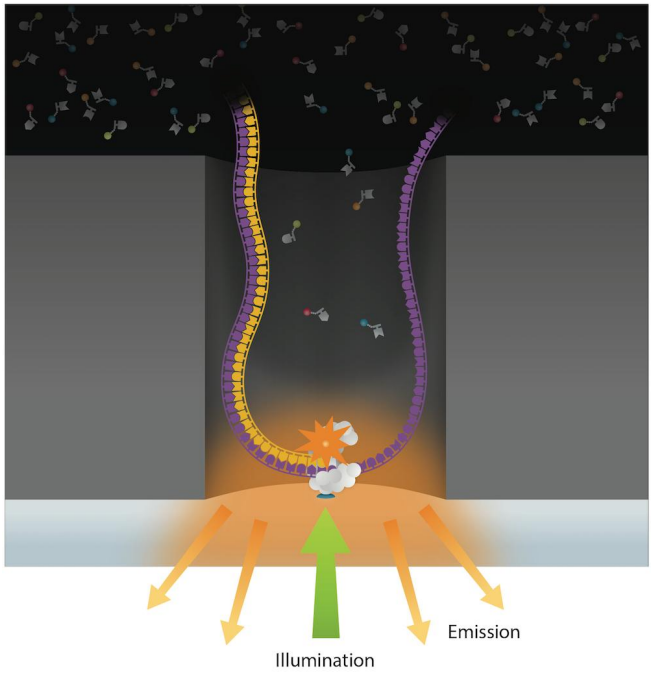
cycles	cat#	cost
300	FC-404-2004	$4,222
150	FC-404-2002	$2,635
75	FC-404-2005	$1,374

24 samples
library prep = $51/sample
sequencing = $58/sample
data output = 15M reads/sample

Long read sequencing

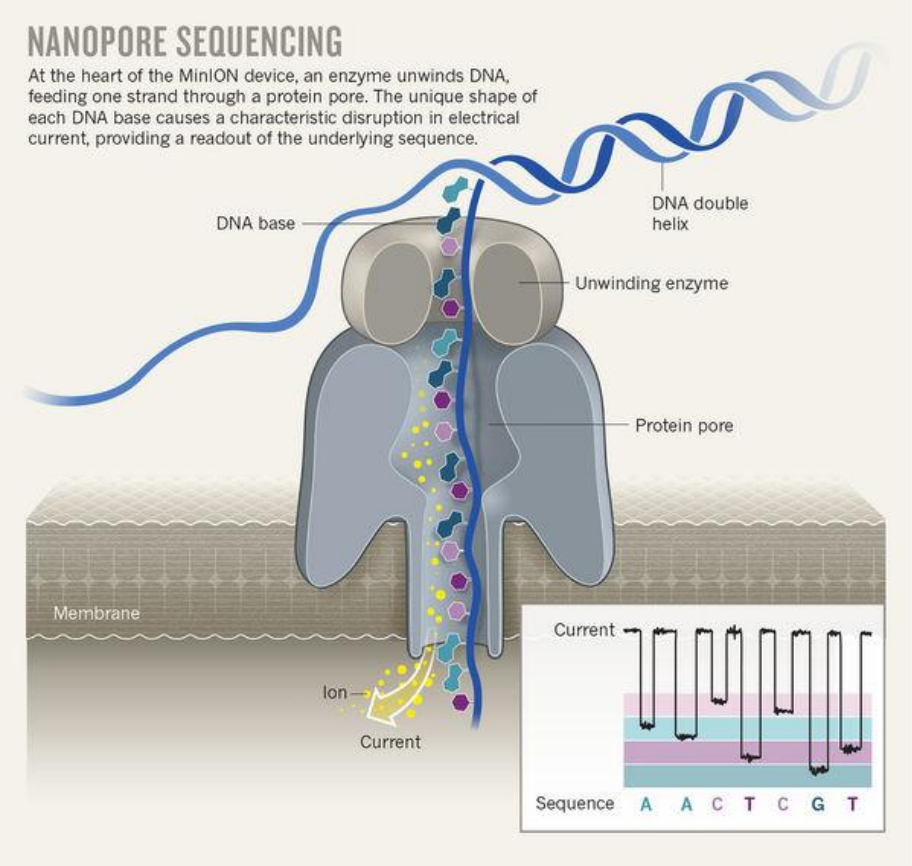
PacBio Sequel 2

SMRT sequencing

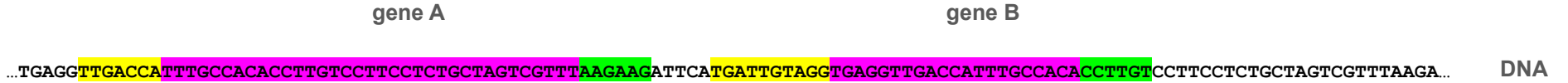


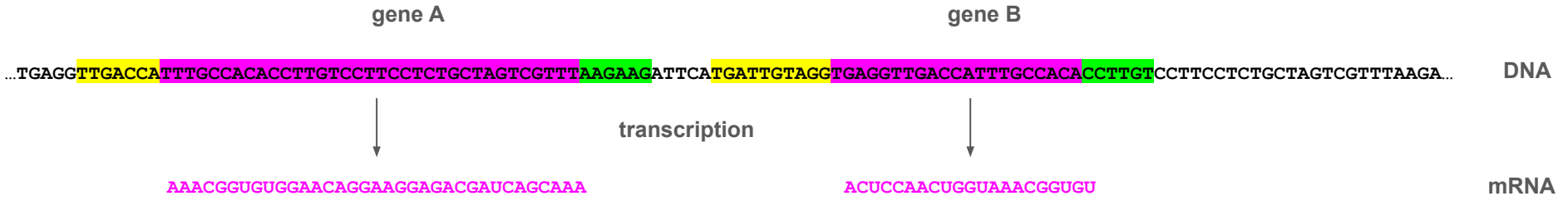
Oxford Nanopore

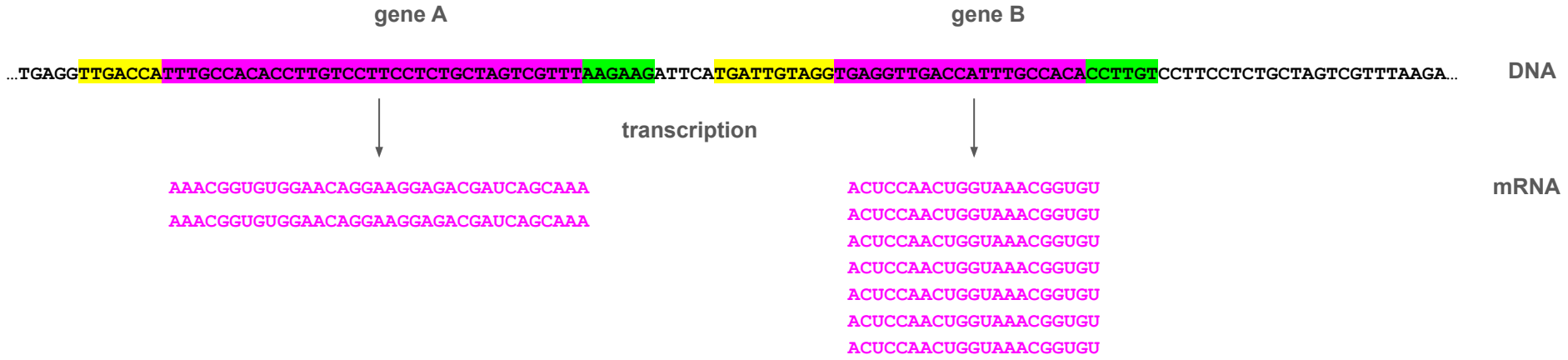
Oxford Nanopore

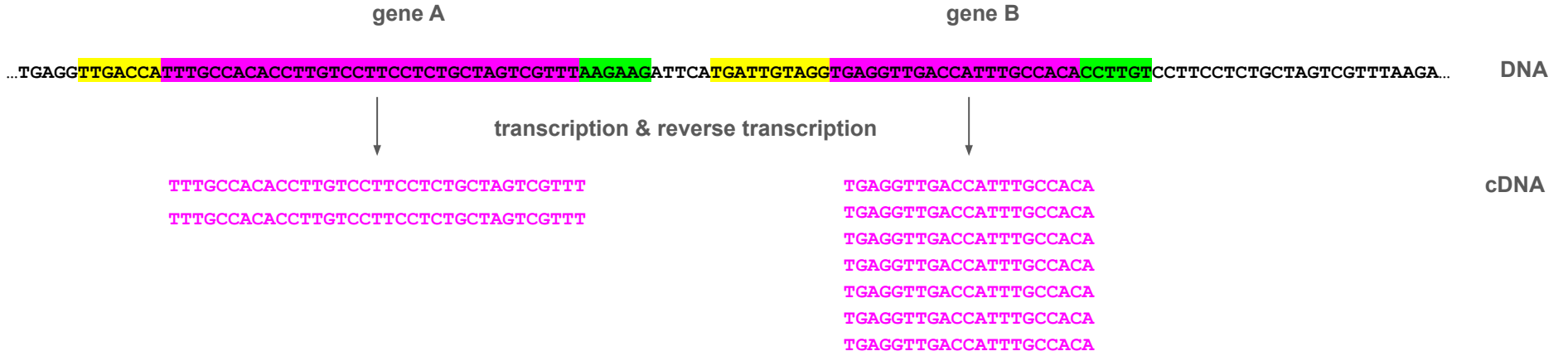


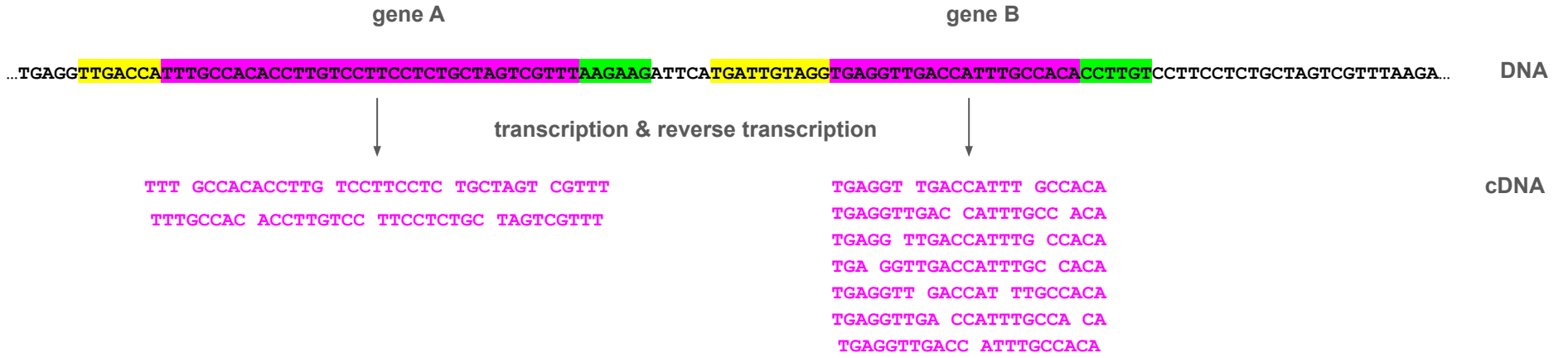
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA… DNA


gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA


gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription
AAACGGUGUGGAACAGGAAGGAGACGAUCAGCAAA
ACUCCAACUGGUAAACGGUGU
mRNA


gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription
AAACGGUGUGGAACAGGAAGGAGACGAUCAGCAAA
AAACGGUGUGGAACAGGAAGGAGACGAUCAGCAAA
ACUCCAACUGGUAAACGGUGU
ACUCCAACUGGUAAACGGUGU
ACUCCAACUGGUAAACGGUGU
ACUCCAACUGGUAAACGGUGU
ACUCCAACUGGUAAACGGUGU
ACUCCAACUGGUAAACGGUGU
ACUCCAACUGGUAAACGGUGU
mRNA


gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription & reverse transcription
TTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTT
TTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTT
TGAGGTTGACCATTTGCCACA
TGAGGTTGACCATTTGCCACA
TGAGGTTGACCATTTGCCACA
TGAGGTTGACCATTTGCCACA
TGAGGTTGACCATTTGCCACA
TGAGGTTGACCATTTGCCACA
TGAGGTTGACCATTTGCCACA
cDNA


gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription & reverse transcription
TTT GCCACACCTTG TCCTTCCTC TGCTAGT CGTTT
TTTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTTT
TGAGGT TGACCATTT GCCACA
TGAGGTTGAC CATTTGCC ACA
TGAGG TTGACCATTTG CCACA
TGA GGTTGACCATTTGC CACA
TGAGGTT GACCAT TTGCCACA
TGAGGTTGA CCATTTGCCA CA
TGAGGTTGACC ATTTGCCACA
cDNA

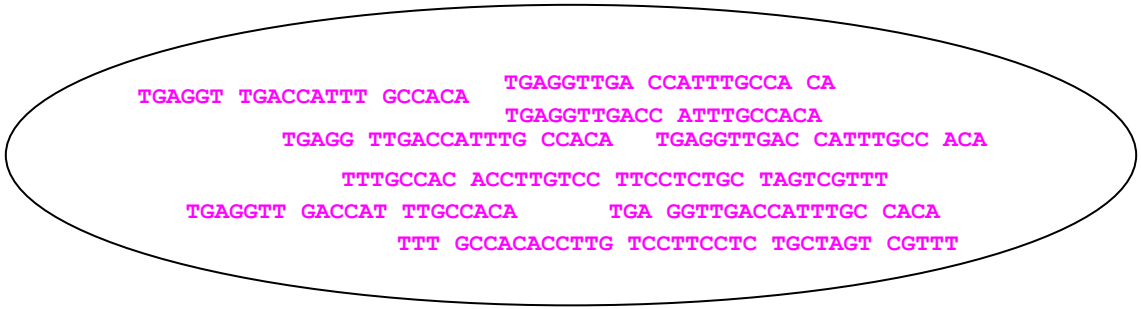
gene A

gene B

…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA… DNA

cDNA

A bag of sequenced reads



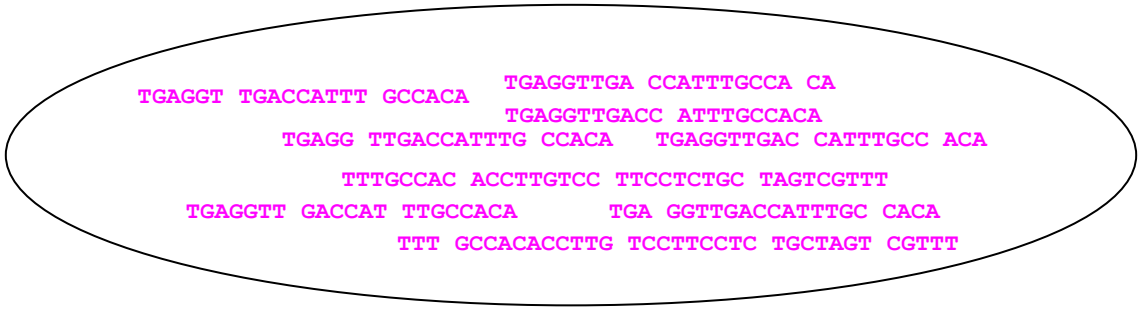
gene A

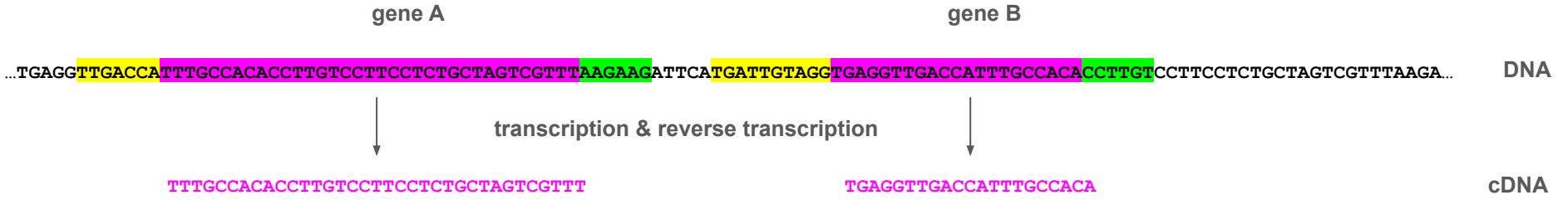
gene B

…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA… DNA

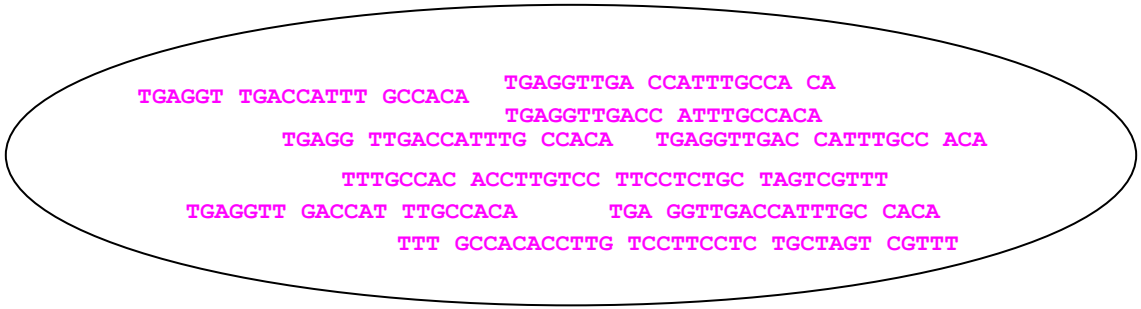
cDNA

A library!



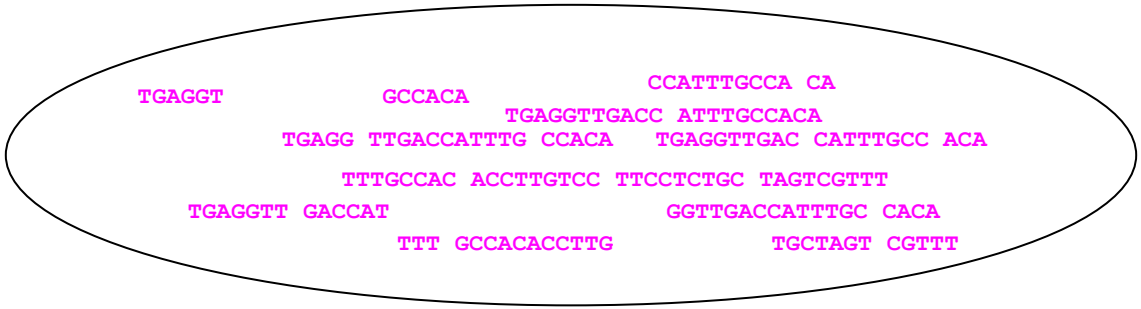

gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription & reverse transcription
TTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTT
TGAGGTTGACCATTTGCCACA
cDNA

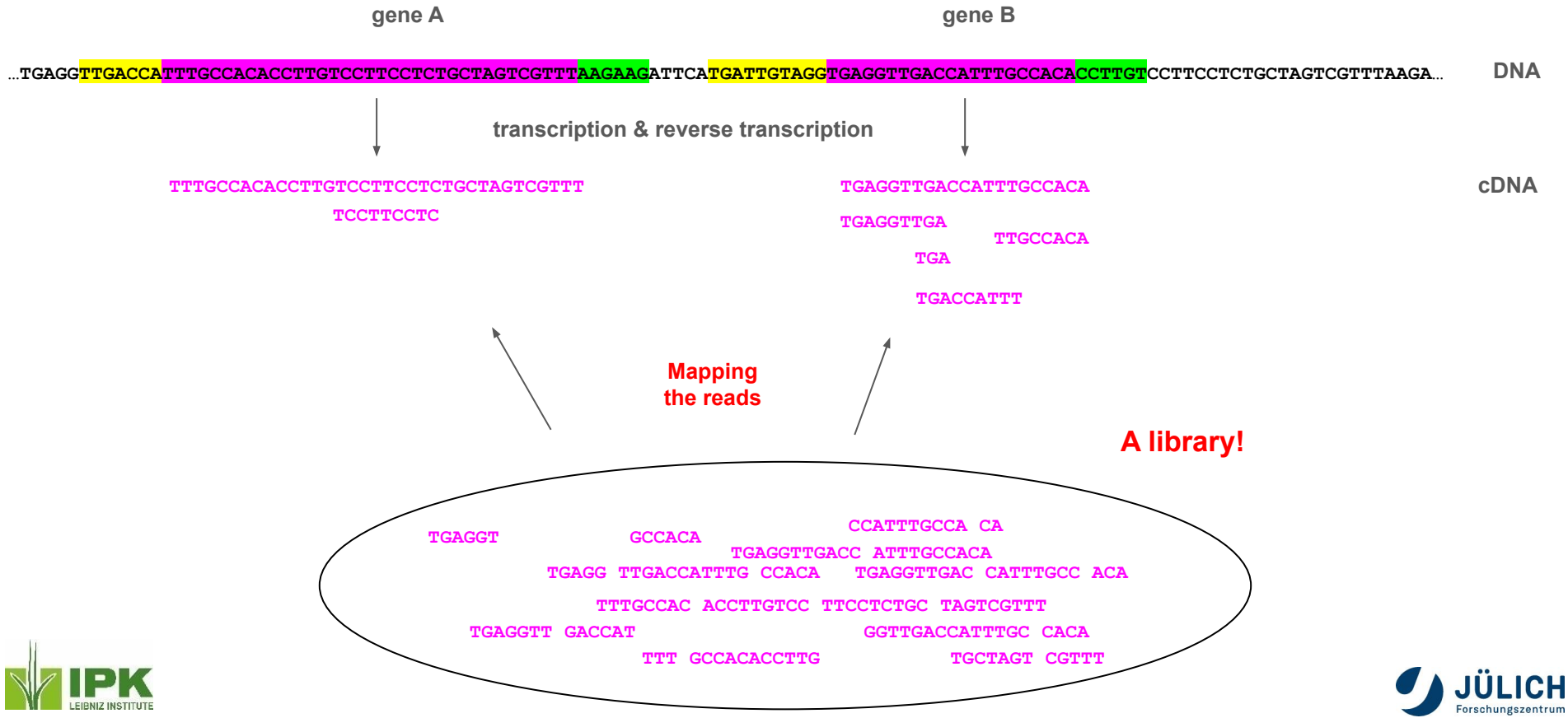
A library!


TGAGGT TGACCATTT GCCACA
TGAGGTTGA CCATTTGCCA CA
TGAGGTTGACC ATTTGCCACA
TGAGG TTGACCATTTG CCACA
TGAGGTTGAC CATTTGCC ACA
TTTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTTT
TGAGGTT GACCAT TTGCCACA
TGA GGTTGACCATTTGC CACA
TTT GCCACACCTTG TCCTTCCTC TGCTAGT CGTTT

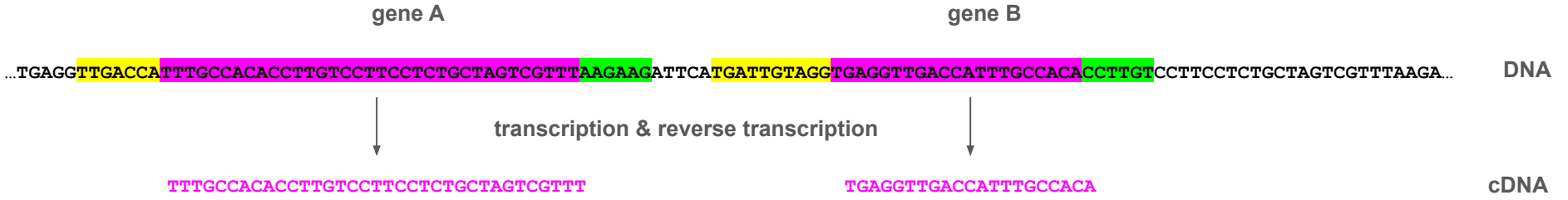

gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription & reverse transcription
TTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTT
TCCTTCCTC
TGAGGTTGACCATTTGCCACA
TGAGGTTGA
TTGCCACA
TGA
TGACCATTT
cDNA

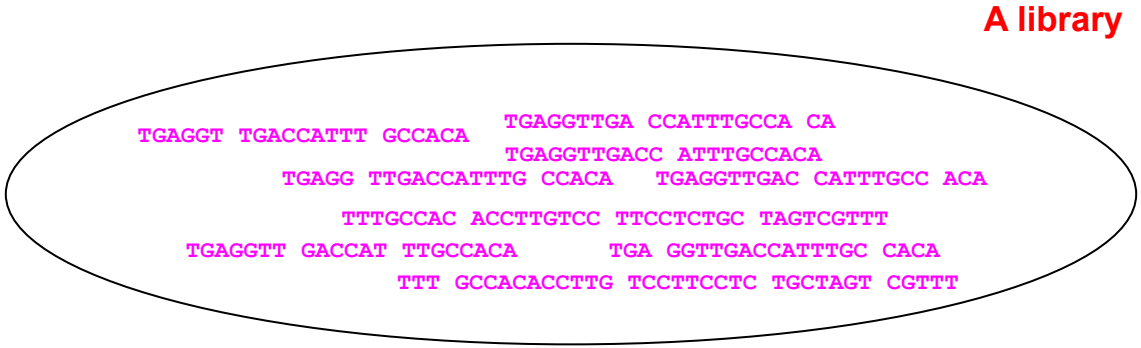
A library!


TGAGGT
GCCACA
CCATTTGCCA CA
TGAGGTTGACC ATTTGCCACA
TGAGG TTGACCATTTG CCACA
TGAGGTTGAC CATTTGCC ACA
TTTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTTT
TGAGGTT GACCAT
GGTTGACCATTTGC CACA
TTT GCCACACCTTG
TGCTAGT CGTTT

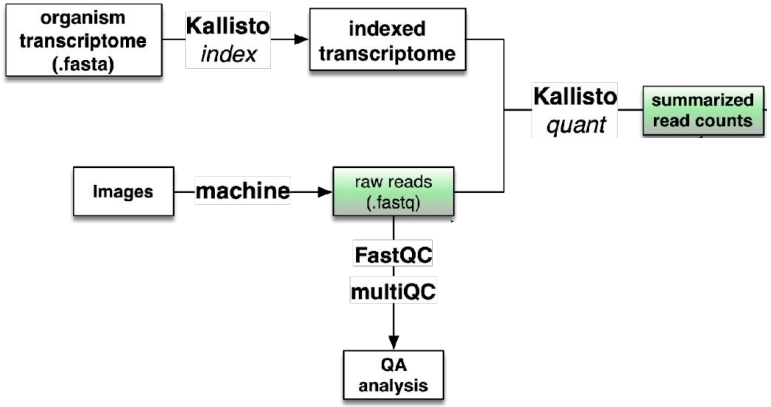

gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription & reverse transcription
TTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTT
TCCTTCCTC
TGAGGTTGACCATTTGCCACA
TGAGGTTGA
TTGCCACA
TGA
TGACCATTT
cDNA
Mapping
the reads
A library!
TGAGGT
GCCACA
CCATTTGCCA CA
TGAGGTTGACC ATTTGCCACA
TGAGG TTGACCATTTG CCACA
TGAGGTTGAC CATTTGCC ACA
TTTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTTT
TGAGGTT GACCAT
GGTTGACCATTTGC CACA
TTT GCCACACCTTG
TGCTAGT CGTTT

What data do we need to start an RNAseq project?


gene A
gene B
…TGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTTAAGA…
DNA
transcription & reverse transcription
TTTGCCACACCTTGTCCTTCCTCTGCTAGTCGTTT
TGAGGTTGACCATTTGCCACA
cDNA

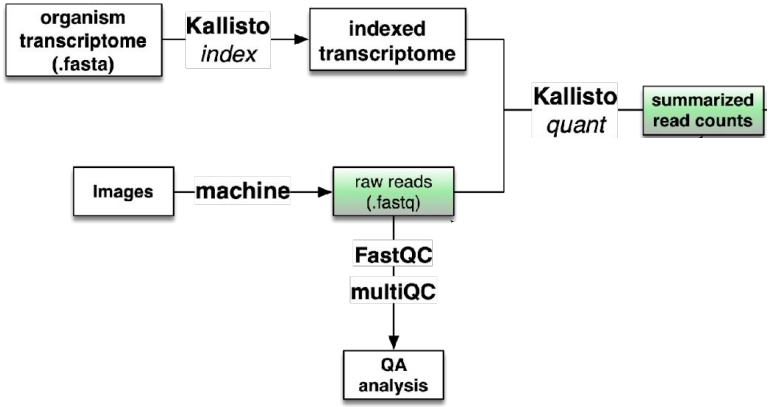

A library
TGAGGT TGACCATTT GCCACA
TGAGGTTGA CCATTTGCCA CA
TGAGGTTGACC ATTTGCCACA
TGAGG TTGACCATTTG CCACA
TGAGGTTGAC CATTTGCC ACA
TTTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTTT
TGAGGTT GACCAT TTGCCACA
TGA GGTTGACCATTTGC CACA
TTT GCCACACCTTG TCCTTCCTC TGCTAGT CGTTT

RNAseq analysis pipeline



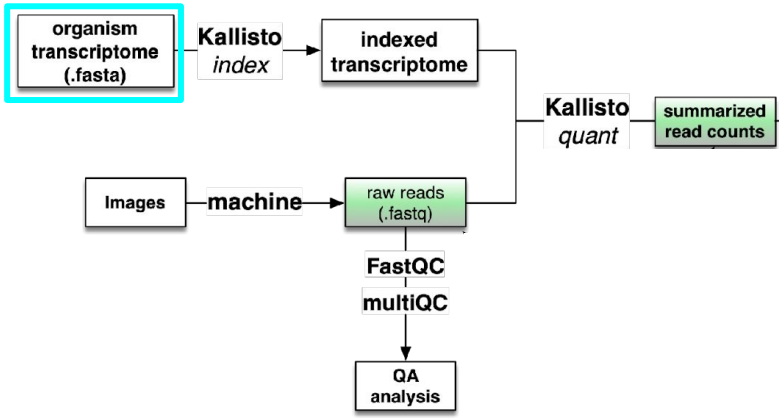
~/Documents/my_study
- scripts
- genome
- raw_reads
- publication
- QC
- mapped_reads

RNAseq analysis pipeline



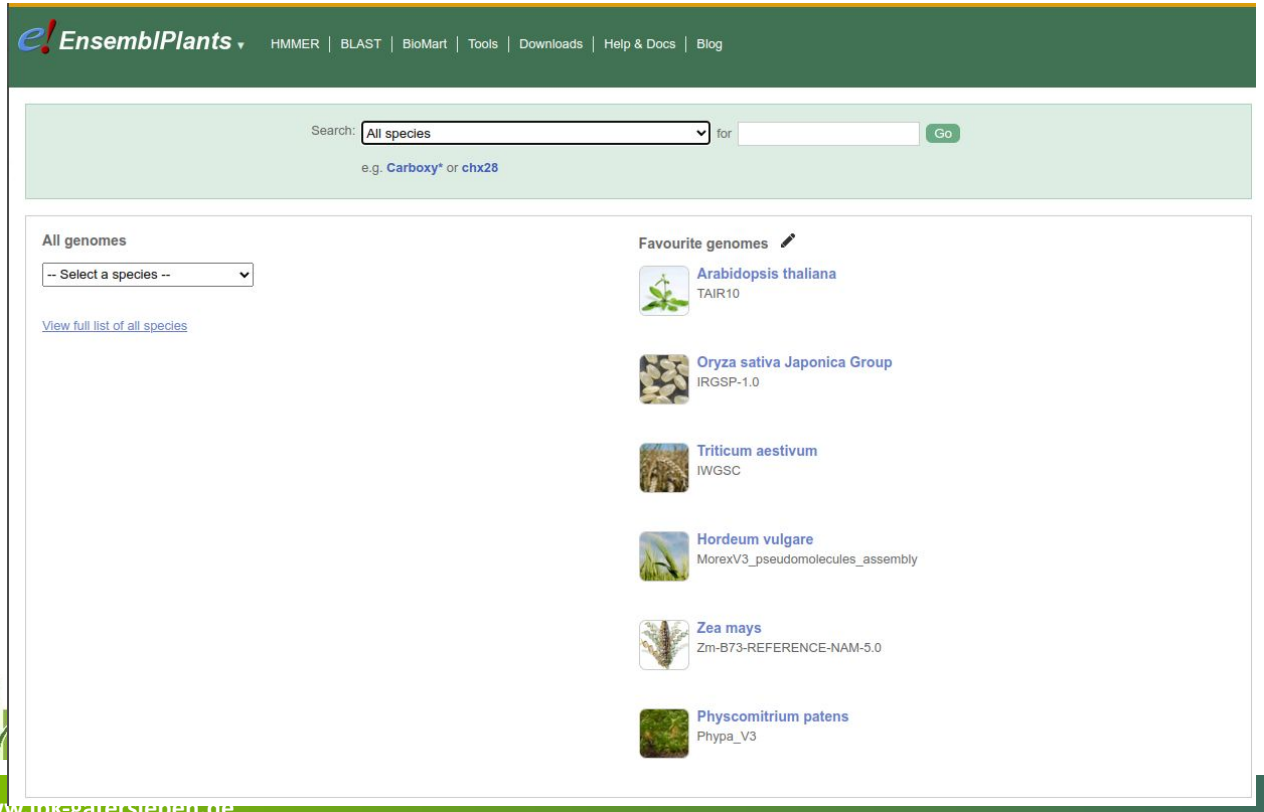
~/Documents/my_study
- scripts
- genome
- raw_reads
- publication
  - manuscript.pdf
- QC
- mapped_reads

RNAseq analysis pipeline

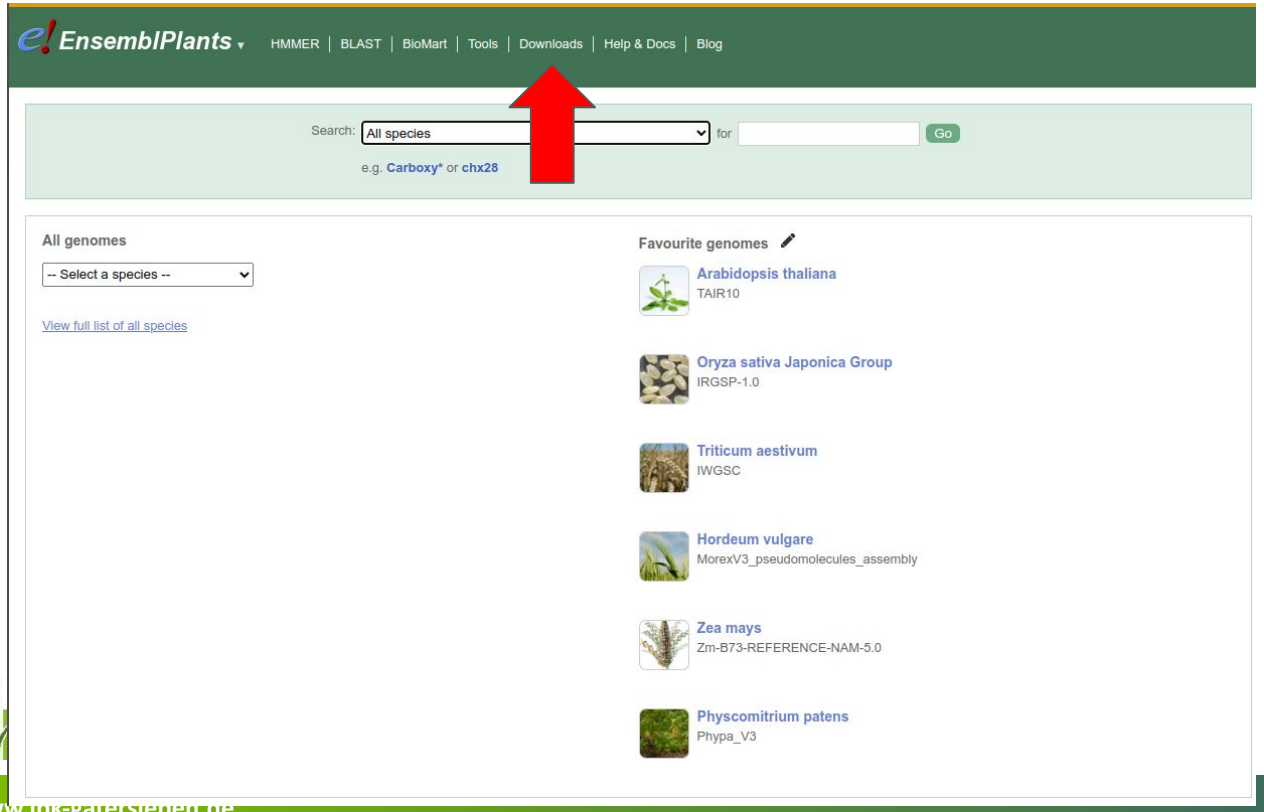


~/Documents/my_study
- scripts
- genome
- raw_reads
- publication
  - manuscript.pdf
- QC
- mapped_reads

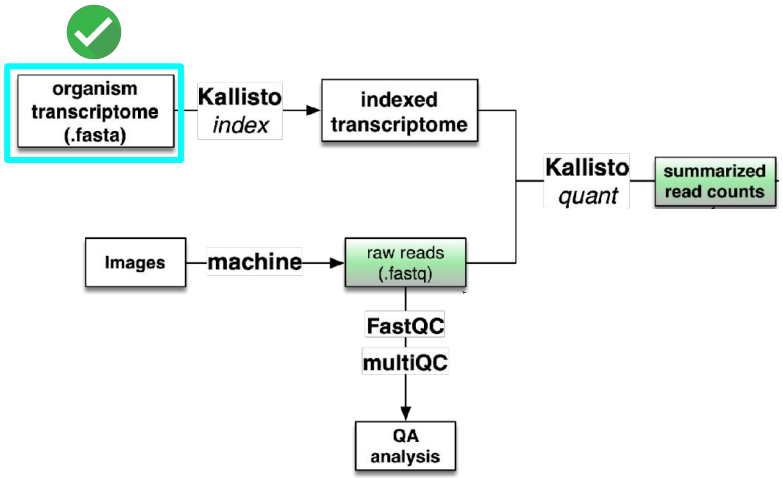
Get the reference genome



Get the reference genome

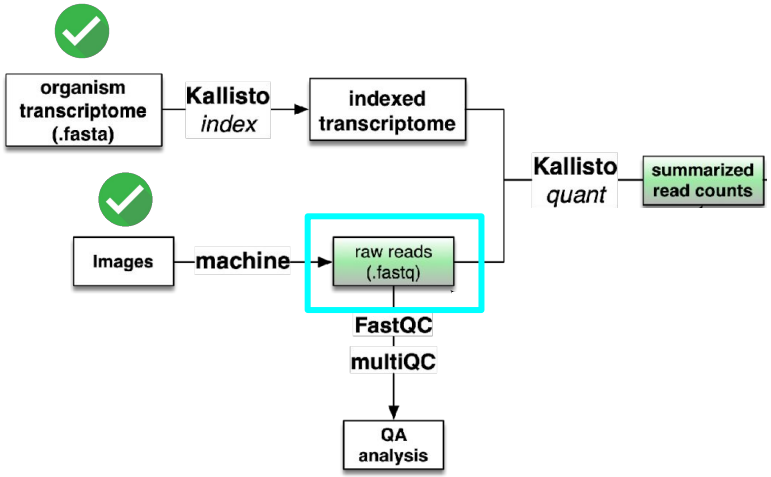


RNAseq analysis pipeline



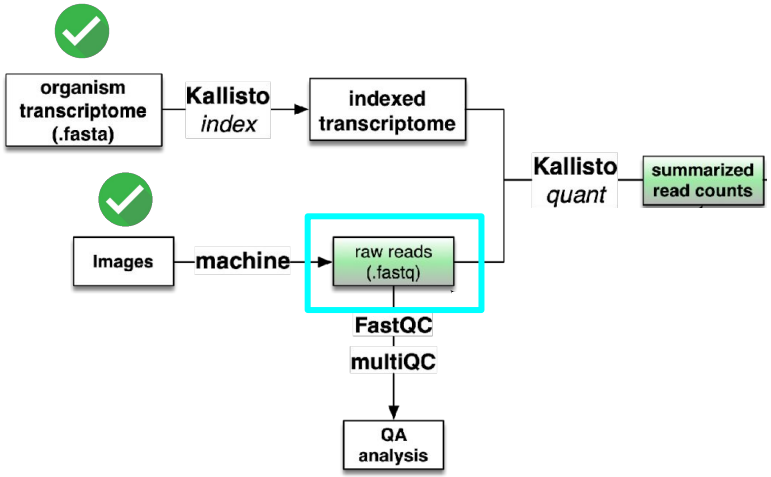

```
~/Documents/my_study
├── scripts
├── genome
│   └── Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
├── raw_reads
├── publication
│   └── manuscript.pdf
├── QC
└── mapped_reads
```

RNAseq analysis pipeline



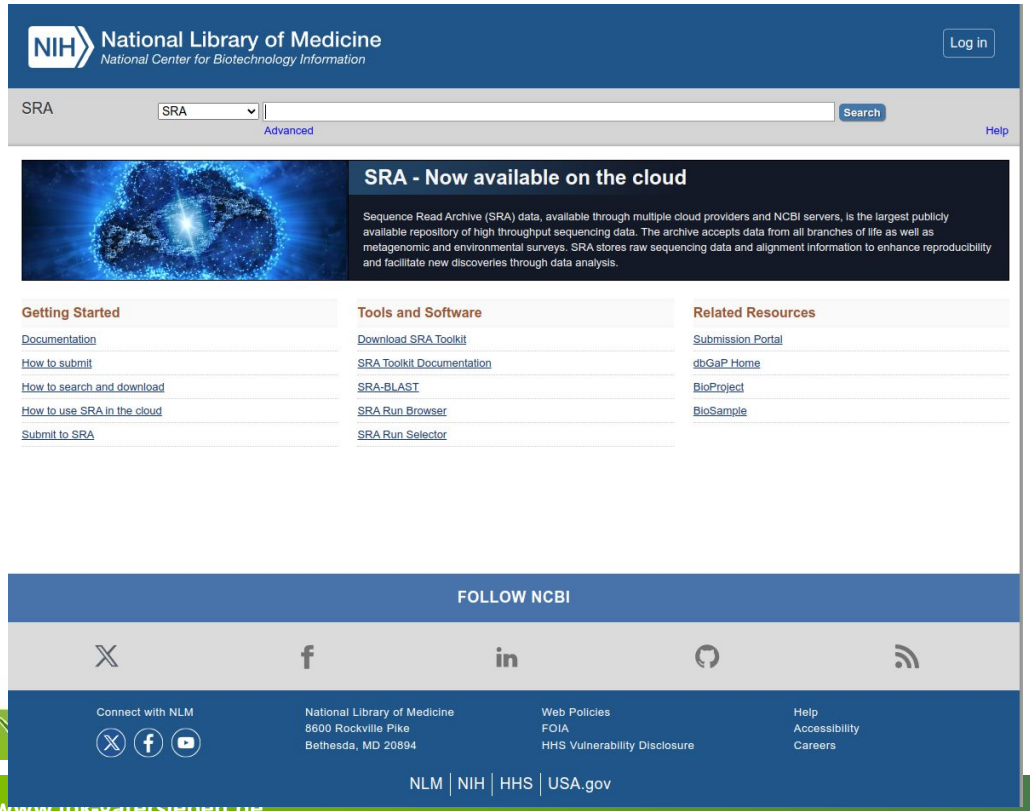
~/Documents/my_study
- scripts
- genome
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
- raw_reads
- publication
  - manuscript.pdf
- QC
- mapped_reads

RNAseq analysis pipeline



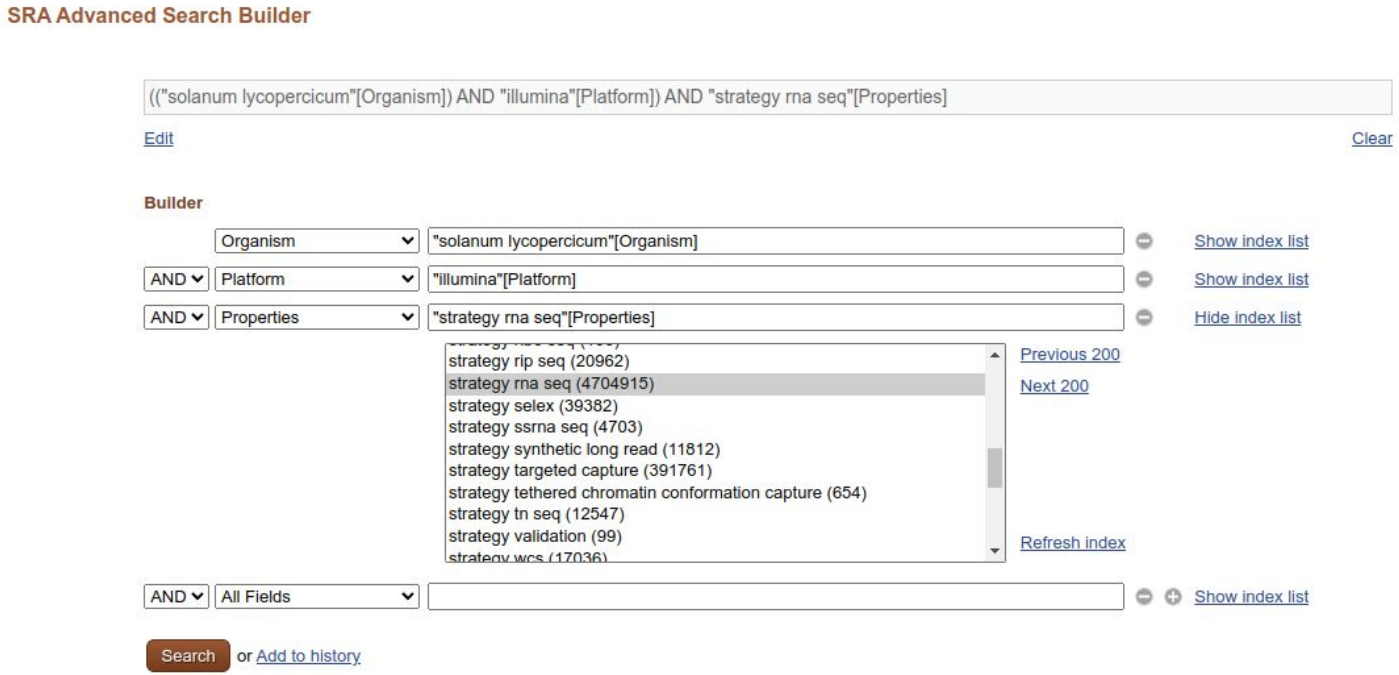
~/Documents/my_study
- scripts
- genome
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
- raw_reads
- publication
  - manuscript.pdf
- QC
- mapped_reads

From SRA search

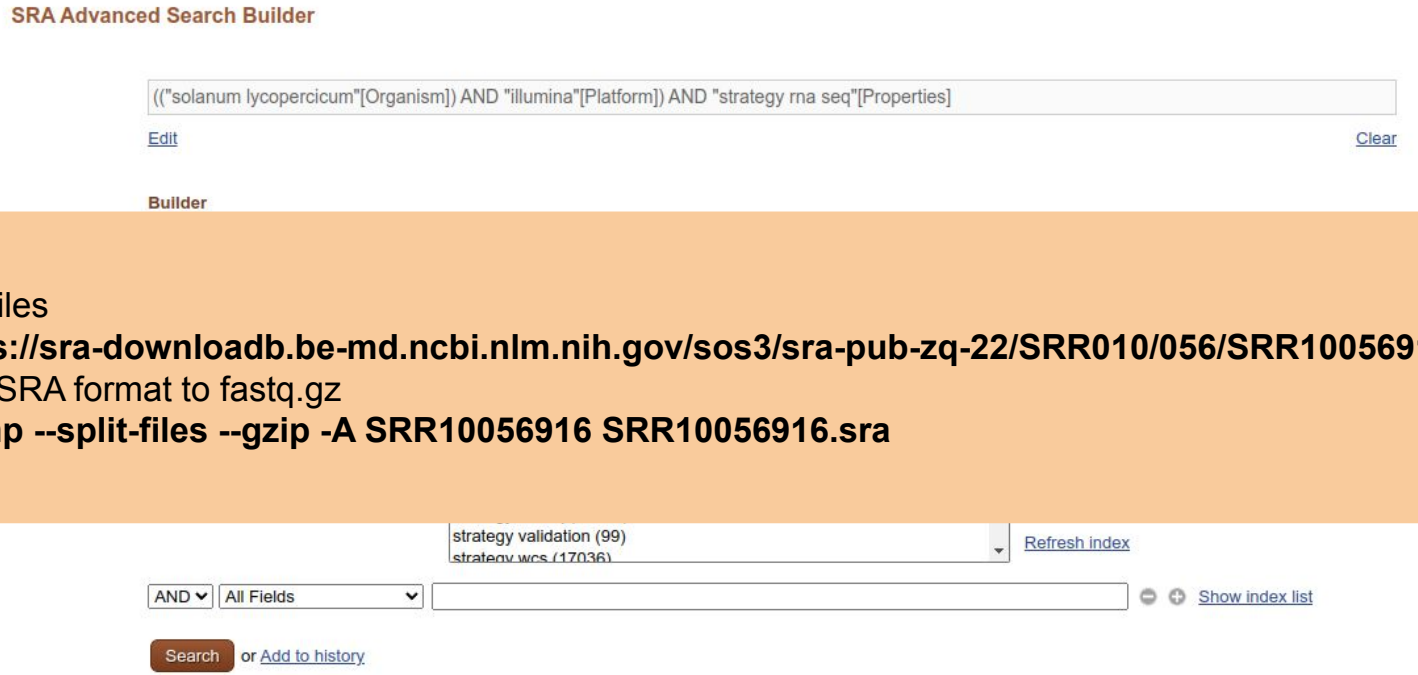


https://www.ncbi.nlm.nih.gov/sra

From SRA search

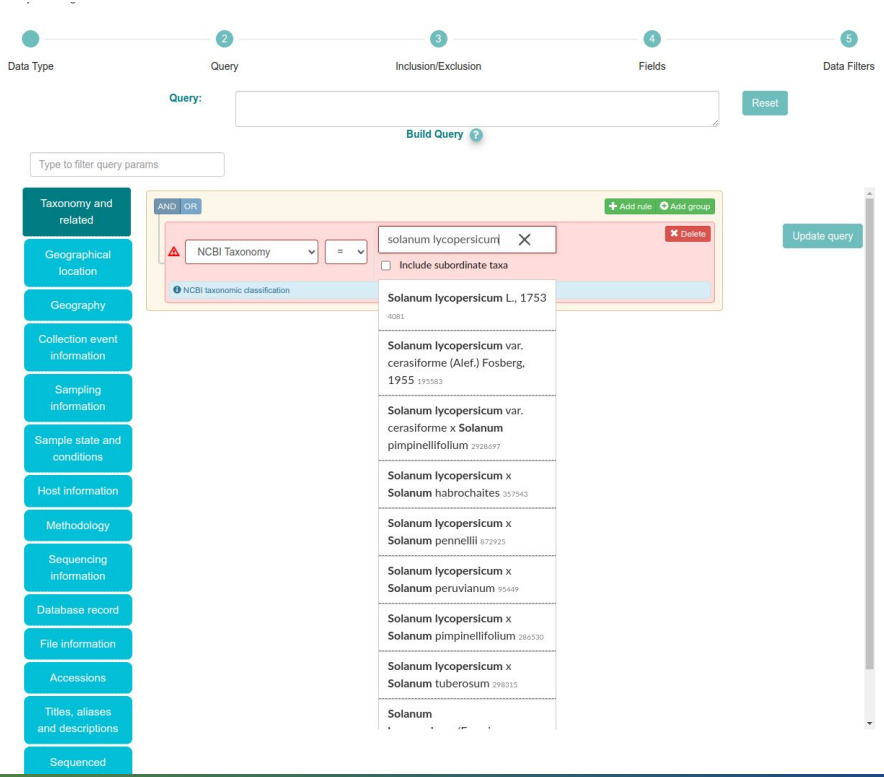


From SRA search



Get the files
wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-zq-22/SRR010/056/SRR10056916.sralite.1
Convert SRA format to fastq.gz
fastq-dump --split-files --gzip -A SRR10056916 SRR10056916.sra

From ENA search



Source data of a manuscript

HsfA2 Controls the Activity of Developmentally and Stress-Regulated Heat Stress Protection Mechanisms in Tomato Male Reproductive Tissues[1][OPEN]

Sotirios Fragkostefanakis, Anida Mesihovic, Stefan Simm, Marine Josephine Paupière, Yangjie Hu, Puneet Paul, Shravan Kumar Mishra[2], Bettina Tschiersch[3], Klaus Theres, Arnaud Bovy, Enrico Schleiff*, and Klaus-Dieter Scharf*

Department of Biosciences, Molecular Cell Biology of Plants, Goethe University, D-60438 Frankfurt am Main, Germany (S.F., A.M., S.S., Y.H., P.P., S.K.M., E.S., K.-D.S.); Cluster of Excellence Frankfurt, Goethe University, D-60438 Frankfurt am Main, Germany (S.S., E.S.); Plant Breeding, Wageningen University, Wageningen 6708PB, The Netherlands (M.J.P., A.B.); Leibniz Institute of Plant Biochemistry, D-06120 Halle (Saale), Germany (B.T., K.-D.S.); Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany (K.T.); and Buchmann Institute for Molecular Life Sciences, Goethe University, D-60438 Frankfurt am Main, Germany (E.S.)

ORCID IDs: 0000-0001-5311-7868 (S.F.); 0000-0001-6277-5141 (A.M.); 0000-0002-0910-6384 (Y.H.); 0000-0001-8220-8021 (P.P.); 0000-0003-3899-0495 (S.K.M.); 0000-0002-6956-8804 (B.T.).

Male reproductive tissues are more sensitive to heat stress (HS) compared to vegetative tissues, but the basis of this phenomenon is poorly understood. Heat stress transcription factors (Hsfs) regulate the transcriptional changes required for protection from HS. In tomato (*Solanum lycopersicum*), HsfA2 acts as coactivator of HsfA1a and is one of the major Hsfs accumulating in response to elevated temperatures. The contribution of HsfA2 in heat stress response (HSR) and thermotolerance was investigated in different tissues of transgenic tomato plants with suppressed HsfA2 levels (*A2AS*). Global transcriptome analysis and immunodetection of two major Hsps in vegetative and reproductive tissues showed that HsfA2 regulates subsets of HS-induced genes in a tissue-specific manner. Accumulation of HsfA2 by a moderate HS treatment enhances the capacity of seedlings to cope with a subsequent severe HS, suggesting an important role for HsfA2 in regulating acquired thermotolerance. In pollen, HsfA2 is an important coactivator of HsfA1a during HSR. HsfA2 suppression reduces the viability and germination rate of pollen that received the stress during the stages of meiosis and microspore formation but had no effect on more advanced stages. In general, pollen meiocytes and microspores are characterized by increased susceptibility to HS due to their lower capacity to induce a strong HSR. This sensitivity is partially mitigated by the developmentally regulated expression of HsfA2 and several HS-responsive genes mediated by HsfA1a under nonstress conditions. Thereby, HsfA2 is an important factor for the priming process that sustains pollen thermotolerance during microsporogenesis.

Source data of a manuscript

HsfA2 Controls the Activity of Developmentally and Stress-Regulated Heat Stress Protection Mechanisms in Tomato Male Reproductive Tissues[1][OPEN]

Sotirios Fragkostefanakis, Anida Mesihovic, Stefan Simm, Marine Josephine Paupière, Yangjie Hu, Puneet Paul, Shravan Kumar Mishra[2], Bettina Tschiersch[3], Klaus Theres, Arnaud Bovy, Enrico Schleiff*, and Klaus-Dieter Scharf*

Department of Biosciences, Molecular Cell Biology of Plants, Goethe University, D-60438 Frankfurt am Main, Germany (S.F., A.M., S.S., Y.H., P.P., S.K.M., E.S., K.-D.S.); Cluster of Excellence Frankfurt, Goethe University, D-60438 Frankfurt am Main, Germany (S.S., E.S.); Plant Breeding, Wageningen University, Wageningen 6708PB, The Netherlands (M.J.P., A.B.); Leibniz Institute of Plant Biochemistry, D-06120 Halle (Saale), Germany (B.T., K.-D.S.); Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany (K.T.); and Buchmann Institute for Molecular Life Sciences, Goethe University, D-60438 Frankfurt am Main, Germany (E.S.)

ORCID IDs: 0000-0001-5311-7868 (S.F.); 0000-0001-6277-5141 (A.M.); 0000-0002-0910-6384 (Y.H.); 0000-0001-8220-8021 (P.P.); 0000-0003-3899-0495 (S.K.M.); 0000-0002-6956-8804 (B.T.).

Male reproductive tissues are more sensitive to heat stress (HS) compared to vegetative tissues, but the basis of this phenomenon is poorly understood. Heat stress transcription factors (Hsfs) regulate the transcriptional changes required for protection from HS. In tomato (*Solanum lycopersicum*), HsfA2 acts as coactivator of HsfA1a and is one of the major Hsfs accumulating in response to elevated temperatures. The contribution of HsfA2 in heat stress response (HSR) and thermotolerance was investigated in different tissues of transgenic tomato plants with suppressed HsfA2 levels (*A2AS*). Global transcriptome analysis and immunodetection of two major Hsps in vegetative and reproductive tissues showed that HsfA2 regulates subsets of HS-induced genes in a tissue-specific manner. Accumulation of HsfA2 by a moderate HS treatment enhances the capacity of seedlings to cope with a subsequent severe HS, suggesting an important role for HsfA2 in regulating acquired thermotolerance. In pollen, HsfA2 is an important coactivator of HsfA1a during HSR. HsfA2 suppression reduces the viability and germination rate of pollen that received the stress during the stages of meiosis and microspore formation but had no effect on more advanced stages. In general, pollen meiocytes and microspores are characterized by increased susceptibility to HS due to their lower capacity to induce a strong HSR. This sensitivity is partially mitigated by the developmentally regulated expression of HsfA2 and several HS-responsive genes mediated by HsfA1a under nonstress conditions. Thereby, HsfA2 is an important factor for the priming process that sustains pollen thermotolerance during microsporogenesis.

...al standards.
...r (Pfaffl et al.,

...mg) was used
...ut 15 to 20 μg
...le buffer and
... proteins from
...Izol reagent
... immunoblot
...Protran nitro-
... for chemilu-
...Perkin-Elmer).
...have been de-
...rchased from

extracted using MSClust software (Tikunov et al., 2012) and subsequently annotated using NIST MSSearch (National Institute of Standards and Technology) and the T_MSRI_ID database of GC-TOF-MS spectra (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html). Compounds were annotated based on their match factor and the delta retention index between the library and the data. Data were further normalized by using total ion count.

Accession Numbers

Sequence data from this article can be found in the GenBank data libraries under acces... GSE68500.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Thermotolerance of etiolated tomato seedlings of

Source data of a manuscript

HsfA2 Controls the Activity of Developmentally and Stress-Regulated Heat Stress Protection Mechanisms in Tomato Male Reproductive Tissues[1][OPEN]

Sotirios Fragkostefanakis, Anida Mesihovic, Stefan Simm, Marine Josephine Paupière, Yangjie Hu,

ıal standards.
r (Pfaffl et al.,

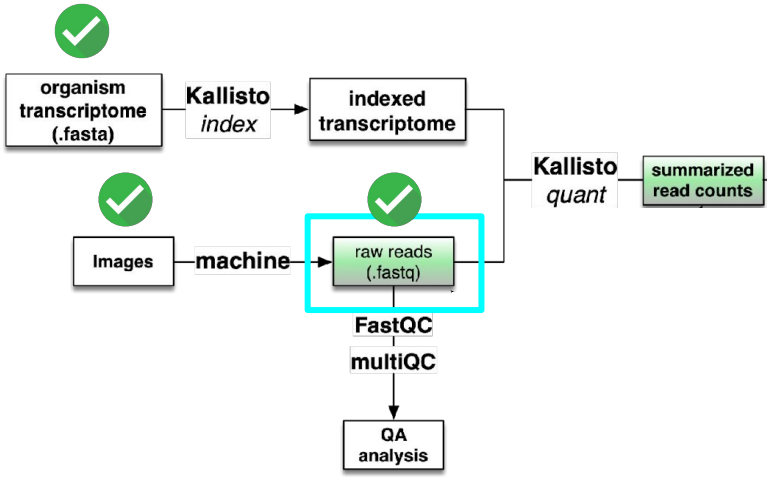
extracted using MSClust software (Tikunov et al., 2012) and subsequently annotated using NIST MSSearch (National Institute of Standards and Technology) and the T_MSRI_ID database of GC-TOF-MS spectra (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html). Compounds were annotated based on their match factor and the delta retention index between the library and the data.

:om/plphys/ar

```
# Get the files
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/003/SRR2006793/SRR2006793.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/007/SRR2006797/SRR2006797.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/002/SRR2006792/SRR2006792.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/004/SRR2006794/SRR2006794.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/001/SRR2006791/SRR2006791.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/005/SRR2006795/SRR2006795.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/006/SRR2006796/SRR2006796.fastq.gz
wget -nc -P data ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/008/SRR2006798/SRR2006798.fastq.gz
```

are characterized by increased susceptibility to HS due to their lower capacity to induce a strong HSR. This sensitivity is partially mitigated by the developmentally regulated expression of HsfA2 and several HS-responsive genes mediated by HsfA1a under nonstress conditions. Thereby, HsfA2 is an important factor for the priming process that sustains pollen thermotolerance during microsporogenesis.

RNAseq analysis pipeline



~/Documents/my_study
- scripts
- genome
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
- raw_reads
  - SRR2006793.fastq.gz
  - SRR2006797.fastq.gz
  - SRR2006792.fastq.gz
  - SRR2006794.fastq.gz
  - etc…
- publication
  - manuscript.pdf
- QC
- mapped_reads

Ingredients for an alignment

our data

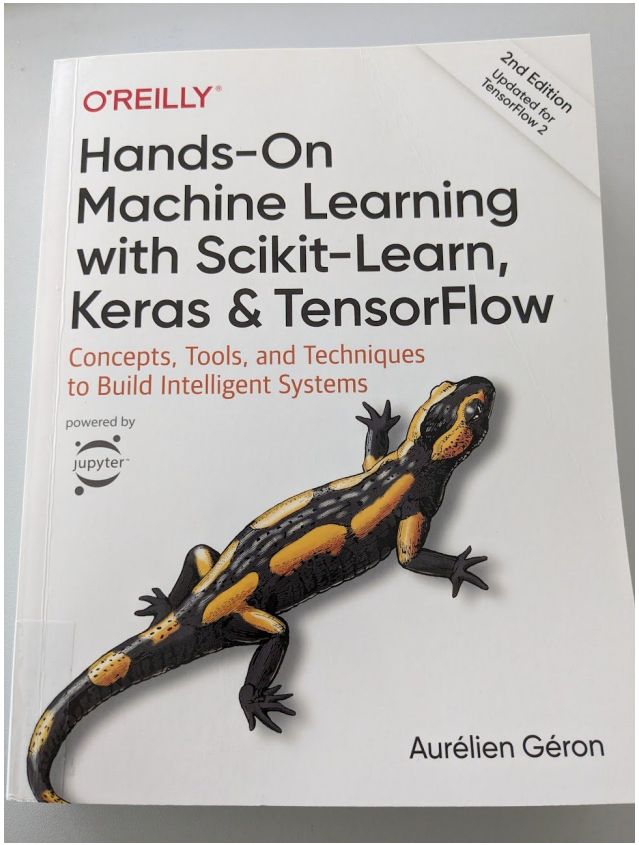
```
@SRR10056916.5114340 5114340 length=101
GATTAACATCATTTTGCATATTGTAAGCTGAATATTCTTGAGCGGTTCTACAACATTGCTTACACCCTCCTCCTCGATAATATCCACCGCAACATCCGTGG

@SRR10056916.5114341 5114341 length=101
GCCGCTTCCGCAGAGAGGTGTTCAGAGCTGGAGTTATCCAAGAGAGTAAACGCCGAAGATTTTTTTGAATCTACTCAAGAAAAGAAGAAAAGGAAGTGTCG

@SRR10056916.5114342 5114342 length=101
GTGAAAGTTACTACCGCCCTGGTGGCCGTGGTAGGGGCCGCGGCCGCGGAAGAGGTAATGGTGGTGGATATGGTGGAATCAGCAGTGTTTCAGCCCCATCC

@SRR10056916.5114343 5114343 length=101
CTAGGGATCTTTTTGATCTCATCAGAGATTTATACGTCAATAGAAAAAGTAGAGATGACCTAAGAGTAAAGAATCTAATATATTGAATATTGAAATGAGGG

@SRR10056916.5114344 5114344 length=100
AGACAGAGCTGTTTTTGCCCTGCAGAACTGCAACGGGATAGTTGAGGAGGCCACCGAGGAGGAGTTGATGGATGCTATGGCTCAGGCAGACTCCACTGGG

@SRR10056916.5114345 5114345 length=99
AGTATGTTAACACCTATCGCCGAACATTCACAAACAAAAAGGGGAAAGAGGTTAGCAAGGCTCCTAAGATCCAGAGGCTGGTAACACCATTGACTCTTC

@SRR10056916.5114346 5114346 length=100
ACTAGAAGATCATCACCCACTTTTACATCTTCAGCAAAACCATCATAGTTAACAGTGATGGTGCGTTCTGGAAGCGGTGGATCAAATGCACGAACAGTGA

@SRR10056916.5114347 5114347 length=101
GCTTAAGAAGGTGAAGAAGAGCAACGGTCAGATGCTGGCTATTAATGAGATTTTTGAGAAGAACCCAACAACAATCAAGAATTATGGTATTTGGCTCCGTT

@SRR10056916.5114348 5114348 length=101
TGAGTGGCGTCCTTGGATTCTGTCCCACACTTCTCTCGATTCCCCGGGTTCATCCTGAACATATAATGTGGAGAAATGCCATGAGGAAGGAGGCCAGTATT
```

.fastq

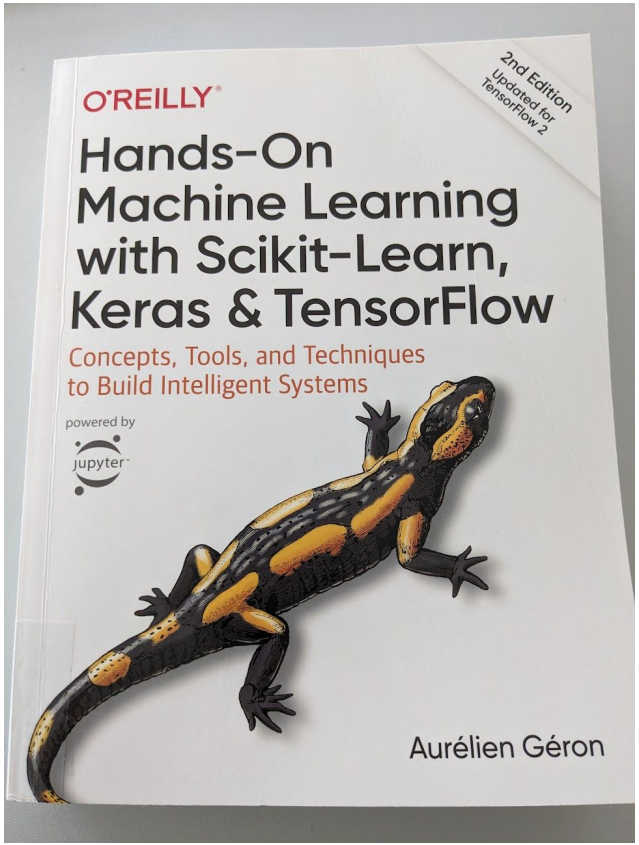
reference

```
ACGTTCTGGAGTGGCTGCAGCTCACCATGTATGCCTGAACCAGGCTGGCCTCATCCAAAG
CCATGCAGGATCACTCAGGATTTCAGTTTCACCCTCTATTTCCAAAGCCATTTACCTCAA
AGGACCCAGCAGCTACACCCCTACAGGCTTCCAGGCACCTCATCAGTCATGCTCCTCCTC
CATTTTACCCCCTACCCATCCTTGATCGGTCATGCCTAGCCTGACCCTTTAGTAAAGCAA
TGAGGTAGGAAGAACAAACCCTTGTCCCTTTGCCATGTGGAGGAAAGTGCCTGCCTCTGG
TCCGAGCCGCCTCGGTTCTGAAGCGAGTGCTCCTGCTTACCTTGCTCTAGGCTGTCTGCA
GAAGCACCTGCCGGTGGCACTCAGCACCTCCTTGTGCTAGAGCCCTCCATCACCTTCACG
CTGTCCCACCATGGGCCAGGAACCAAACCAGCACTGGGTTCTACTGCTGTGGGGTAAACT
AACTCAGTGGAATAGGGCTGGTTACTTTGGGCTGTCCAACTCATAAGTTTGGCTGCATTT
TGAAAAAAGCTGATCTAAATAAAGGCATGTGTATGGCTGGTCCCC
>ENST00000397163.8 cdna chromosome:GRCh38:15:42359501:42412317:1 gene:ENSG00000092529.25
gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:CAPN3 description:calpain 3
[Source:HGNC Symbol;Acc:HGNC:1480]
ACTCTCTTTCTCTCTCCCTCTGGCATGCATGCTGCTGGTAGGAGACCCCCAAGTCAACAT
TGCTTCAGAAATCCTTTAGCACTCATTTCTCAGGAGAACTTATGGCTTCAGAATCACAGC
TCGGTTTTTAAGATGGACATAACCTGTACGACCTTCTGATGGGCTTTCAACTTTGAACTG
GATGTGGACACTTTTCTCTCAGATGACAGAATTACTCCAACTTCCCCTTTGCAGTTGCTT
CCTTTCCTTGAAGGTAGCTGTATCTTATTTTCTTTAAAAAGCTTTTTCTTCCAAAGCCAC
TTGCCATGCCGACCGTCATTAGCGCATCTGTGGCTCCAAGGACAGCGGCTGAGCCCCGGT
CCCCAGGGCCAGTTCCTCACCCGGCCCAGAGCAAGGCCACTGAGGCTGGGGGTGGAAACC
CAAGTGGCATCTATTCAGCCATCATCAGCCGCAATTTTCCTATTATCGGAGTGAAAGAGA
AGACATTCGAGCAACTTCACAAGAAATGTCTAGAAAAGAAAGTTCTTTATGTGGACCCTG
AGTTCCCACCGGATGAGACCTCTCTCTTTTATAGCCAGAAGTTCCCCATCCAGTTCGTCT
GGAAGAGACCTCCGGAAATTTGCGAGAATCCCCGATTTATCATTGATGGAGCCAACAGAA
CTGACATCTGTCAAGGAGAGCTAGGGGACTGCTGGTTTCTCGCAGCCATTGCCTGCCTGA
CCCTGAACCAGCACCTTCTTTTCCGAGTCATACCCCATGATCAAAGTTTCATCGAAAACT
```

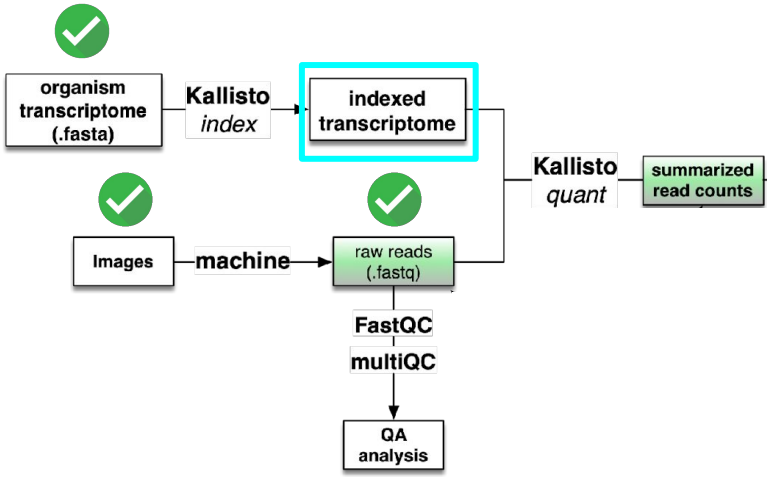
.fasta



“tensor arrays”



RNAseq analysis pipeline



~/Documents/my_study

- **scripts**
- **genome**
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
- **raw_reads**
  - SRR2006793.fastq.gz
  - SRR2006797.fastq.gz
  - SRR2006792.fastq.gz
  - SRR2006794.fastq.gz
  - etc…
- **publication**
  - manuscript.pdf
- **QC**
- **mapped_reads**

Building an index

Use Kallisto to build index from reference fasta

```
kallisto index -i inputFastaName.index inputFastaName.fa
```

Only need to build an index once

Be careful with file names!

- *Long and meaningful file names are better than short ambiguous ones.*
- *No spaces. ‘my_file_name’, or ’myFileName’, or ‘my.file.name’….just not, ‘my file name’*

Let’s construct an index from our reference fasta file

Building an index

Use Kallisto to build index from reference fasta

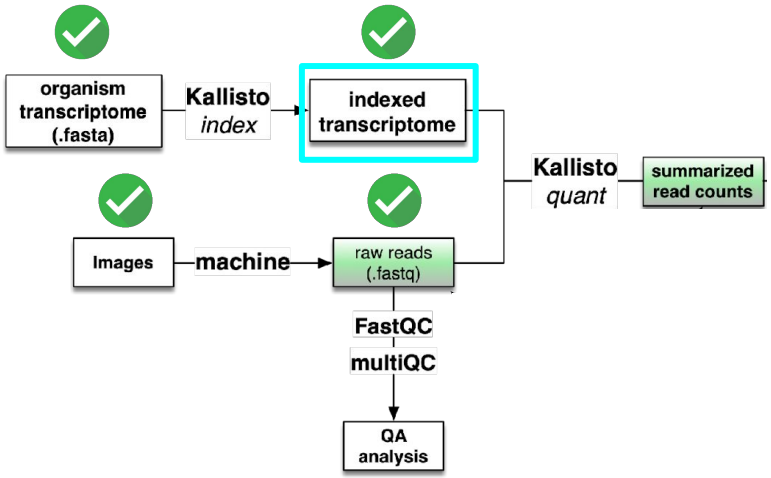
```
kallisto index -i inputFastaName.index inputFastaName.fa
```

```
kallisto index -i Solanum_lycopersicum.SL3.0.cdna.all.fa.index Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
```

No spaces: "my_file_name", or "myFileName", or "my.file.name"...just not, "my file name"

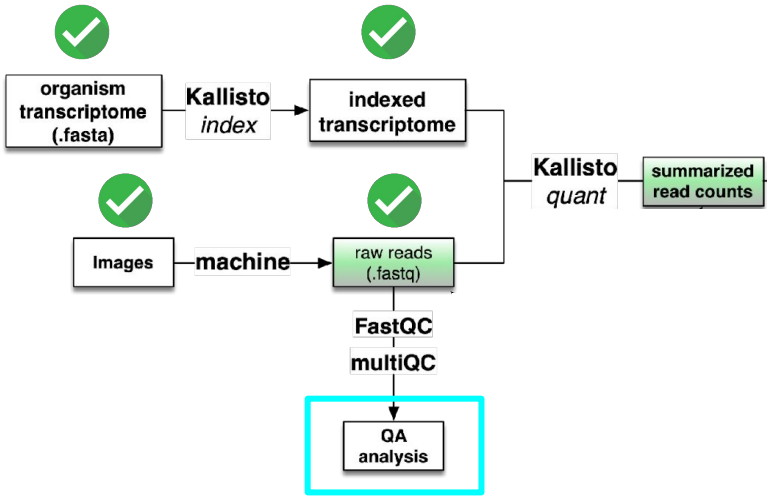
Let's construct an index from our reference fasta file

RNAseq analysis pipeline



~/Documents/my_study
- scripts
- genome
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.index
- raw_reads
  - SRR2006793.fastq.gz
  - SRR2006797.fastq.gz
  - SRR2006792.fastq.gz
  - SRR2006794.fastq.gz
  - etc…
- publication
  - manuscript.pdf
- QC
- mapped_reads

RNAseq analysis pipeline



~/Documents/my_study
- **scripts**
- **genome**
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.index
- **raw_reads**
  - SRR2006793.fastq.gz
  - SRR2006797.fastq.gz
  - SRR2006792.fastq.gz
  - SRR2006794.fastq.gz
  - etc…
- **publication**
  - manuscript.pdf
- **QC**
- **mapped_reads**

High-throughput sequencing data is stored in .fastq format

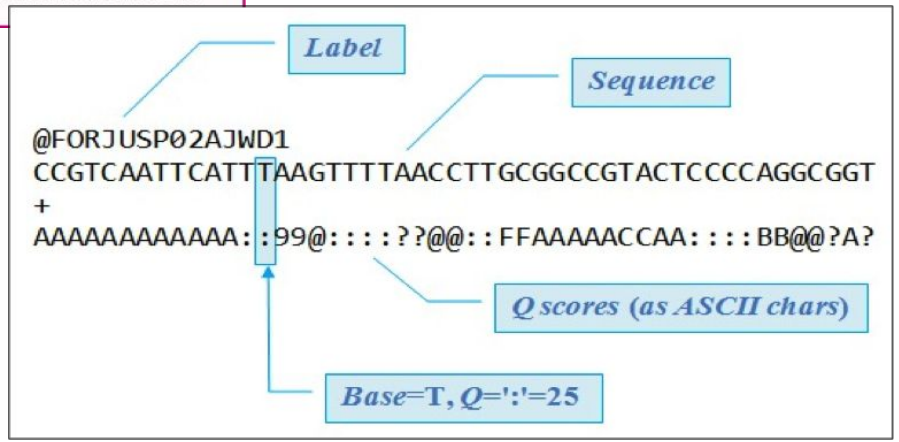
Fasta

Header
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

These are just text files

Fastq



ASCII characters

!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H	I
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Quality score

ASCII characters

!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H	I
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Quality score

1/1000

1/10000

ASCII characters

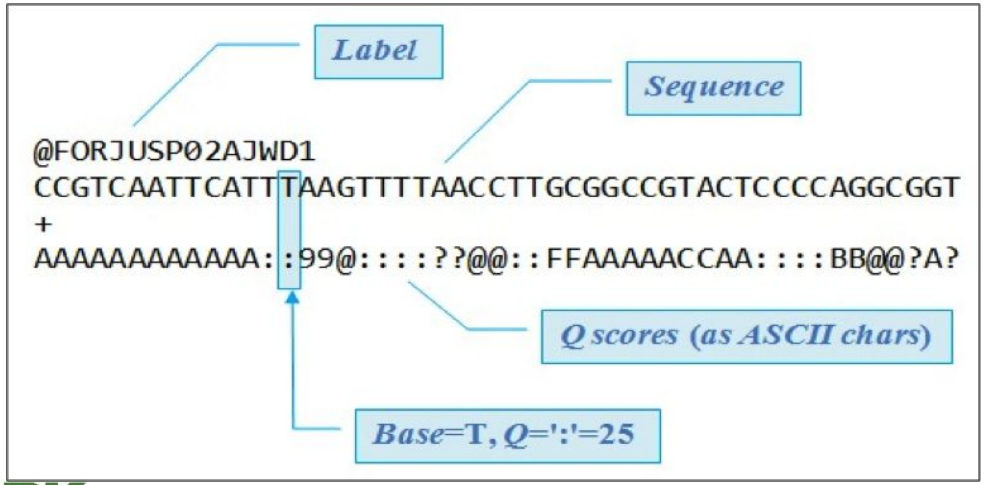
!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H	I
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

1/1000

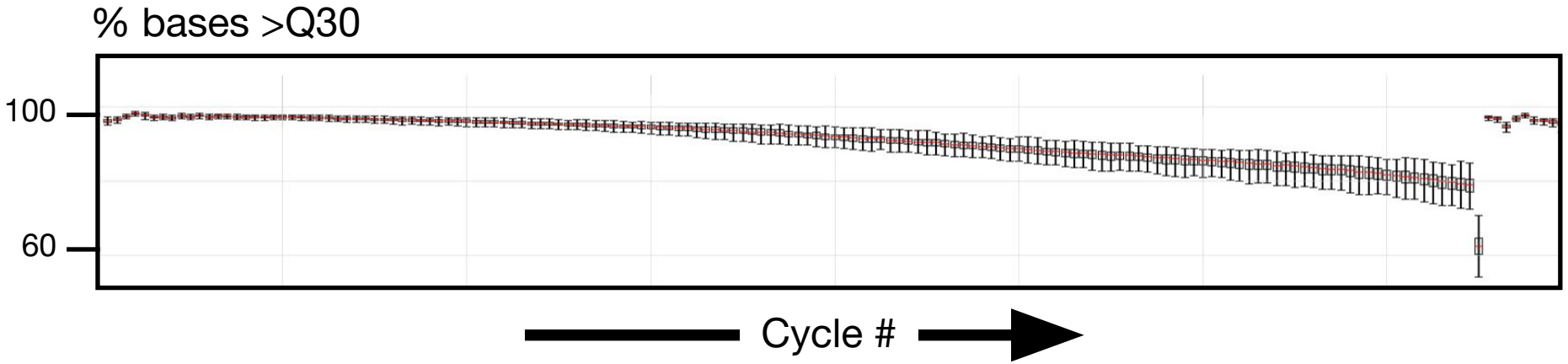
1/10000

Quality score

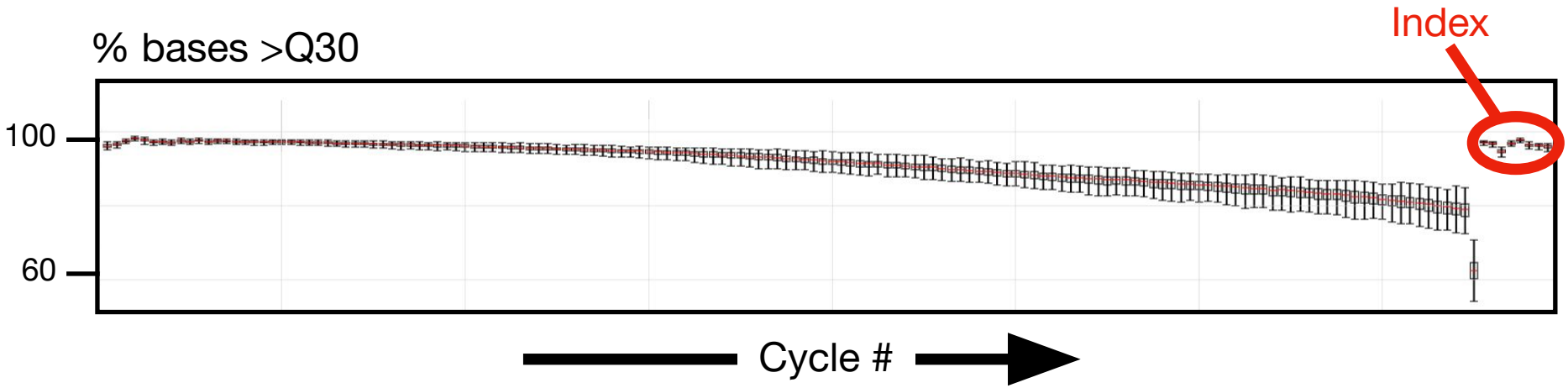
Fastq



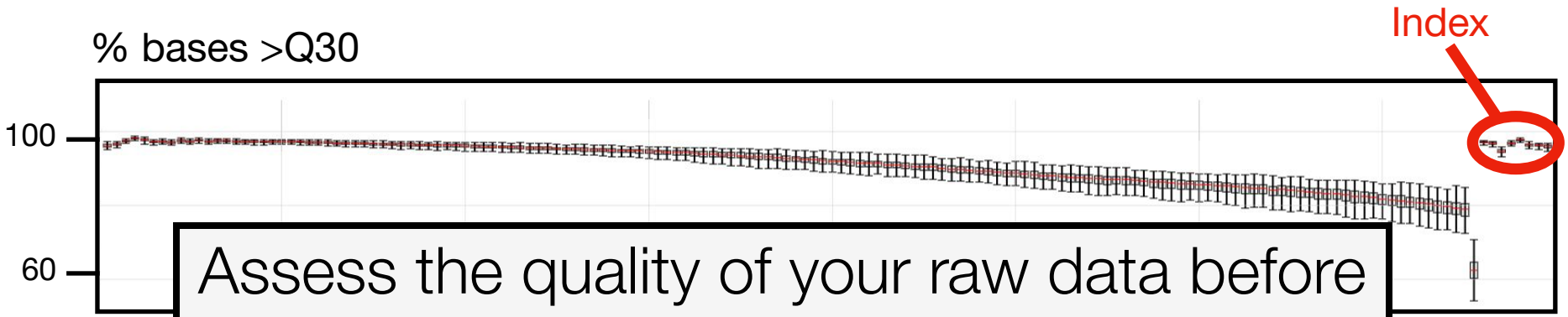
Illumina sequences are short for a reason



Illumina sequences are short for a reason

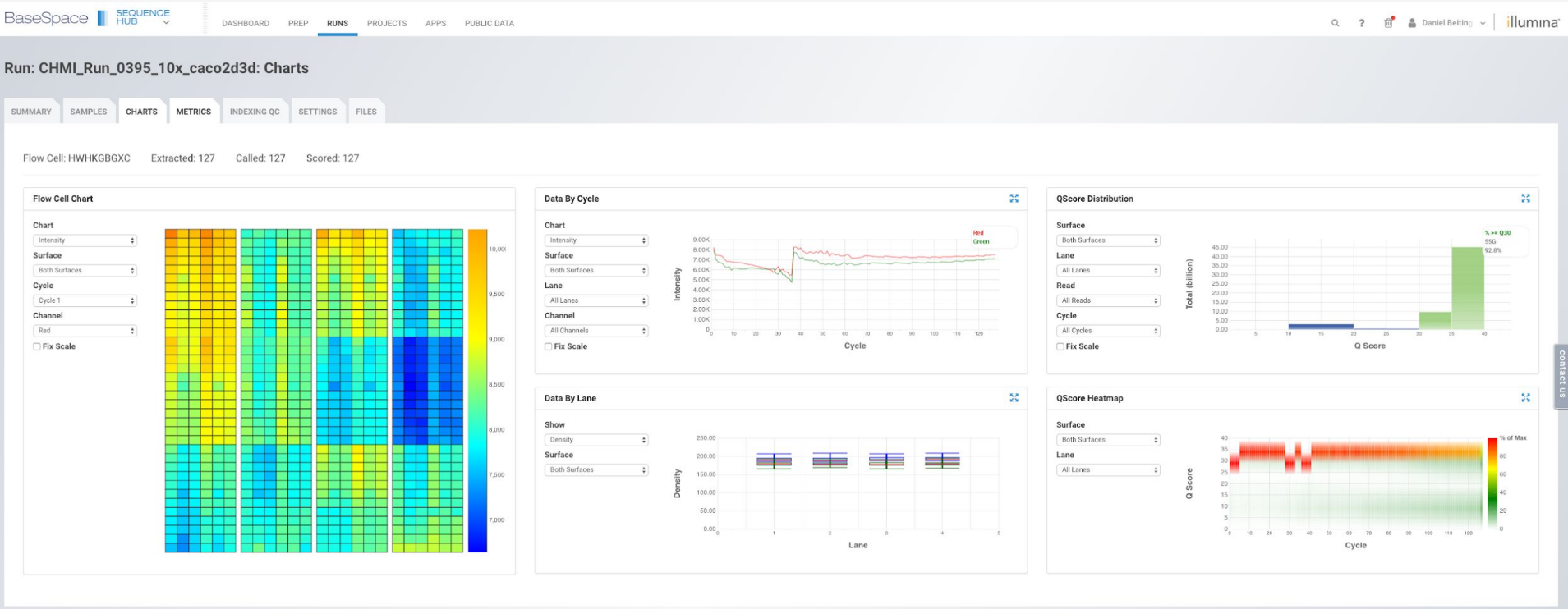


Illumina sequences are short for a reason



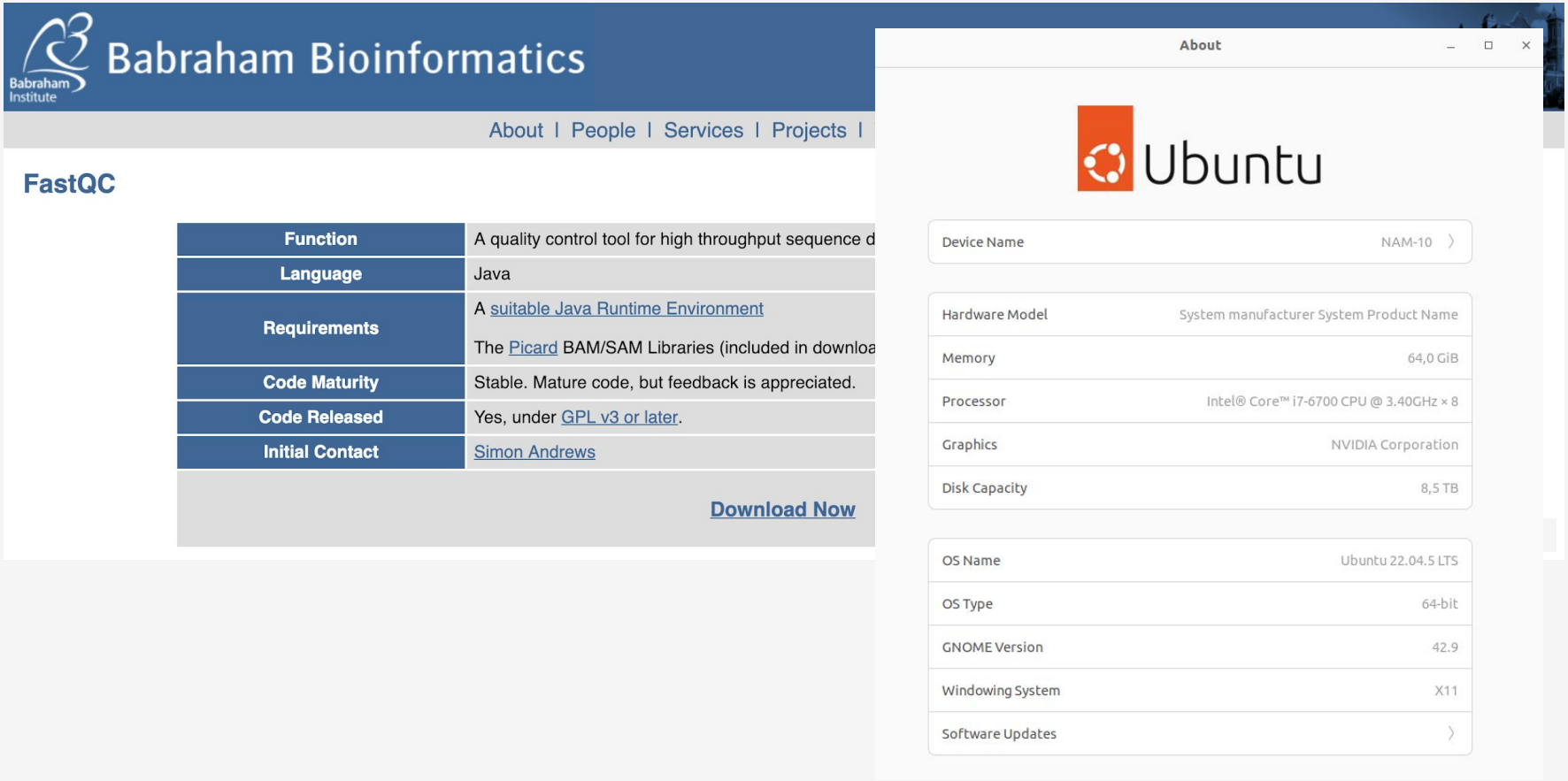
Assess the quality of your raw data before moving forward with analysis

Quality assessment via Illumina's BaseSpace

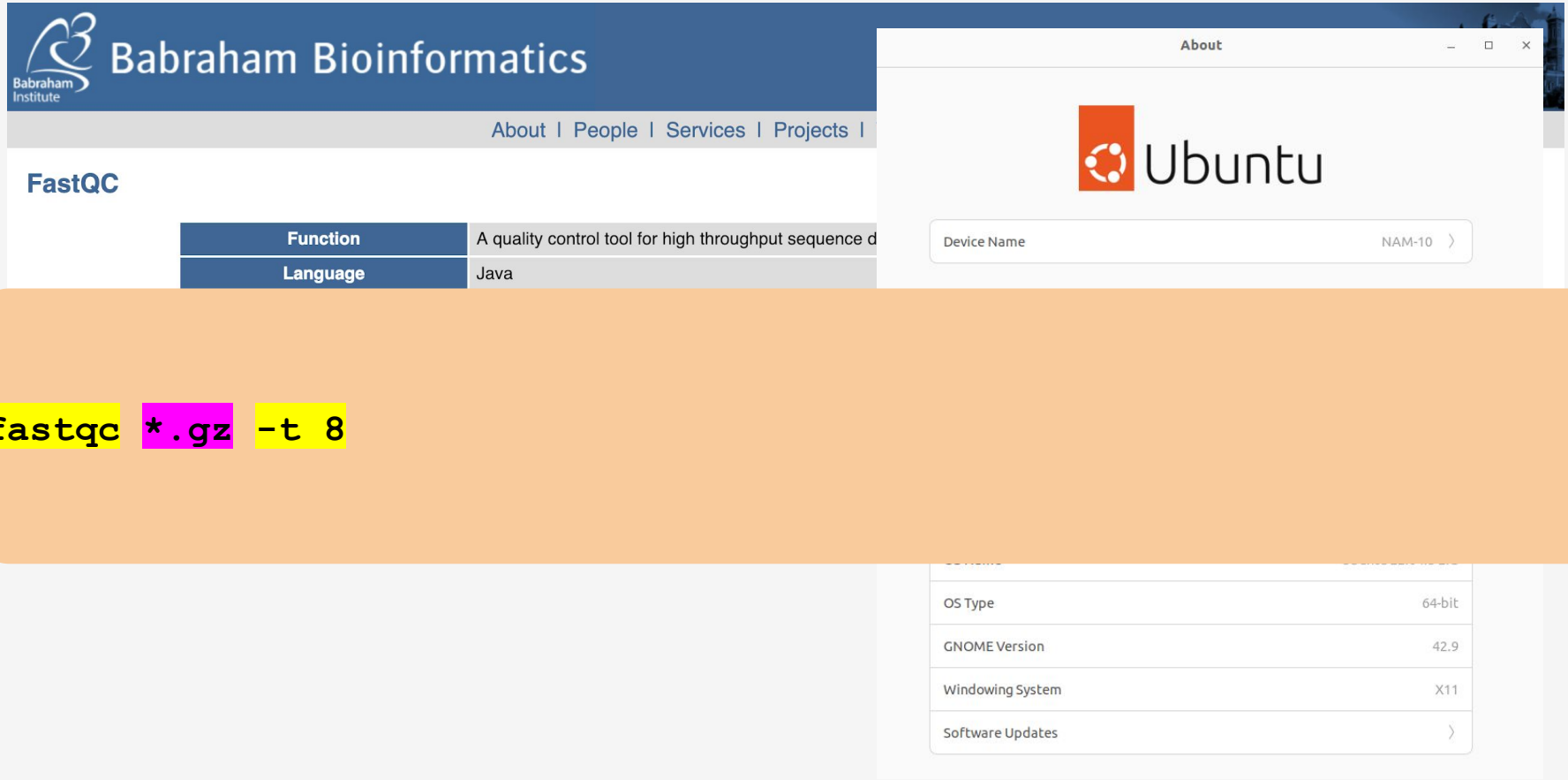


Quality assessment of fastq files using fastqc

Quality assessment of fastq files using fastqc

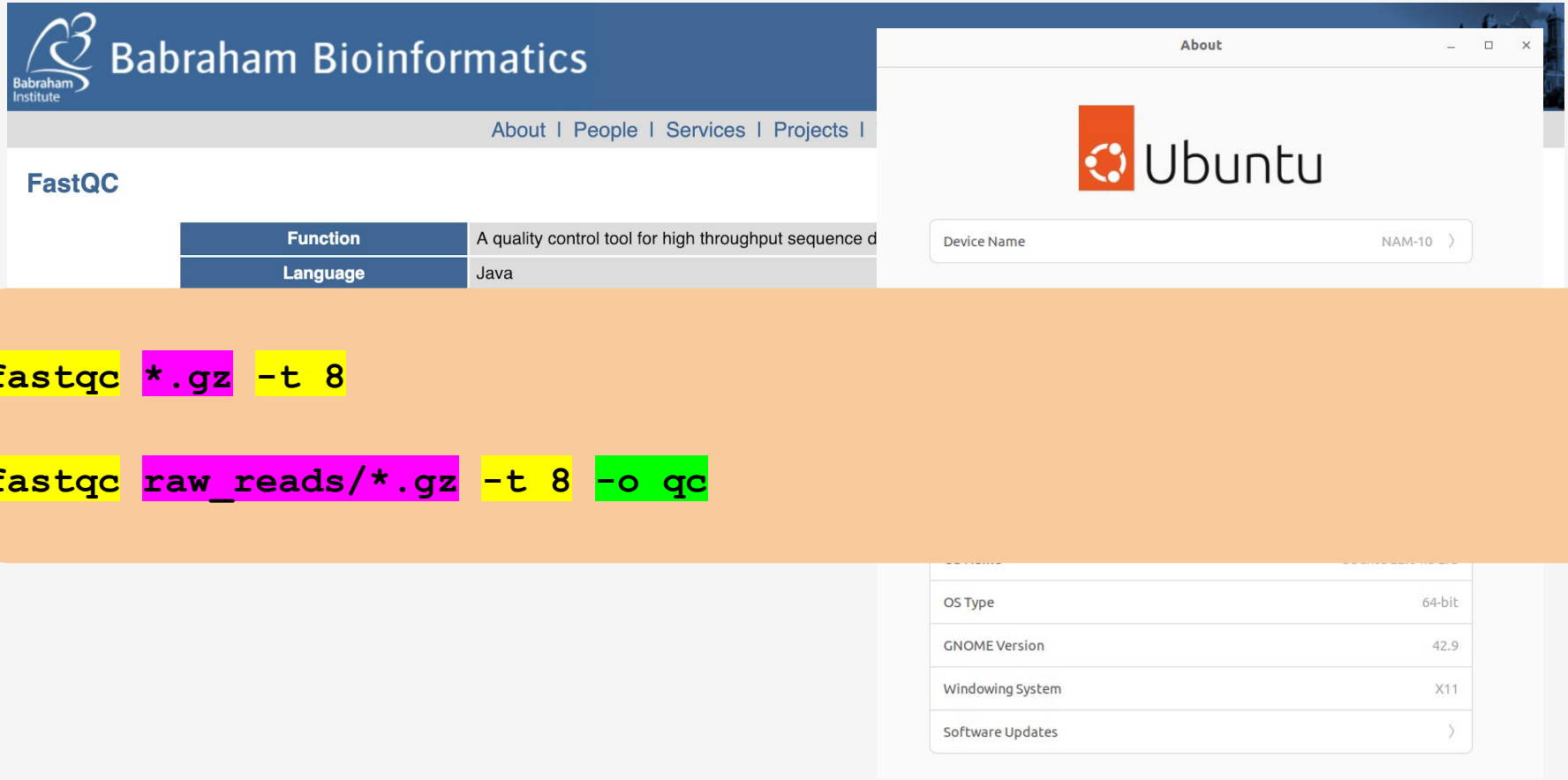


Quality assessment of fastq files using fastqc



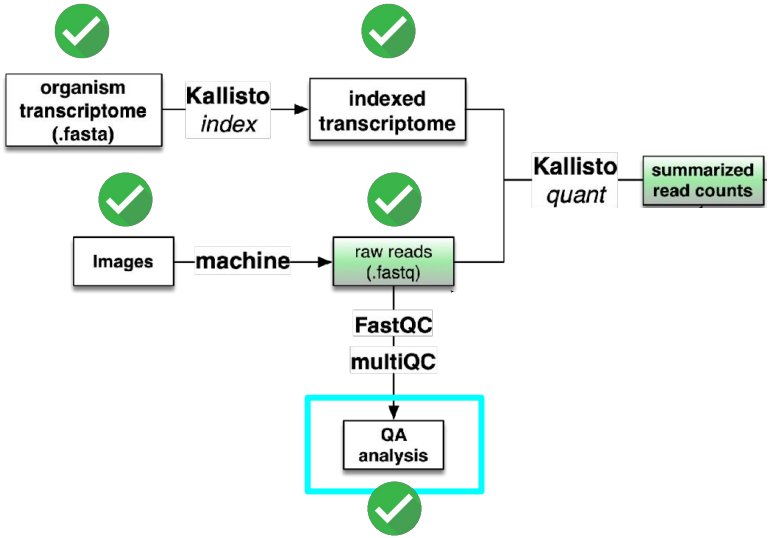

```
fastqc *.gz -t 8
```

Quality assessment of fastq files using fastqc



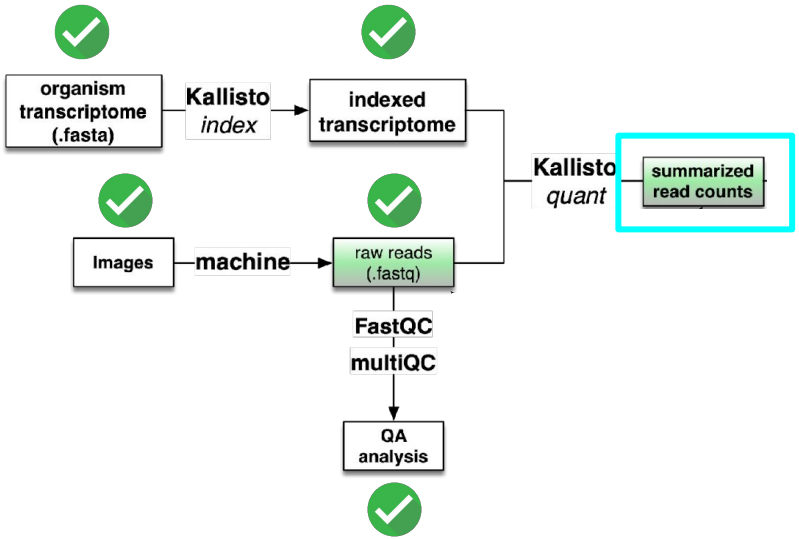

```
fastqc *.gz -t 8

fastqc raw_reads/*.gz -t 8 -o qc
```

RNAseq analysis pipeline



~/Documents/my_study
- scripts
- genome
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.index
- raw_reads
  - SRR2006793.fastq.gz
  - SRR2006797.fastq.gz
  - SRR2006792.fastq.gz
  - SRR2006794.fastq.gz
  - etc…
- publication
  - manuscript.pdf
- QC
  - qc_SRR2006793.html
  - etc….
- mapped_reads

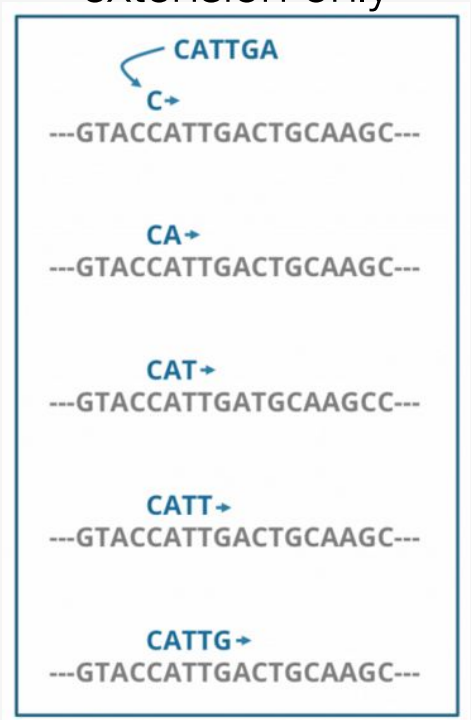
RNAseq analysis pipeline



~/Documents/my_study
- scripts
- genome
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
  - Solanum_lycopersicum.SL3.0.cdna.all.fa.index
- raw_reads
  - SRR2006793.fastq.gz
  - SRR2006797.fastq.gz
  - SRR2006792.fastq.gz
  - SRR2006794.fastq.gz
  - etc…
- publication
  - manuscript.pdf
- QC
  - qc_SRR2006793.html
  - etc….
- mapped_reads

‘Seed and Extend’ alignments

extension only



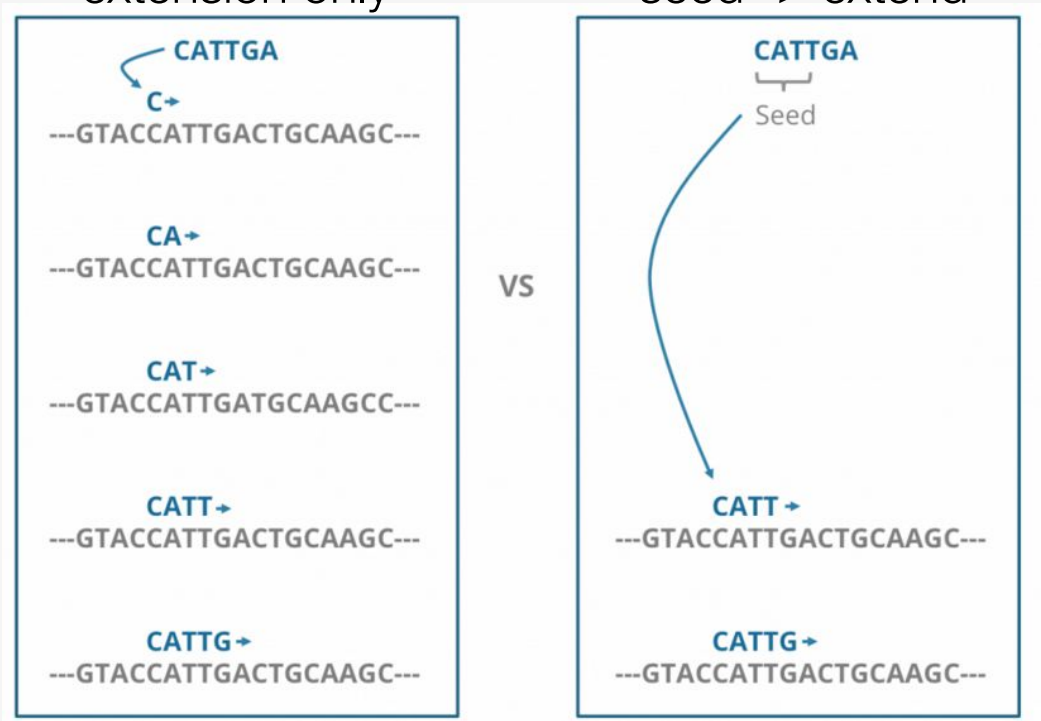
brute force, but slow

https://blog.sbgenomics.com/short-read-alignment-seeding/

‘Seed and Extend’ alignments

extension only

seed -> extend

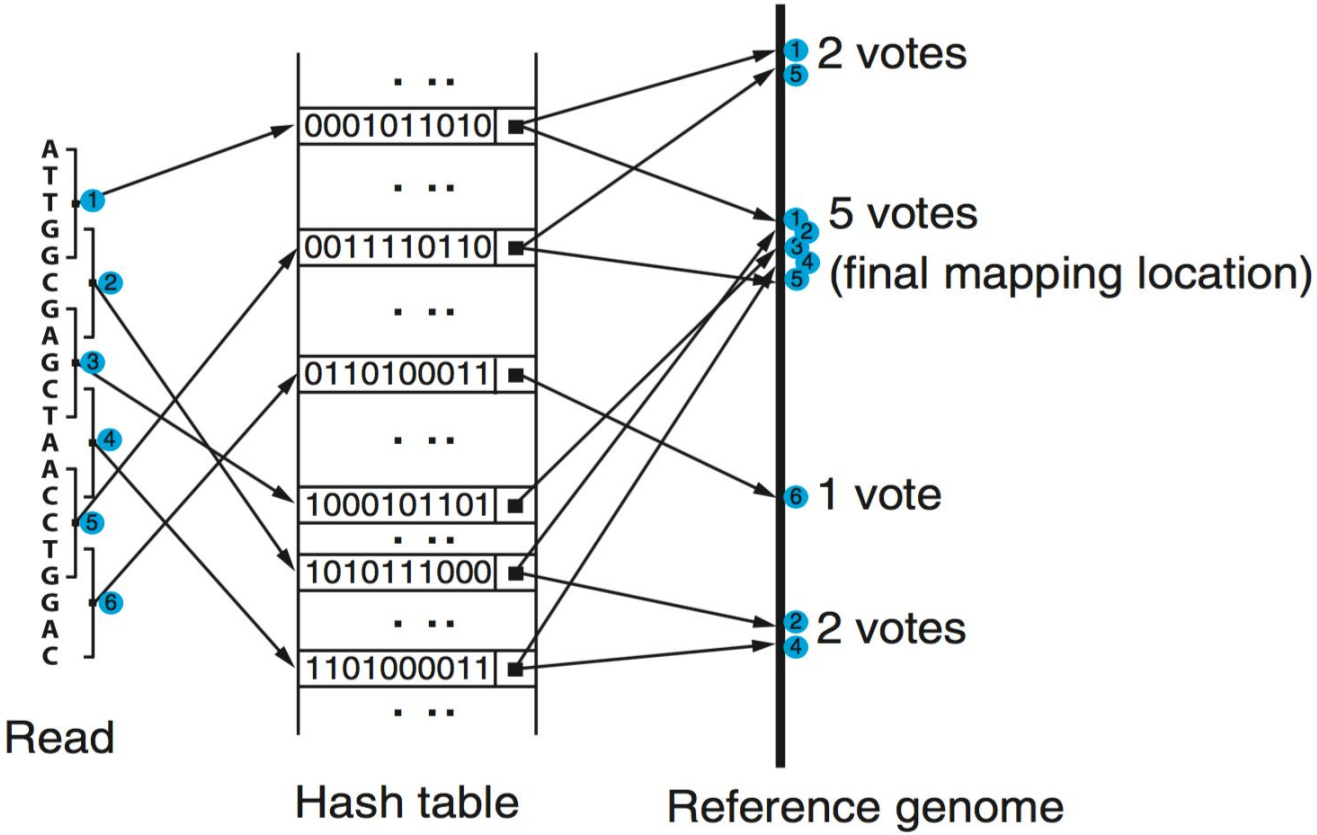


brute force, but slow

fast(er)

https://blog.sbgenomics.com/short-read-alignment-seeding/

R Subread aligner



Liao et al., NAR, 2013

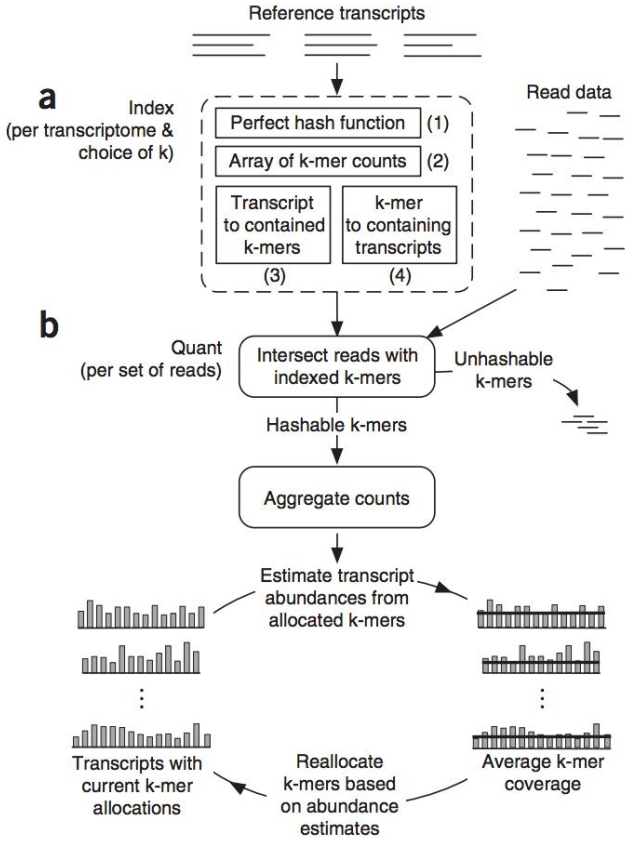
Reads are hard to align, but that's a function of their length. Why not approach the problem by shredding reads into Kmers - enter "Sailfish"

Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms

Rob Patro[1], Stephen M Mount[2,3] & Carl Kingsford[1]

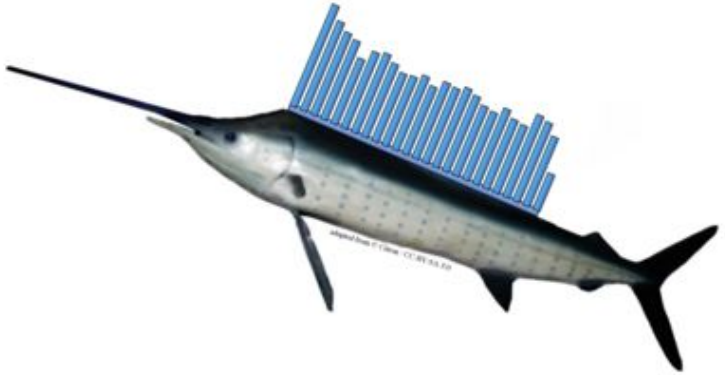
Nature Biotechnology **32**, 462–464 (2014)

Sailfish replaces *approximate* alignment of (error prone) *reads* with *exact* alignment of short *k-mers*



Sailfish

'lightweight read alignment'

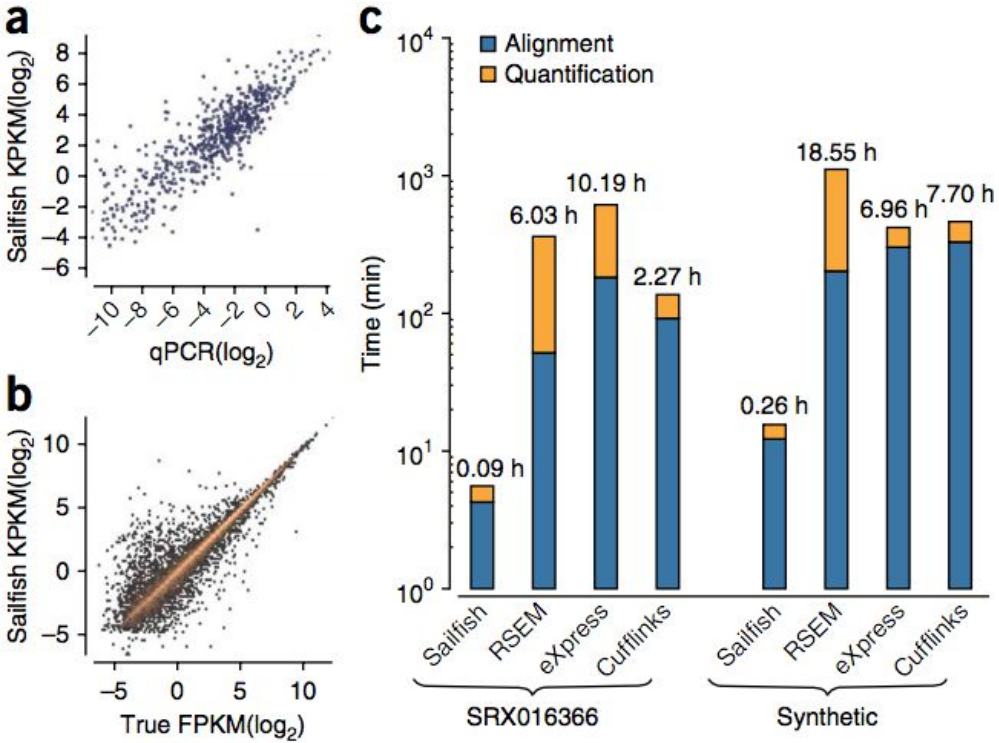


"By not requiring read mapping, **Sailfish avoids parameters specifying**, for example, the number of mismatches to tolerate, total allowable quality of mismatched bases, gap open and extension penalties, whether and how much to trim reads, number and quality of alignments to report from the aligner and pass into the estimation procedure."

Patro et al., Nat. Biotech., 2014

Sailfish

'lightweight read alignment'



Patro et al., Nat. Biotech., 2014

What if the idea of alignment was altogether **abandoned** enter “Kallisto”

nature biotechnology

Nature Biotechnology **34**, 525–527 (2016)

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray[1], Harold Pimentel[2], Páll Melsted[3] & Lior Pachter[2,4,5]

We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. We use kallisto to analyze 30 million unaligned paired-end RNA-seq reads in <10 min on a standard laptop computer. This removes a major computational bottleneck in RNA-seq analysis.

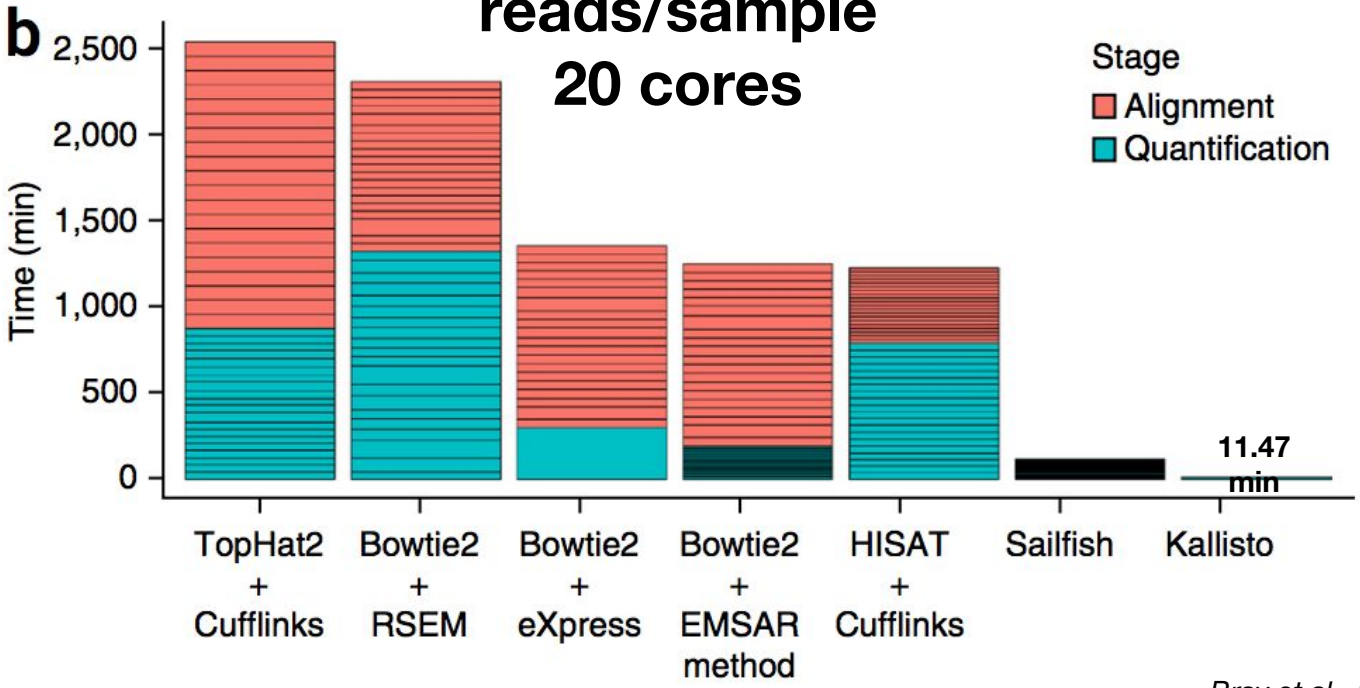
The first two steps in typical transcript-level RNA-seq processing

this information, we develop a method based on pseudoalignment of reads and fragments, which focuses only on identifying the transcripts from which the reads could have originated and does not try to pinpoint exactly how the sequences of the reads and transcripts align.

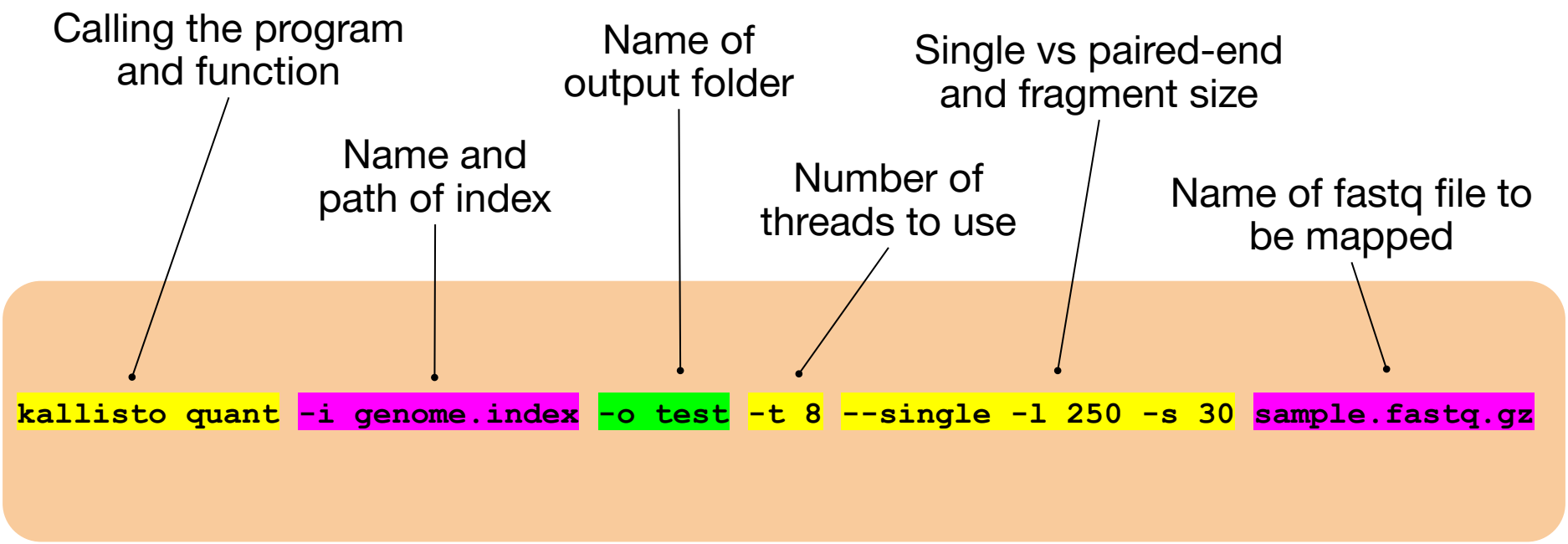
A pseudoalignment of a read to a set of transcripts, T, is a subset, $S \subseteq T$, without specific coordinates mapping each base in the read to specific positions in each of the transcripts in S. Accurate pseudoalignments of reads to a transcriptome can be obtained using fast hashing of k-mers together with the transcriptome de Bruijn graph (T-DBG). de Bruijn graphs have been crucial for DNA and RNA assembly[8], where they are usually constructed from reads. Kallisto uses a T-DBG, which is a de Bruijn graph constructed from k-mers present in the transcriptome (**Fig. 1a**), and a path covering of the graph, a set of paths whose union covers all edges of the graph, where the paths correspond to transcripts (**Fig. 1b**). This path covering of a T-DBG induces multi-sets on the vertices, called k-compatibility classes. A compatibility class can be associated to an error-free read by representing it as a path in the graph and defining the k-compatibility

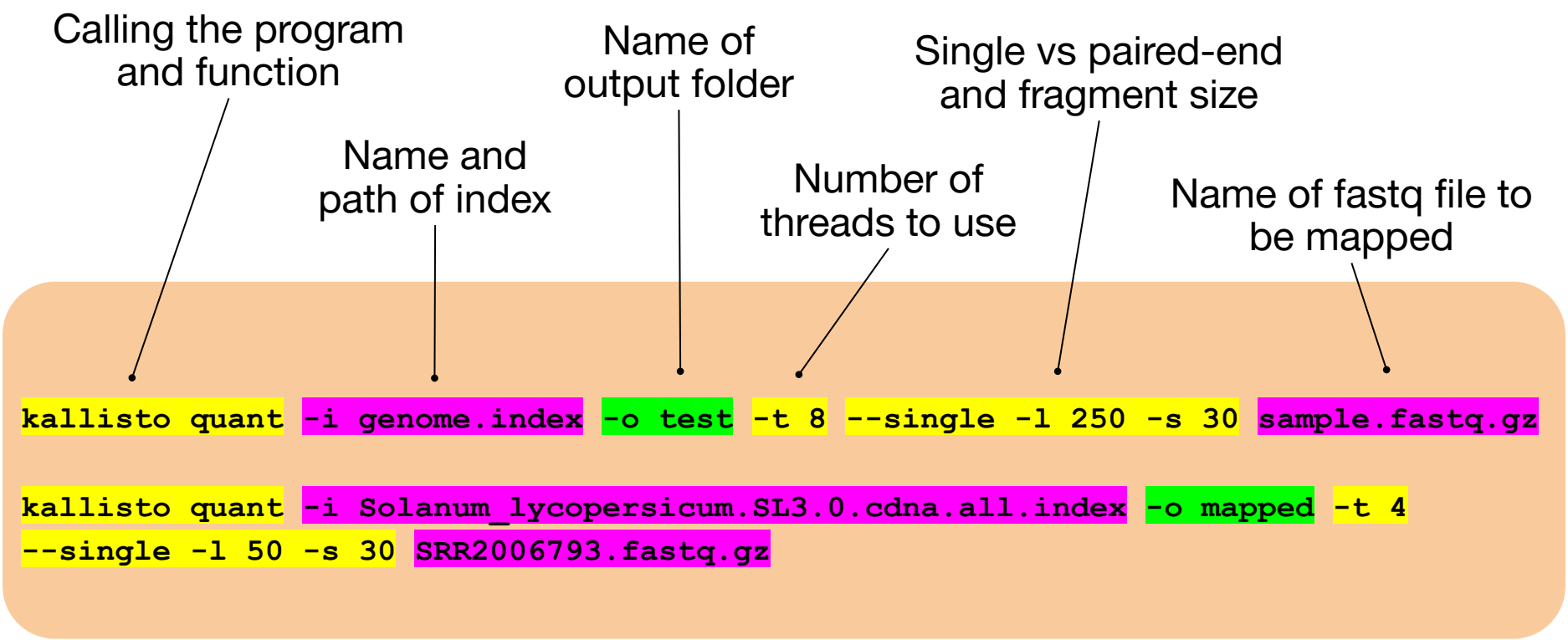
Kallisto - it's fast

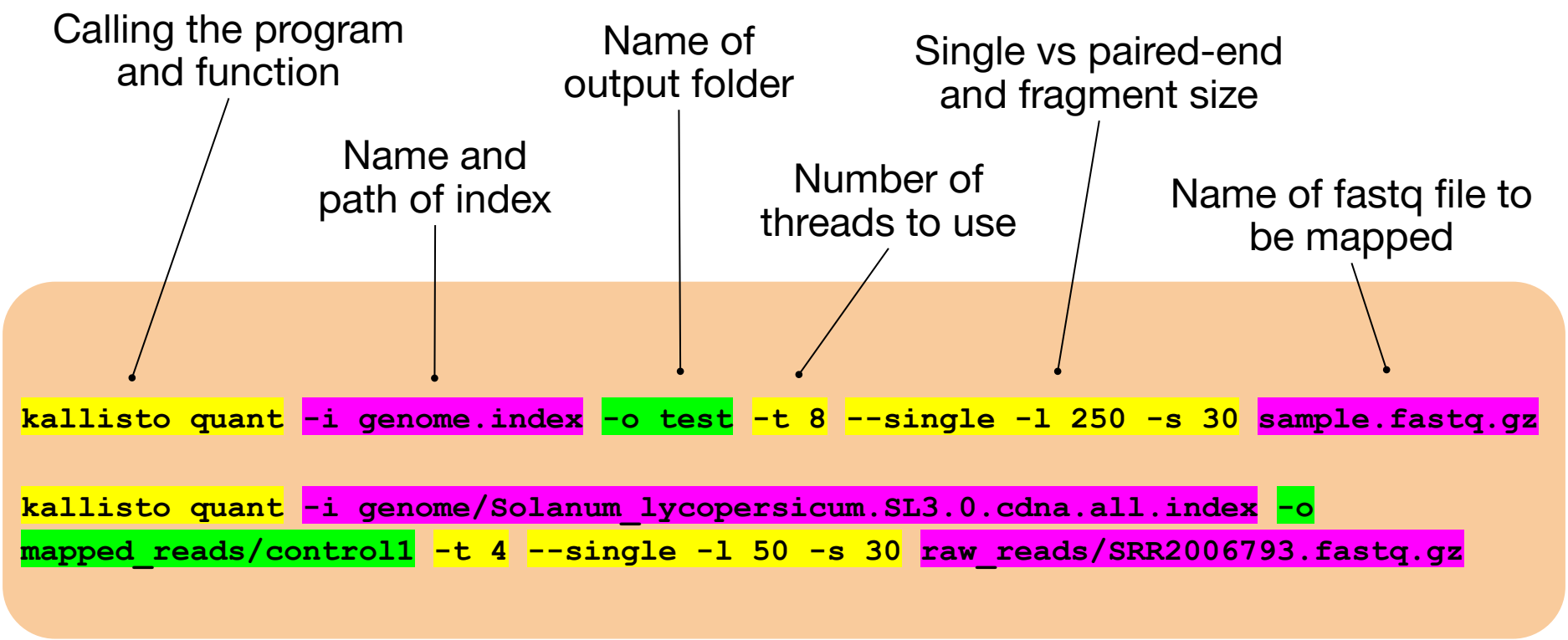
20 samples
30 million PE reads/sample
20 cores

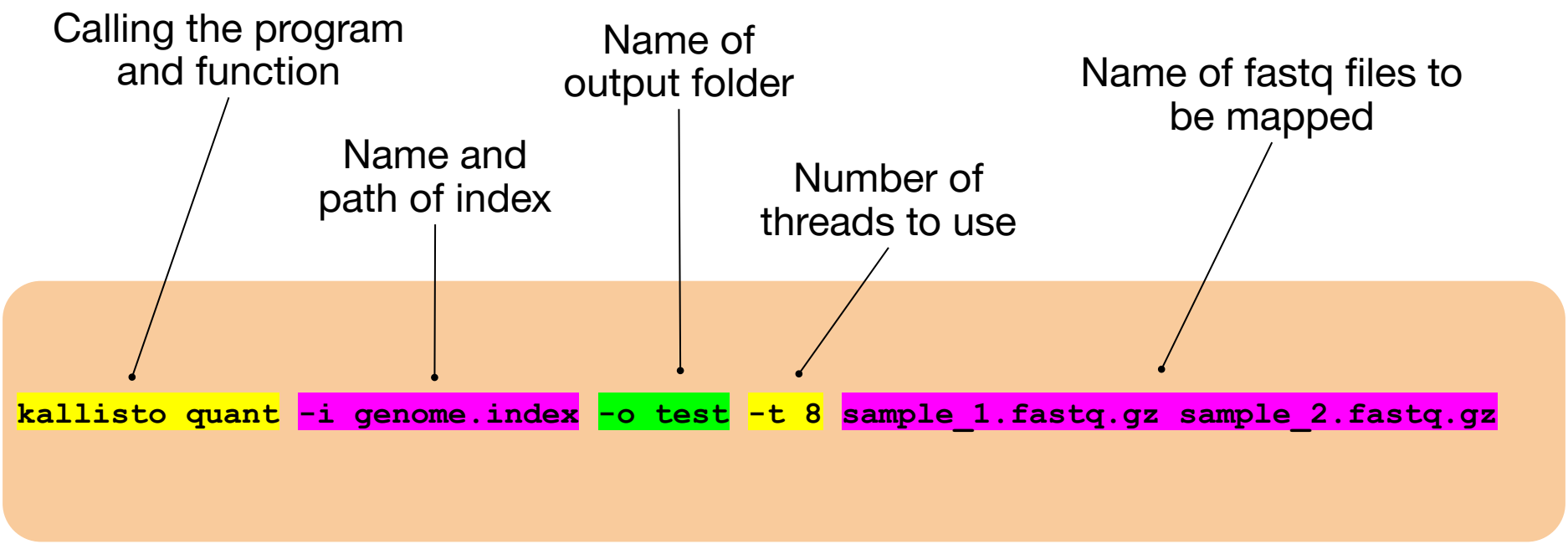


Bray et al., Nat. Biotech., 2016

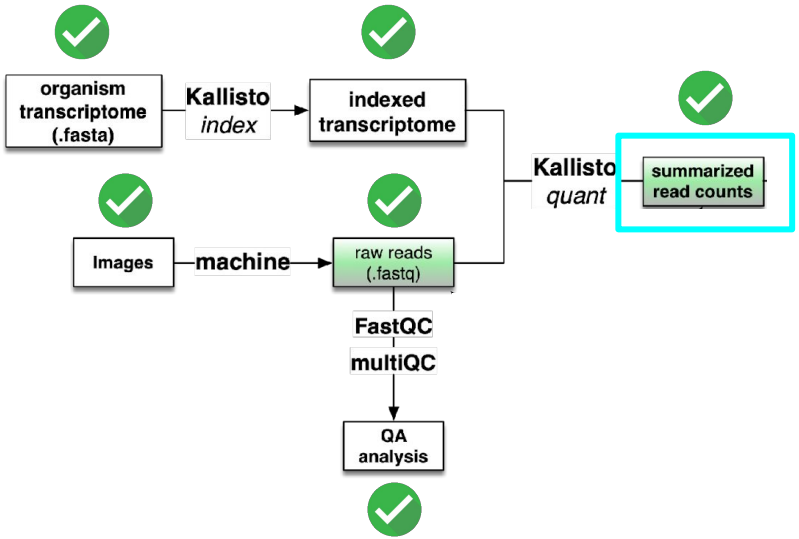

Calling the program and function
Name and path of index
Name of output folder
Number of threads to use
Single vs paired-end and fragment size
Name of fastq file to be mapped
kallisto quant -i genome.index -o test -t 8 --single -l 250 -s 30 sample.fastq.gz


Calling the program and function
Name and path of index
Name of output folder
Number of threads to use
Single vs paired-end and fragment size
Name of fastq file to be mapped
kallisto quant -i genome.index -o test -t 8 --single -l 250 -s 30 sample.fastq.gz
kallisto quant -i Solanum_lycopersicum.SL3.0.cdna.all.index -o mapped -t 4 --single -l 50 -s 30 SRR2006793.fastq.gz

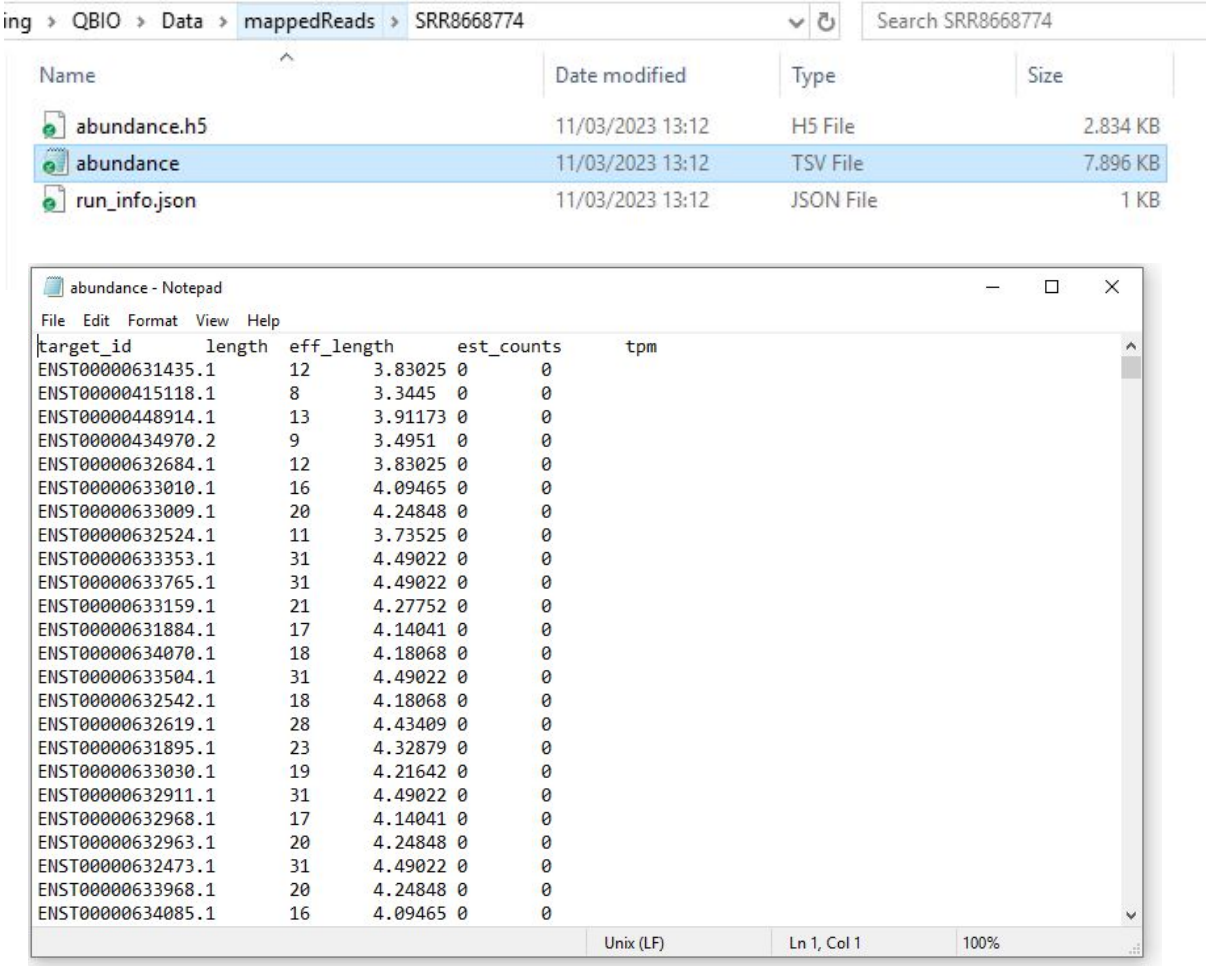

Calling the program and function
Name and path of index
Name of output folder
Number of threads to use
Single vs paired-end and fragment size
Name of fastq file to be mapped
kallisto quant -i genome.index -o test -t 8 --single -l 250 -s 30 sample.fastq.gz
kallisto quant -i genome/Solanum_lycopersicum.SL3.0.cdna.all.index -o mapped_reads/control1 -t 4 --single -l 50 -s 30 raw_reads/SRR2006793.fastq.gz

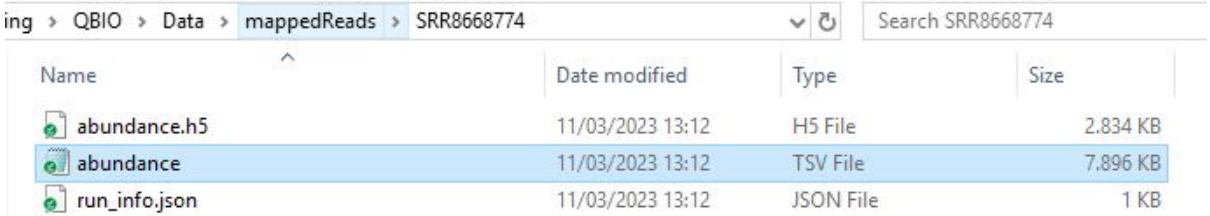

Calling the program and function
Name and path of index
Name of output folder
Number of threads to use
Name of fastq files to be mapped
kallisto quant -i genome.index -o test -t 8 sample_1.fastq.gz sample_2.fastq.gz

RNAseq analysis pipeline




```
~/Documents/my_study
├── scripts
├── genome
│   ├── Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
│   └── Solanum_lycopersicum.SL3.0.cdna.all.fa.index
├── raw_reads
│   ├── SRR2006793.fastq.gz
│   ├── SRR2006797.fastq.gz
│   ├── SRR2006792.fastq.gz
│   ├── SRR2006794.fastq.gz
│   └── etc...
├── publication
│   └── manuscript.pdf
├── QC
│   ├── qc_SRR2006793.html
│   └── etc....
└── mapped_reads
    ├── control1
    │   ├── abundance.h5
    │   ├── abundance.tsv
    │   └── run_info.json
    ├── control2
    │   ├── abundance.h5
    │   ├── abundance.tsv
    │   └── run_info.json
    └── etc...
        ├── abundance.h5
        ├── abundance.tsv
        └── run_info.json
```


ing › QBIO › Data › mappedReads › SRR8668774
Search SRR8668774
Name
Date modified
Type
Size
abundance.h5
11/03/2023 13:12
H5 File
2.834 KB
abundance
11/03/2023 13:12
TSV File
7.896 KB
run_info.json
11/03/2023 13:12
JSON File
1 KB
abundance - Notepad
File Edit Format View Help
target_id length eff_length est_counts tpm
ENST00000631435.1 12 3.83025 0 0
ENST00000415118.1 8 3.3445 0 0
ENST00000448914.1 13 3.91173 0 0
ENST00000434970.2 9 3.4951 0 0
ENST00000632684.1 12 3.83025 0 0
ENST00000633010.1 16 4.09465 0 0
ENST00000633009.1 20 4.24848 0 0
ENST00000632524.1 11 3.73525 0 0
ENST00000633353.1 31 4.49022 0 0
ENST00000633765.1 31 4.49022 0 0
ENST00000633159.1 21 4.27752 0 0
ENST00000631884.1 17 4.14041 0 0
ENST00000634070.1 18 4.18068 0 0
ENST00000633504.1 31 4.49022 0 0
ENST00000632542.1 18 4.18068 0 0
ENST00000632619.1 28 4.43409 0 0
ENST00000631895.1 23 4.32879 0 0
ENST00000633030.1 19 4.21642 0 0
ENST00000632911.1 31 4.49022 0 0
ENST00000632968.1 17 4.14041 0 0
ENST00000632963.1 20 4.24848 0 0
ENST00000632473.1 31 4.49022 0 0
ENST00000633968.1 20 4.24848 0 0
ENST00000634085.1 16 4.09465 0 0
Unix (LF)
Ln 1, Col 1
100%



abundance - Notepad

File Edit Format View Help

target_id	length	eff_length	est_counts	tpm
ENST00000631435.1	12	3.83025	0	0
ENST00000415118.1	8	3.3445	0	0
ENST00000448914.1	13	3.91173	0	0
ENST00000434970.2	9	3.4951	0	0
ENST00000632684.1	12	3.83025	0	0
ENST00000633010.1	16	4.09465	0	0
ENST00000633009.1	20	4.24848	0	0
ENST00000632524.1	11	3.73525	0	0
ENST00000633353.1	31	4.49022	0	0
ENST00000633765.1	31	4.49022	0	0
ENST00000633159.1	21	4.27752	0	0
ENST00000631884.1	17	4.14041	0	0
ENST00000634070.1	18	4.18068	0	0
ENST00000633504.1	31	4.49022	0	0
ENST00000632542.1	18	4.18068	0	0
ENST00000632619.1	28	4.43409	0	0
ENST00000631895.1	23	4.32879	0	0
ENST00000633030.1	19	4.21642	0	0
ENST00000632911.1	31	4.49022	0	0
ENST00000632968.1	17	4.14041	0	0
ENST00000632963.1	20	4.24848	0	0
ENST00000632473.1	31	4.49022	0	0
ENST00000633968.1	20	4.24848	0	0
ENST00000634085.1	16	4.09465	0	0

Unix (LF)

abundance - Notepad

File Edit Format View Help

ENST00000371754.8	5235	4986	1208.55	4.16607
ENST00000469991.1	640	391	14.5506	0.639615
ENST00000371752.5	7212	6963	6866.66	16.9498
ENST00000396105.6	7209	6960	12876.7	31.7988
ENST00000371744.5	2964	2715	1972.76	12.4887
ENST00000455070.1	2122	1873	34.6566	0.318026
ENST00000509231.1	2224	1975	0	0
ENST00000646235.1	3341	3092	0	0
ENST00000644105.2	2821	2572	0	0
ENST00000646118.1	3411	3162	3	0.016307
ENST00000647304.1	3707	3458	1075.39	5.3451
ENST00000395409.7	2192	1943	0	0
ENST00000540171.2	1307	1058	2	0.0324907
ENST00000675743.1	648	399	0	0
ENST00000674723.1	931	682	0	0
ENST00000674869.1	837	588	0	0
ENST00000675248.1	577	328	0	0
ENST00000676427.1	883	634	32.2398	0.874013
ENST00000330634.11	7578	7329	289.566	0.679076
ENST00000392634.9	7623	7374	0.956435	0.00222929
ENST00000675482.1	1020	771	20.6712	0.460816
ENST00000398337.8	1689	1440	224.596	2.68074
ENST00000674966.1	859	610	8.63282	0.243242
ENST00000675207.1	7732	7483	58.9235	0.13534
ENST00000675616.1	431	182	0	0

Unix (LF) Ln 1, Col 1 100%

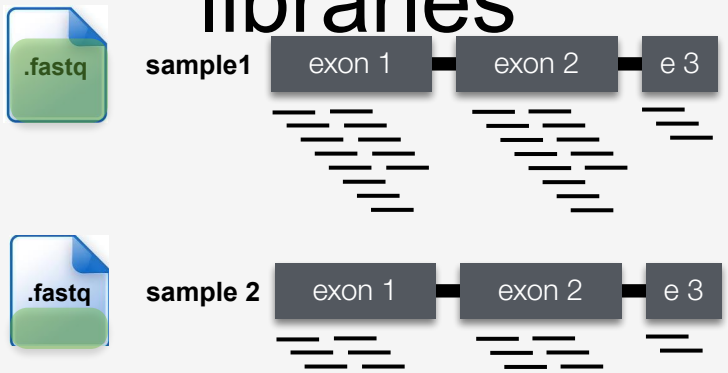
The ‘effective length’ of a transcript

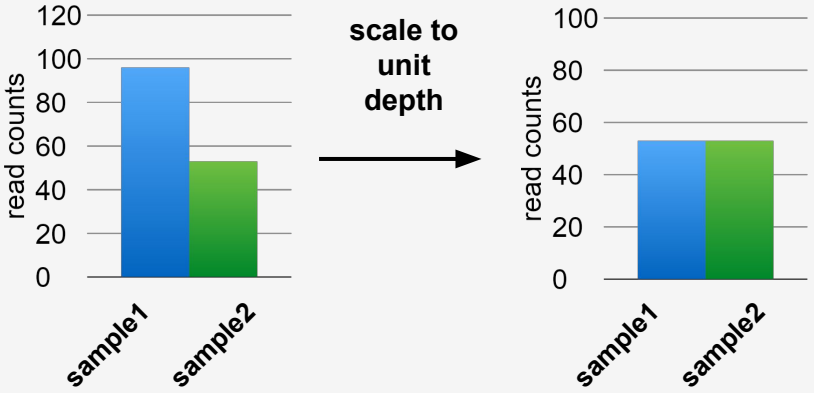
The length of a transcript after adjusting for the total number of possible positions a fragment of size X could originate from

$$L_{effective} = L_{actual} - L_{fragment} + 1$$

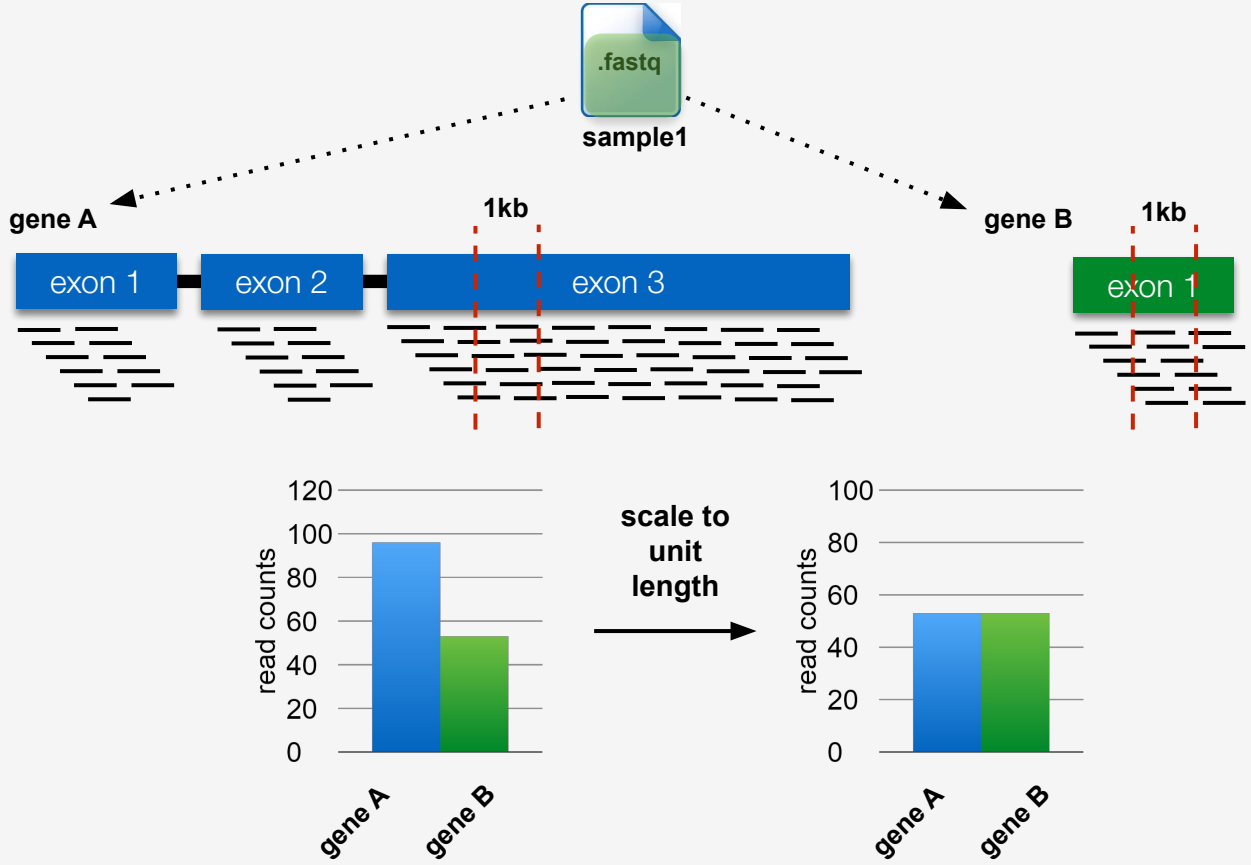
transcript	ATGCGTAACATG	$L_{actual} = 12$	$\mathbf{L_{effective} = 10}$
fragment	NNN	$L_{fragment} = 3$	

scaling units for between libraries

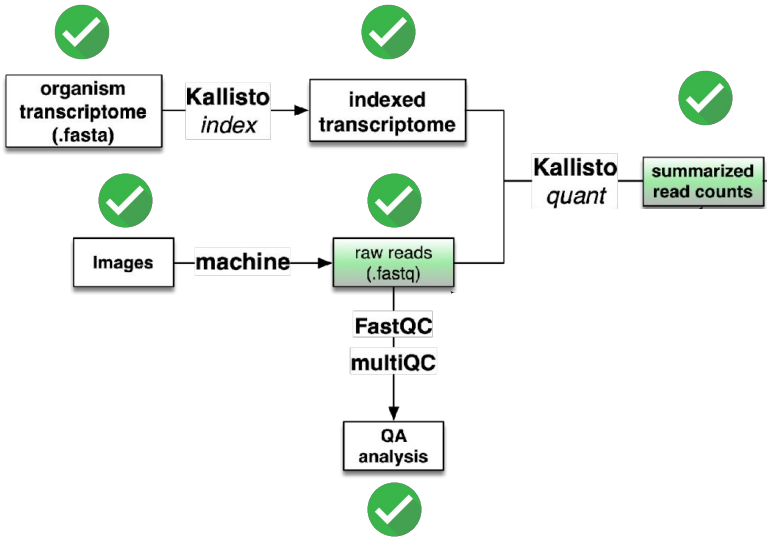




scaling units for within-sample comparisons



RNAseq analysis pipeline



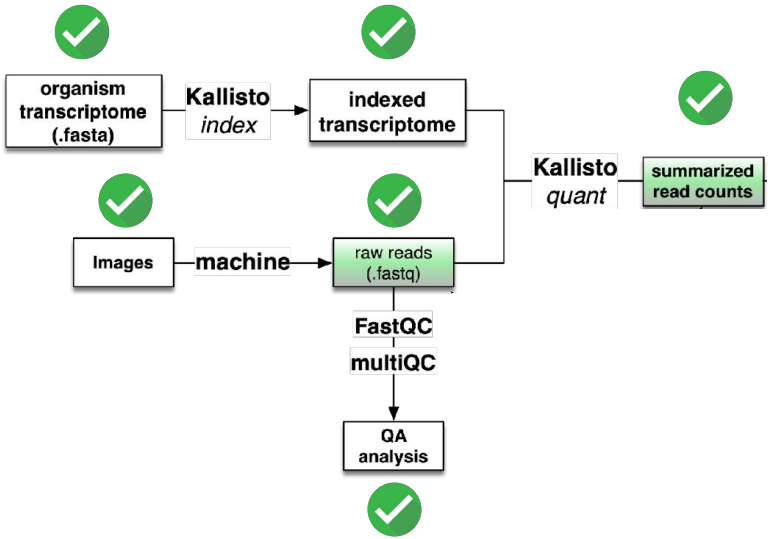
abundance - Notepad

File Edit Format View Help

ENST00000371754.8	5235	4986	1208.55	4.16607
ENST00000469991.1	640	391	14.5506	0.639615
ENST00000371752.5	7212	6963	6866.66	16.9498
ENST00000396105.6	7209	6960	12876.7	31.7988
ENST00000371744.5	2964	2715	1972.76	12.4887
ENST00000455070.1	2122	1873	34.6566	0.318026
ENST00000509231.1	2224	1975	0	0
ENST00000646235.1	3341	3092	0	0
ENST00000644105.2	2821	2572	0	0
ENST00000646118.1	3411	3162	3	0.016307
ENST00000647304.1	3707	3458	1075.39	5.3451
ENST00000395409.7	2192	1943	0	0
ENST00000540171.2	1307	1058	2	0.0324907
ENST00000675743.1	648	399	0	0
ENST00000674723.1	931	682	0	0
ENST00000674869.1	837	588	0	0
ENST00000675248.1	577	328	0	0
ENST00000676427.1	883	634	32.2398	0.874013
ENST00000330634.11	7578	7329	289.566	0.679076
ENST00000392634.9	7623	7374	0.956435	0.00222929
ENST00000675482.1	1020	771	20.6712	0.460816
ENST00000398337.8	1689	1440	224.596	2.68074
ENST00000674966.1	859	610	8.63282	0.243242
ENST00000675207.1	7732	7483	58.9235	0.13534
ENST00000675616.1	431	182	0	0

Unix (LF)

RNAseq analysis pipeline



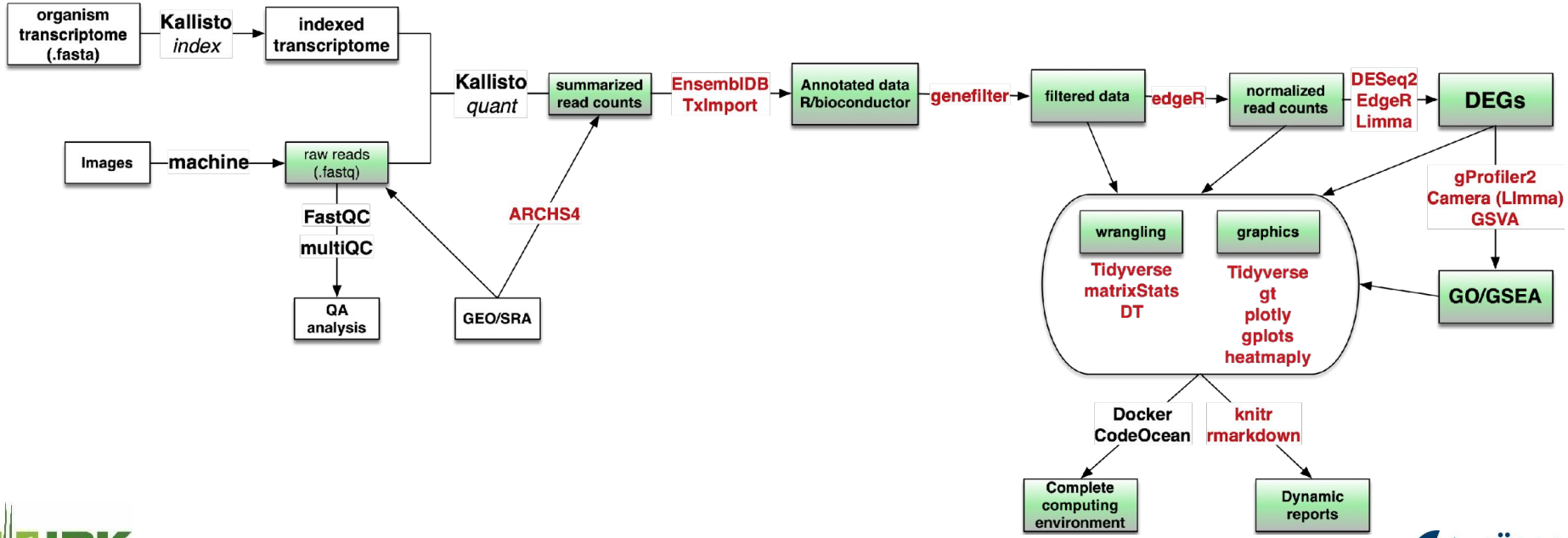
abundance - Notepad

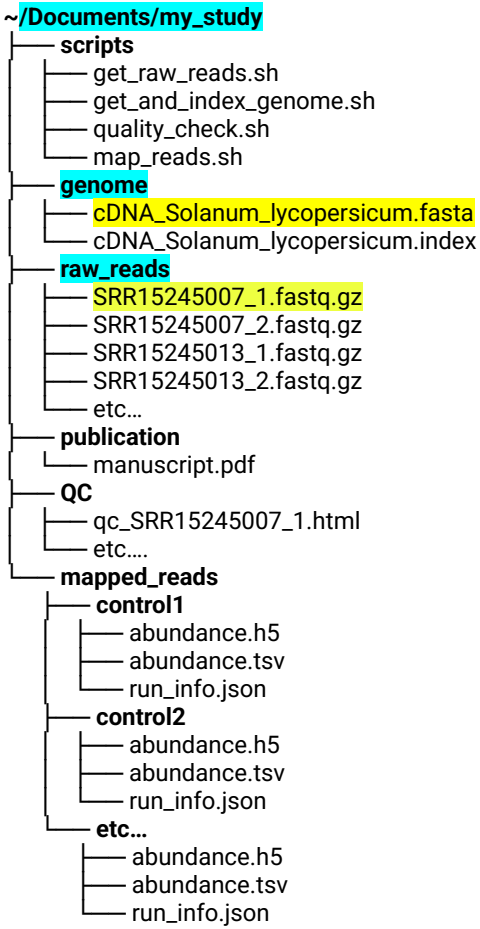
File Edit Format View Help

ENST00000371754.8	5235	4986	1208.55	4.16607
ENST00000469991.1	640	391	14.5506	0.639615
ENST00000371752.5	7212	6963	6866.66	16.9498
ENST00000396105.6	7209	6960	12876.7	31.7988
ENST00000371744.5	2964	2715	1972.76	12.4887
ENST00000455070.1	2122	1873	34.6566	0.318026
ENST00000509231.1	2224	1975	0	0
ENST00000646235.1	3341	3092	0	0
ENST00000644105.2	2821	2572	0	0
ENST00000646118.1	3411	3162	3	0.016307
ENST00000647304.1	3707	3458	1075.39	5.3451
ENST00000395409.7	2192	1943	0	0
ENST00000540171.2	1307	1058	2	0.0324907
ENST00000675743.1	648	399	0	0
ENST00000674723.1	931	682	0	0
ENST00000674869.1	837	588	0	0
ENST00000675248.1	577	328	0	0
ENST00000676427.1	883	634	32.2398	0.874013
ENST00000330634.11	7578	7329	289.566	0.679076
ENST00000392634.9	7623	7374	0.956435	0.00222929
ENST00000675482.1	1020	771	20.6712	0.460816
ENST00000398337.8	1689	1440	224.596	2.68074
ENST00000674966.1	859	610	8.63282	0.243242
ENST00000675207.1	7732	7483	58.9235	0.13534
ENST00000675616.1	431	182	0	0

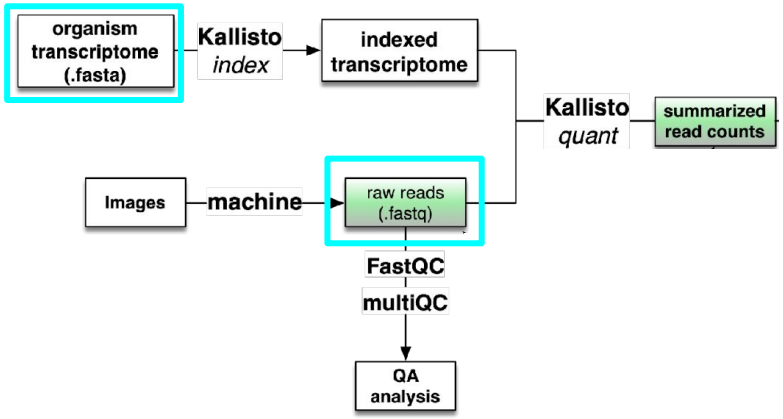
Unix (LF)

The RNAseq analysis pipeline



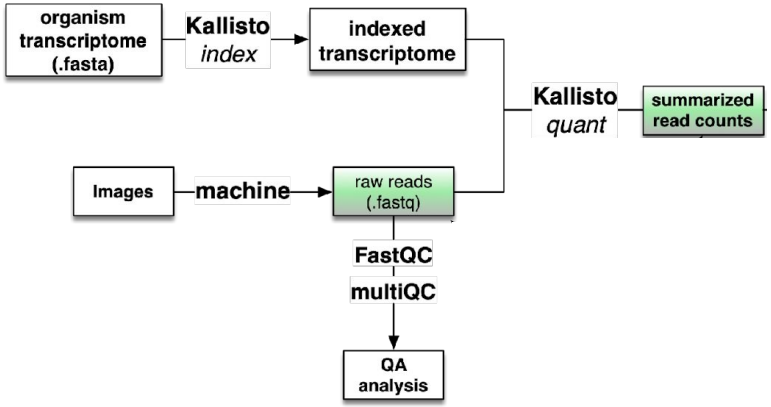

~/Documents/my_study
scripts
get_raw_reads.sh
get_and_index_genome.sh
quality_check.sh
map_reads.sh
genome
cDNA_Solanum_lycopersicum.fasta
cDNA_Solanum_lycopersicum.index
raw_reads
SRR15245007_1.fastq.gz
SRR15245007_2.fastq.gz
SRR15245013_1.fastq.gz
SRR15245013_2.fastq.gz
etc…
publication
manuscript.pdf
QC
qc_SRR15245007_1.html
etc….
mapped_reads
control1
abundance.h5
abundance.tsv
run_info.json
control2
abundance.h5
abundance.tsv
run_info.json
etc…
abundance.h5
abundance.tsv
run_info.json

RNAseq analysis pipeline




```
~/Documents/my_study
├── scripts
│   ├── get_raw_reads.sh
│   ├── get_and_index_genome.sh
│   ├── quality_check.sh
│   └── map_reads.sh
├── genome
│   ├── cDNA_Solanum_lycopersicum.fasta
│   └── cDNA_Solanum_lycopersicum.index
├── raw_reads
│   ├── SRR15245007_1.fastq.gz
│   ├── SRR15245007_2.fastq.gz
│   ├── SRR15245013_1.fastq.gz
│   ├── SRR15245013_2.fastq.gz
│   └── etc…
├── publication
│   └── manuscript.pdf
├── QC
│   ├── qc_SRR15245007_1.html
│   └── etc….
└── mapped_reads
    ├── control1
    │   ├── abundance.h5
    │   ├── abundance.tsv
    │   └── run_info.json
    ├── control2
    │   ├── abundance.h5
    │   ├── abundance.tsv
    │   └── run_info.json
    └── etc…
        ├── abundance.h5
        ├── abundance.tsv
        └── run_info.json
```

The RNAseq analysis pipeline



The End