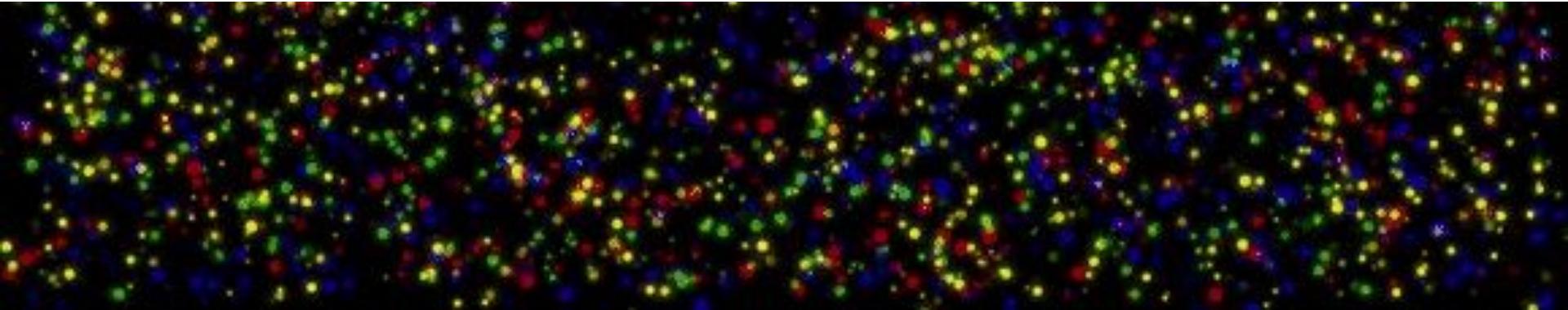


RNA-seq for Plant Science



Jędrzej Jakub Szymański



@NAMlab



www.szymanskilab.com



www.ipk-gatersleben.de



CEPLAS
Cluster of Excellence on Plant Sciences



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

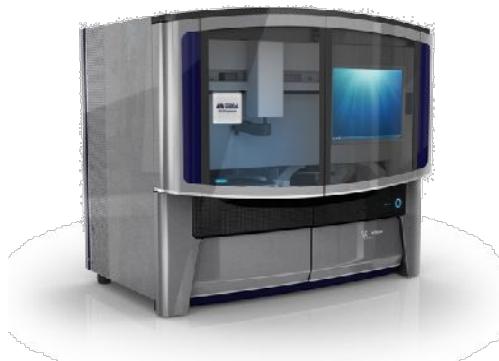
Roche 454 'pyrosequencing'



Life Technologies
Ion Torrent



Life Technologies SOLiD

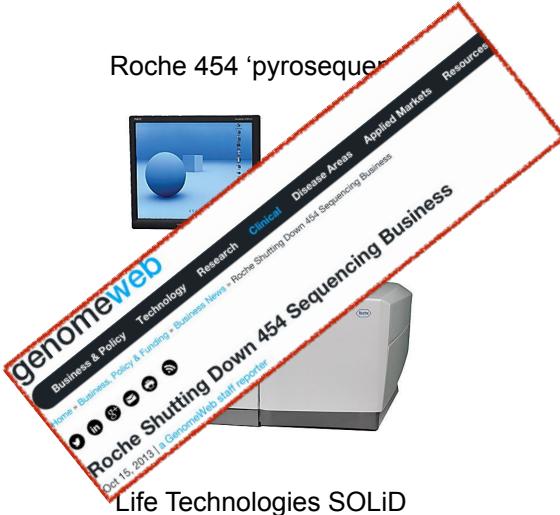


Illumina



Sequencing circa 2012

Roche 454 'pyrosequencer'



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.

SPECIAL FEATURE | METHOD OF THE YEAR

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."

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



Illumina's (Solexa) Genome Analyzer.

© 2008 Nature Publishing Group <http://www.nature.com/naturemethods>

npg

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.

SPECIAL FEATURE | METHOD OF THE YEAR

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."

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



Illumina's (Solexa) Genome Analyzer.

© 2008 Nature Publishing Group <http://www.nature.com/naturemethods>

npg

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.

SPECIAL FEATURE | METHOD OF THE YEAR

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."

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



Illumina's (Solexa) Genome Analyzer.

© 2008 Nature Publishing Group <http://www.nature.com/naturemethods>

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



MiniSeq System

MiSeq Series

NextSeq Series

HiSeq Series

HiSeq X Series

NovaSeq Series

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



MiniSeq System

MiSeq Series

NextSeq Series

HiSeq Series

HiSeq X Series

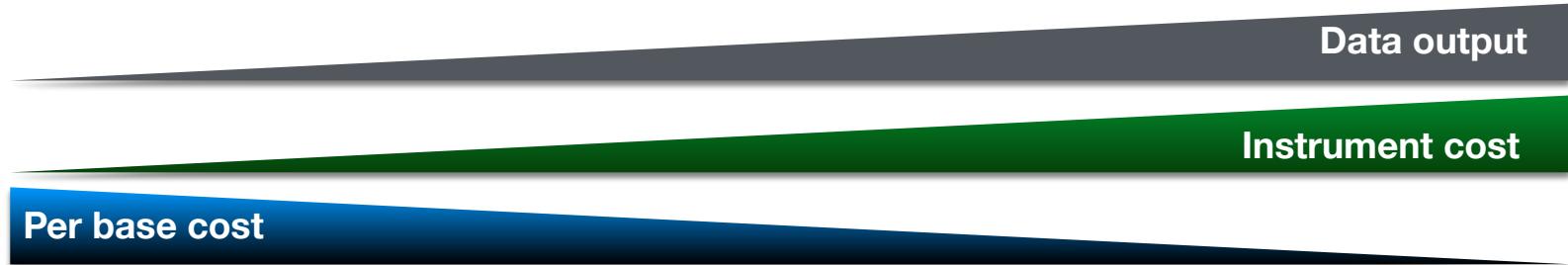
NovaSeq Series

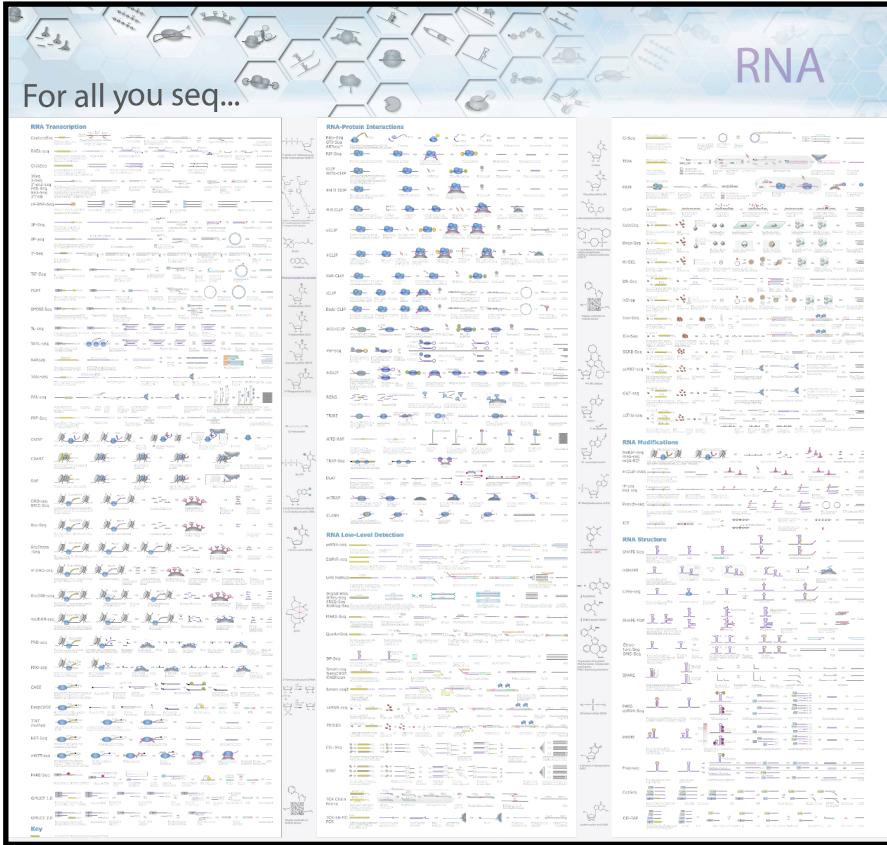
Data output

Instrument cost

Per base cost

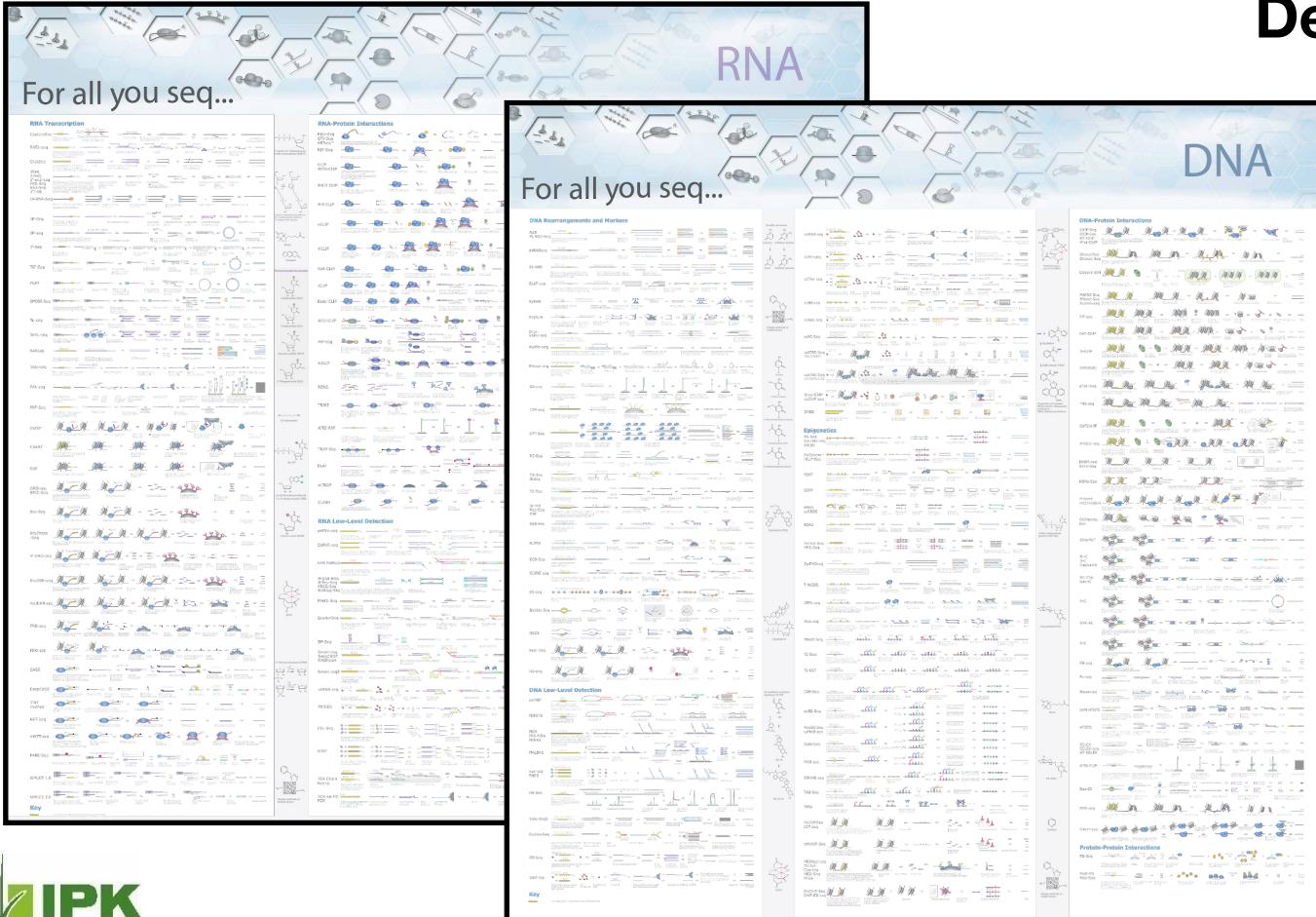
Democratization of sequencing





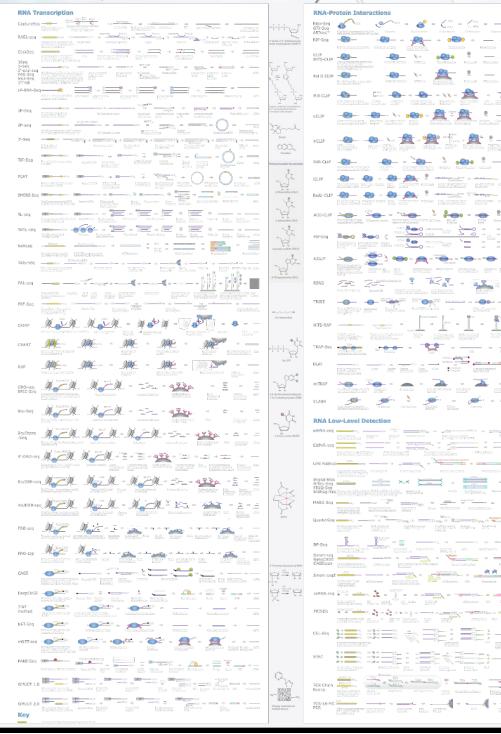
Democratization of sequencing

Democratization of sequencing

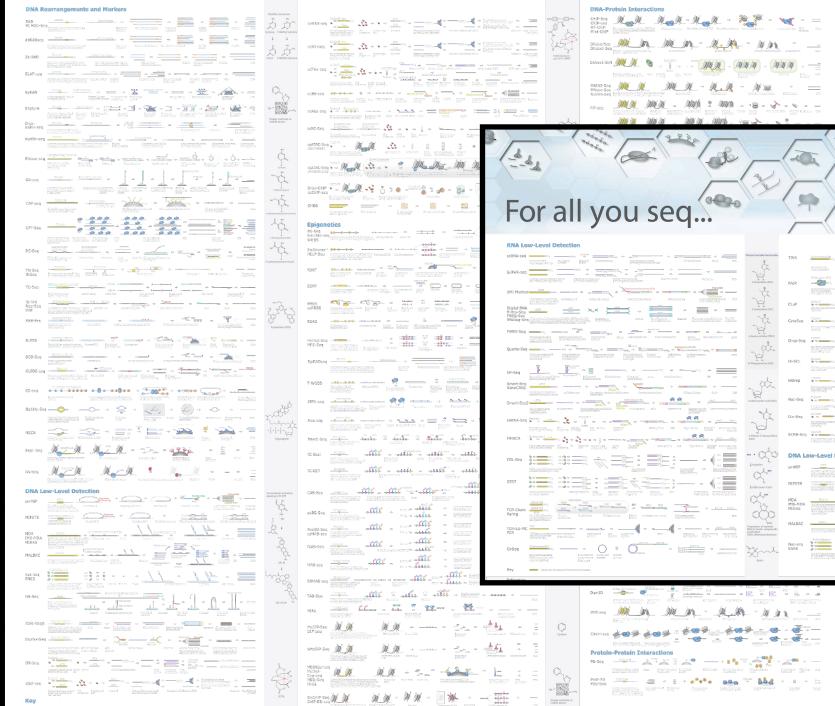




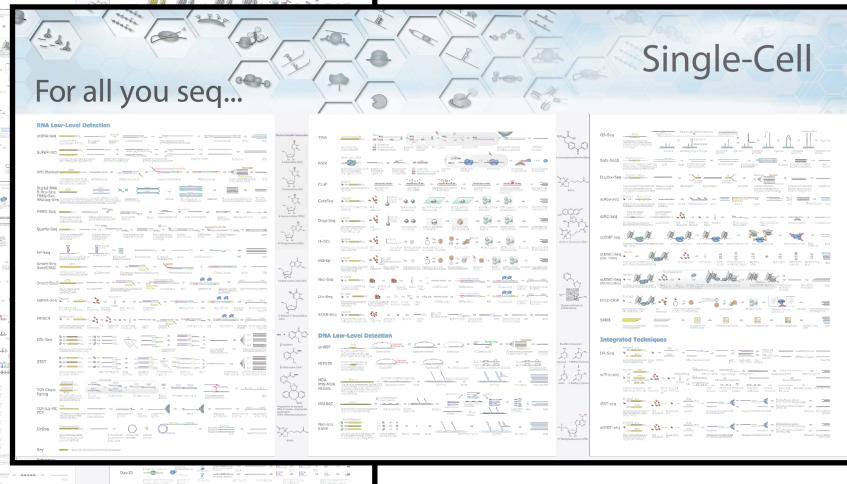
For all you seq...



For all you seq...



Democratization of sequencing



For all you seq...

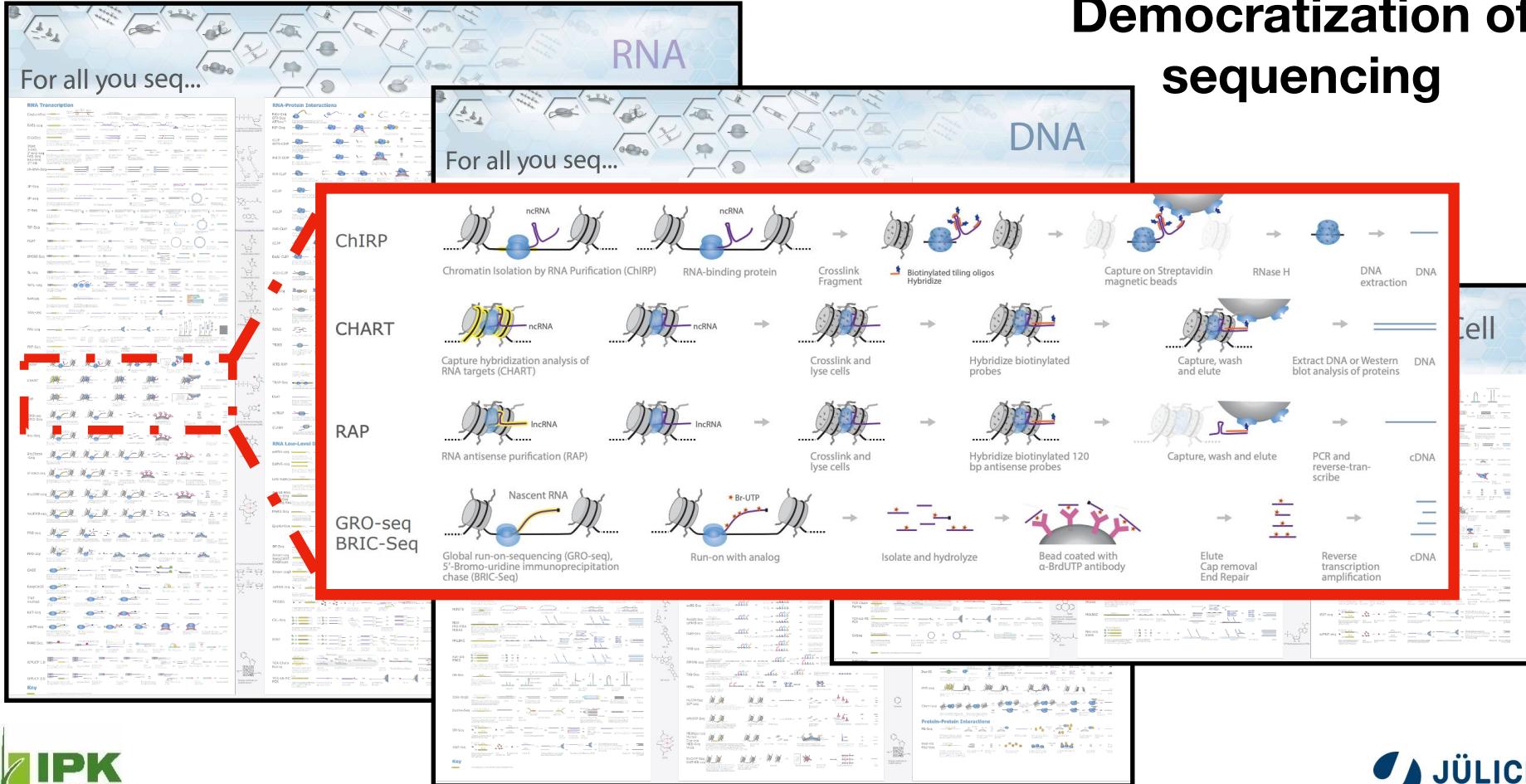


www.ipk-gatersleben.de

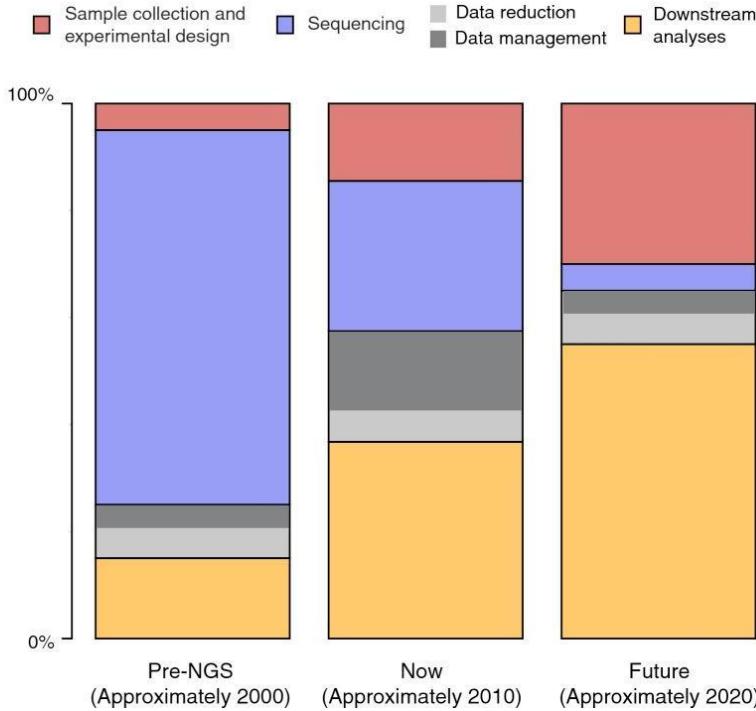
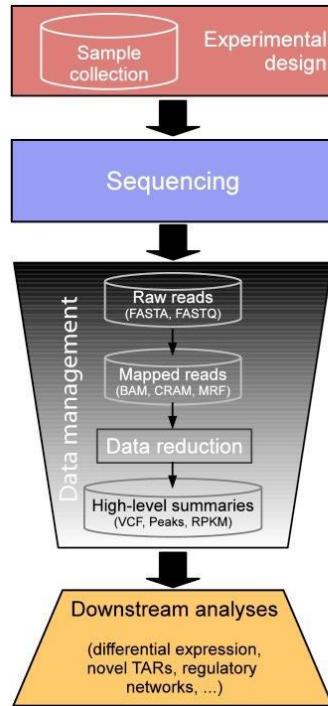


www.fz-juelich.de

Democratization of sequencing

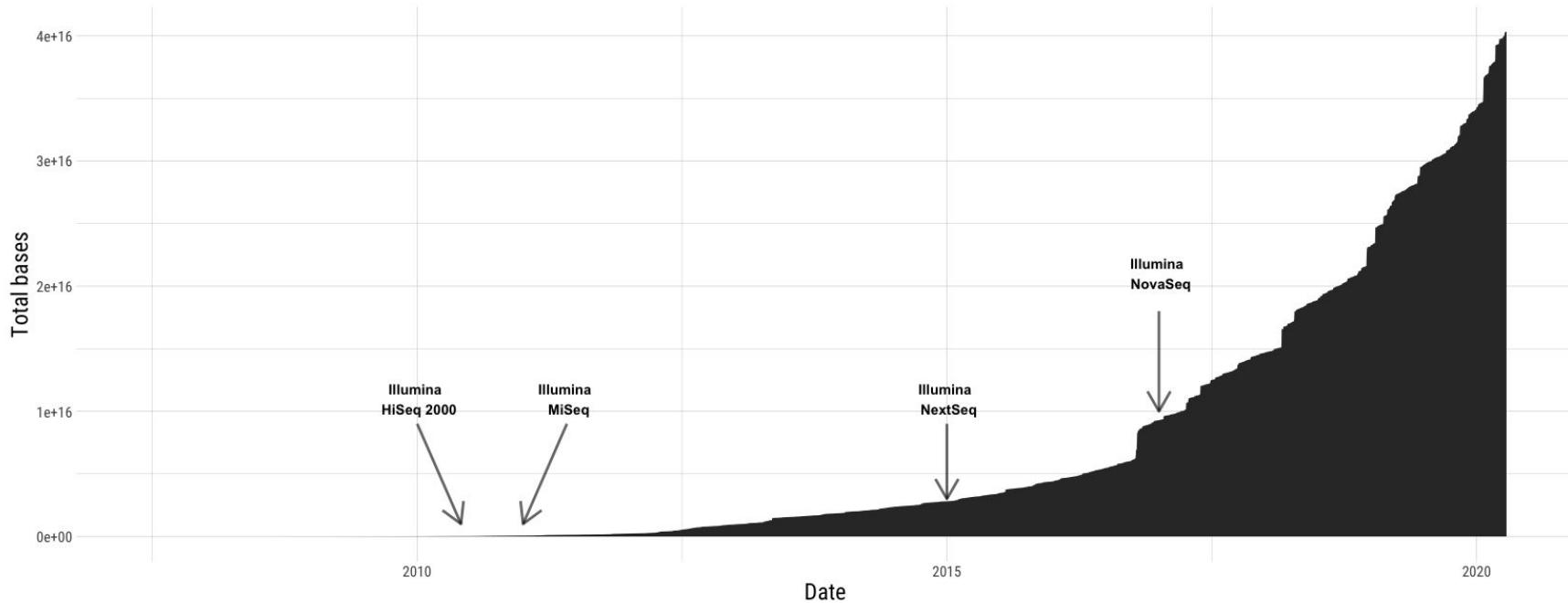


A shift in how time and money are spent on genomics experiments



Sboner A., *Genome Biology*, 2011

NCBI Sequence Read Archive (SRA)

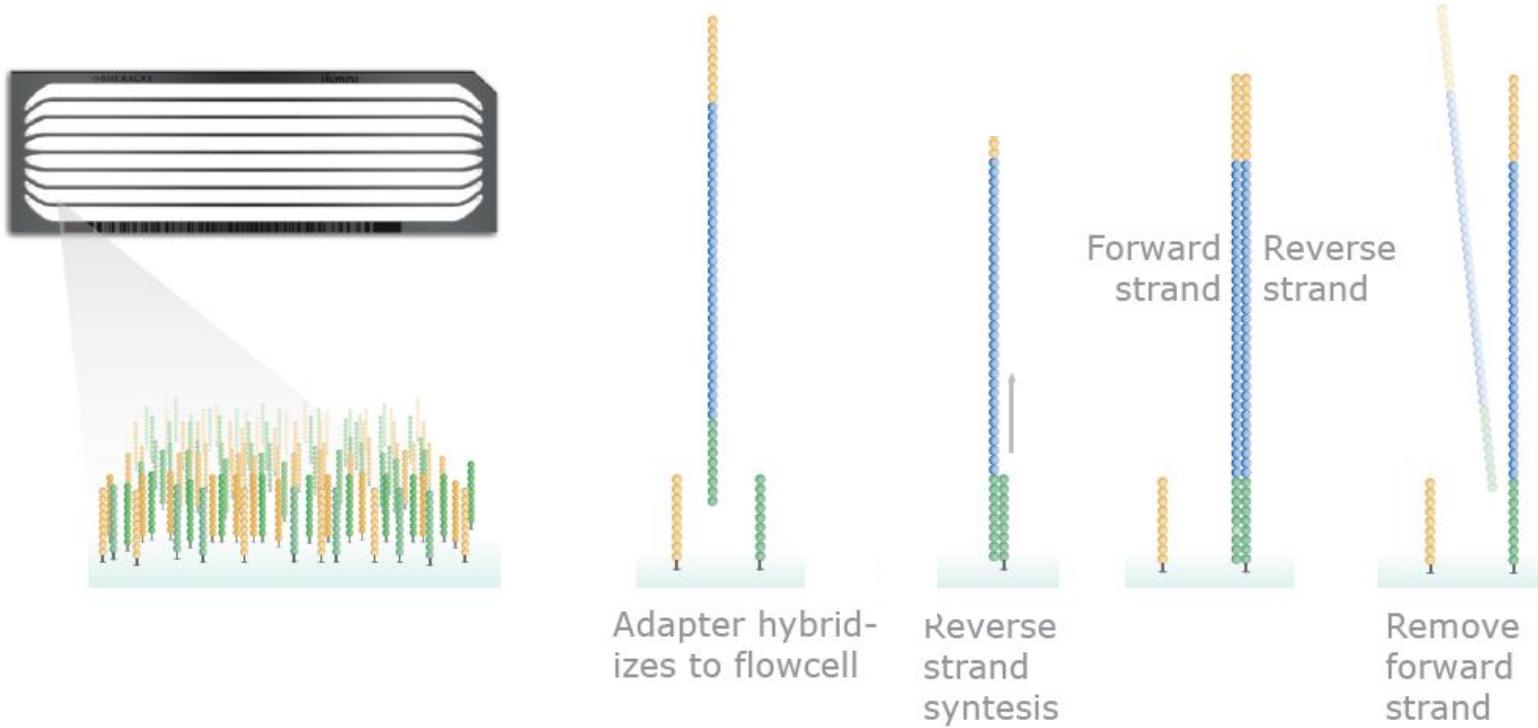


Illumina sequencing



'Sequenceng By Synthesis' (SBS) technology

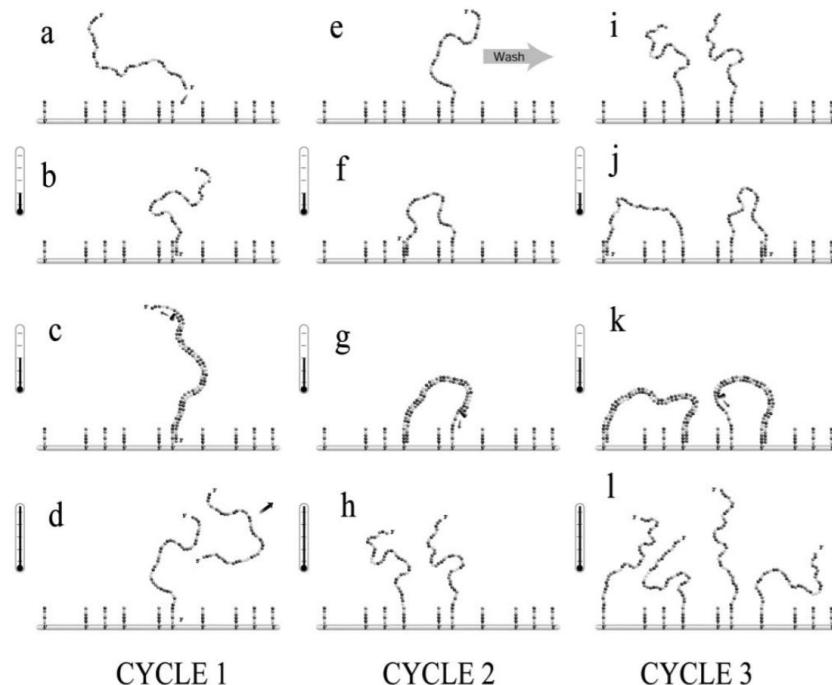
Sequencing by Synthesis



Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model

Jean-Francois Mercier,* Gary W. Slater,* and Pascal Mayer†

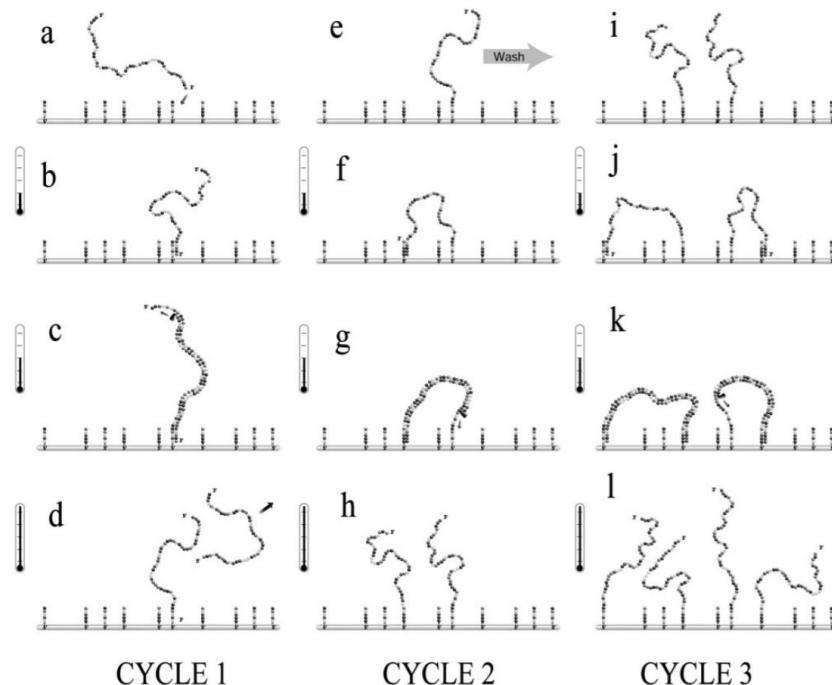
*Department of Physics, University of Ottawa, Ottawa, Ontario, Canada; and †Manteia Predictive Medicine S.A., Coinsins, Switzerland



Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model

Jean-Francois Mercier,* Gary W. Slater,* and Pascal Mayer[†]

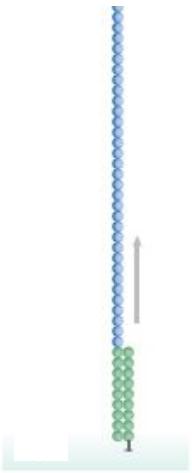
*Department of Physics, University of Ottawa, Ottawa, Ontario, Canada; and [†]Manteia Predictive Medicine S.A., Coinsins, Switzerland



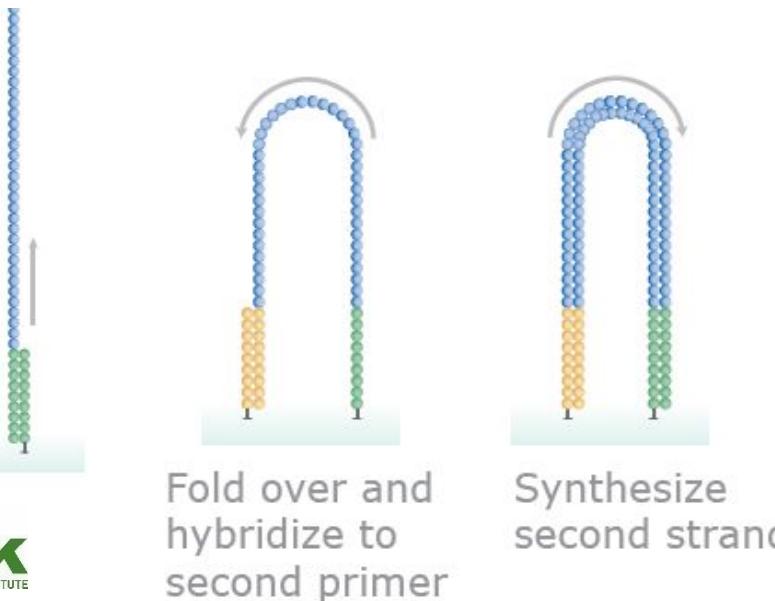
"solid phase DNA amplification leads to the growth of a colony of molecules attached to the surface and located in the same region."

This characteristic could easily be exploited in the design of DNA microarrays."

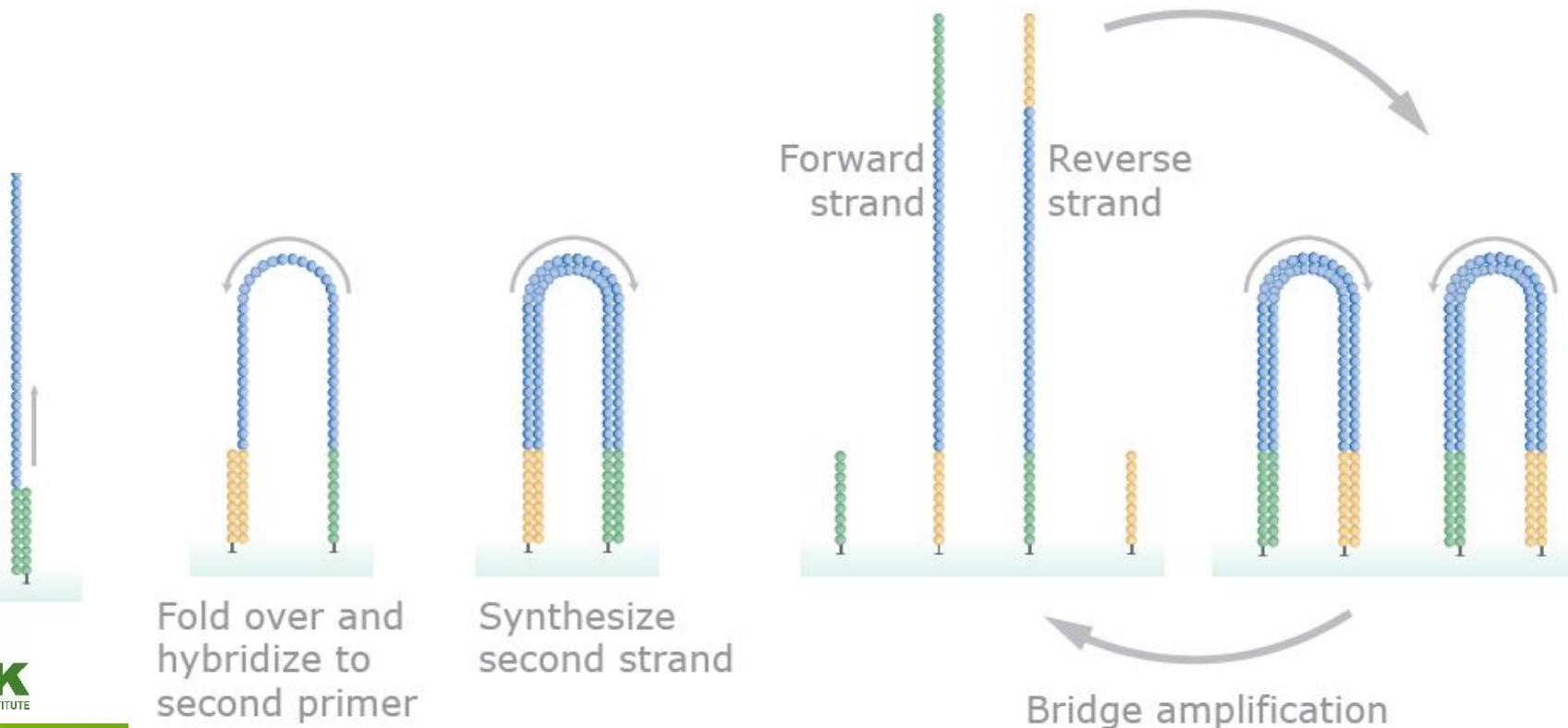
Sequencing by Synthesis (SBS)



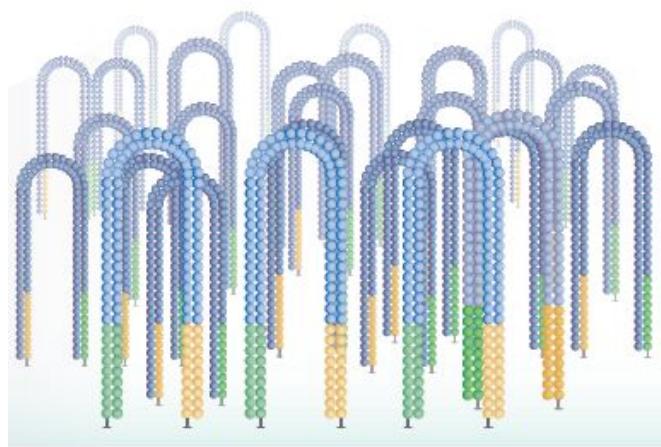
Sequencing by Synthesis (SBS)



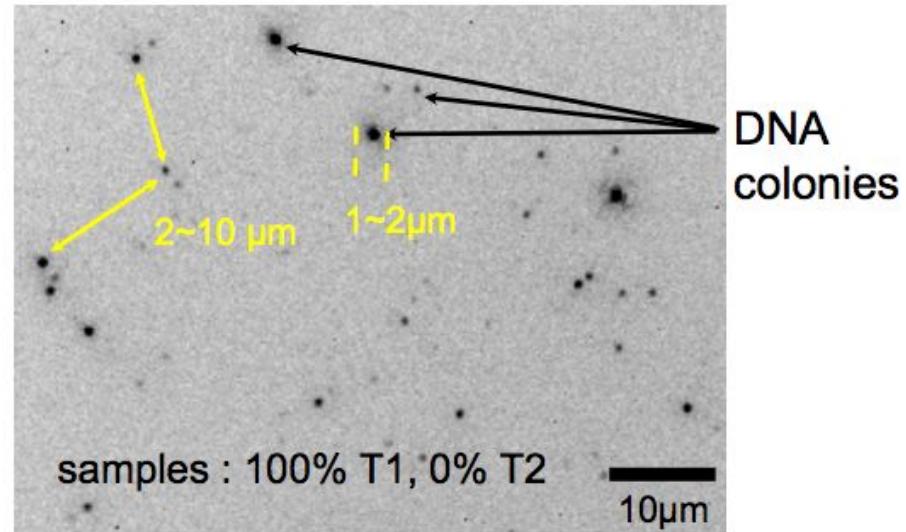
Sequencing by Synthesis (SBS)



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

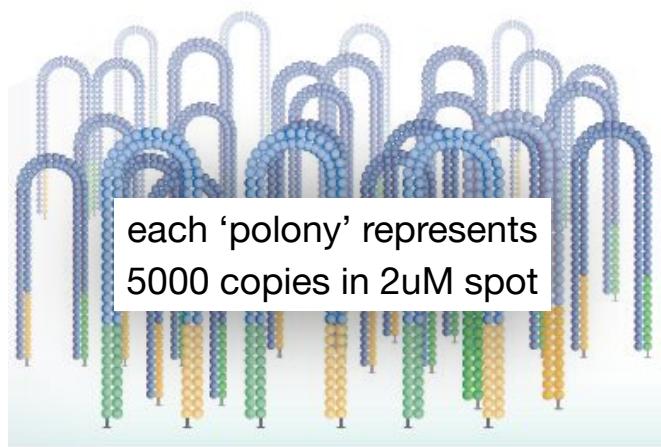


Thousands of molecules are amplified in parallel

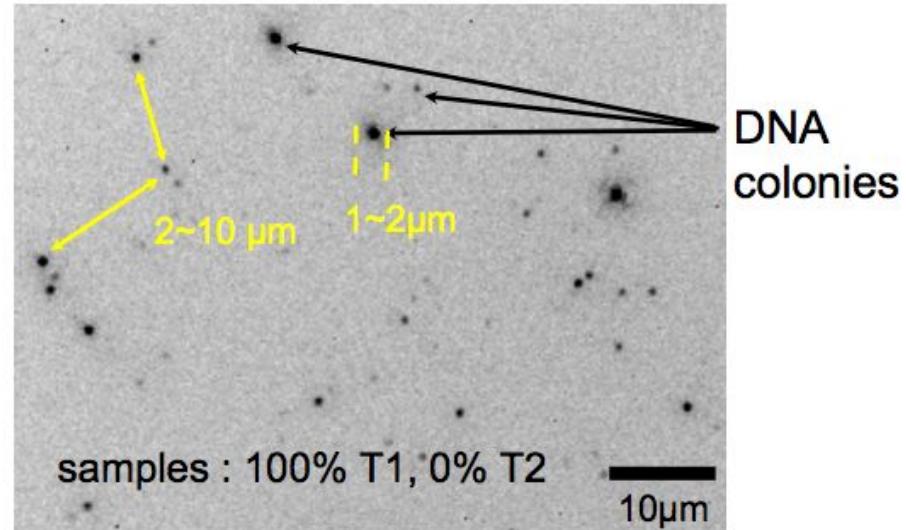


Mayer et al., Presentation 1998

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

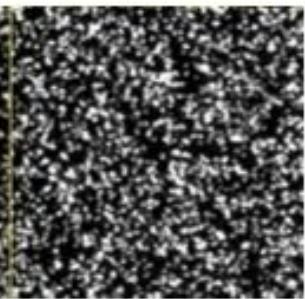


Thousands of molecules are amplified in parallel

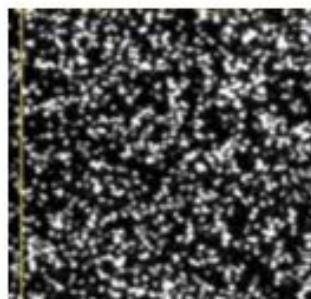


Mayer et al., Presentation 1998

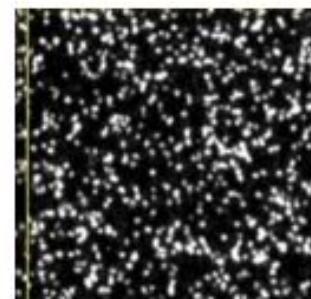
Clustering density is key to data output and quality



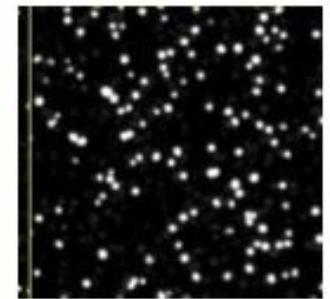
20pM



10pM



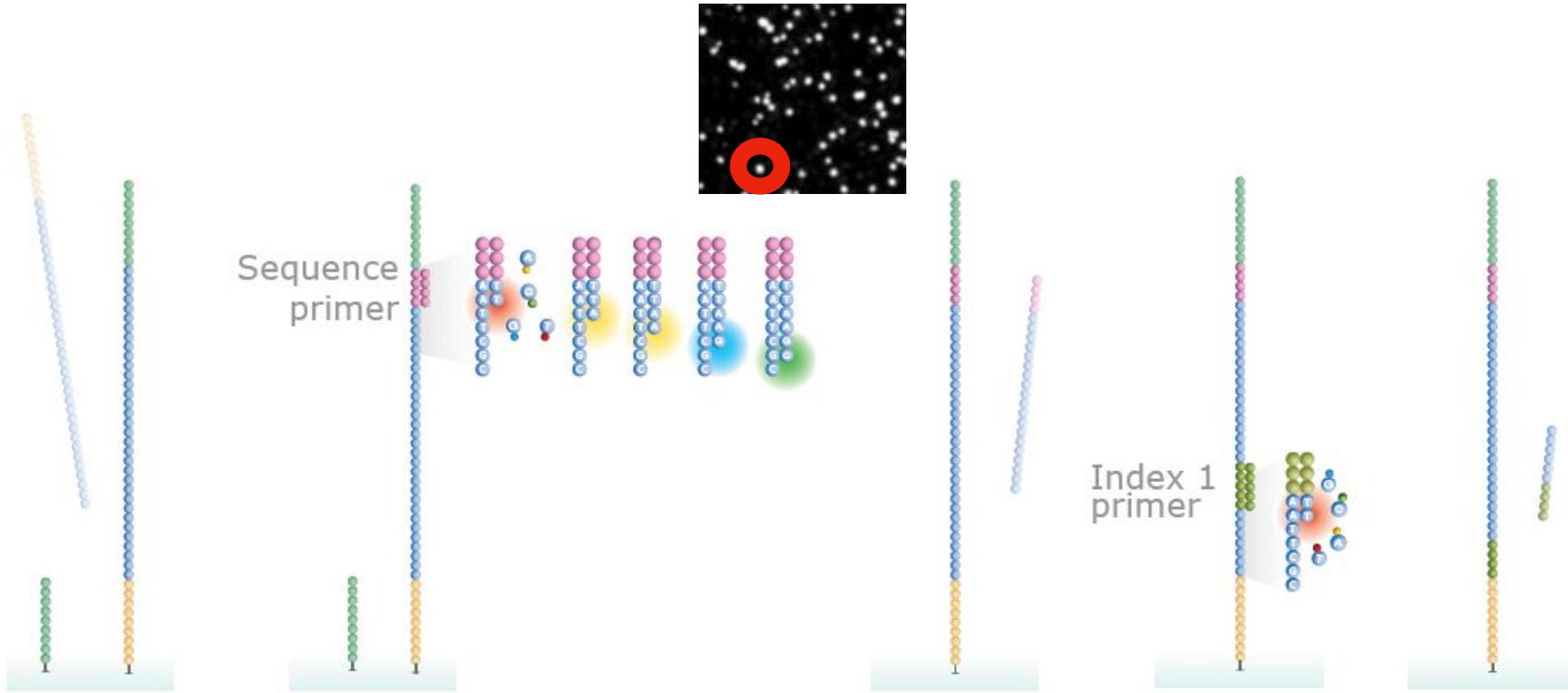
5pM



1pM

HiSeq: 950-1050K clusters/mm²

NextSeq: 170-220K clusters/mm²



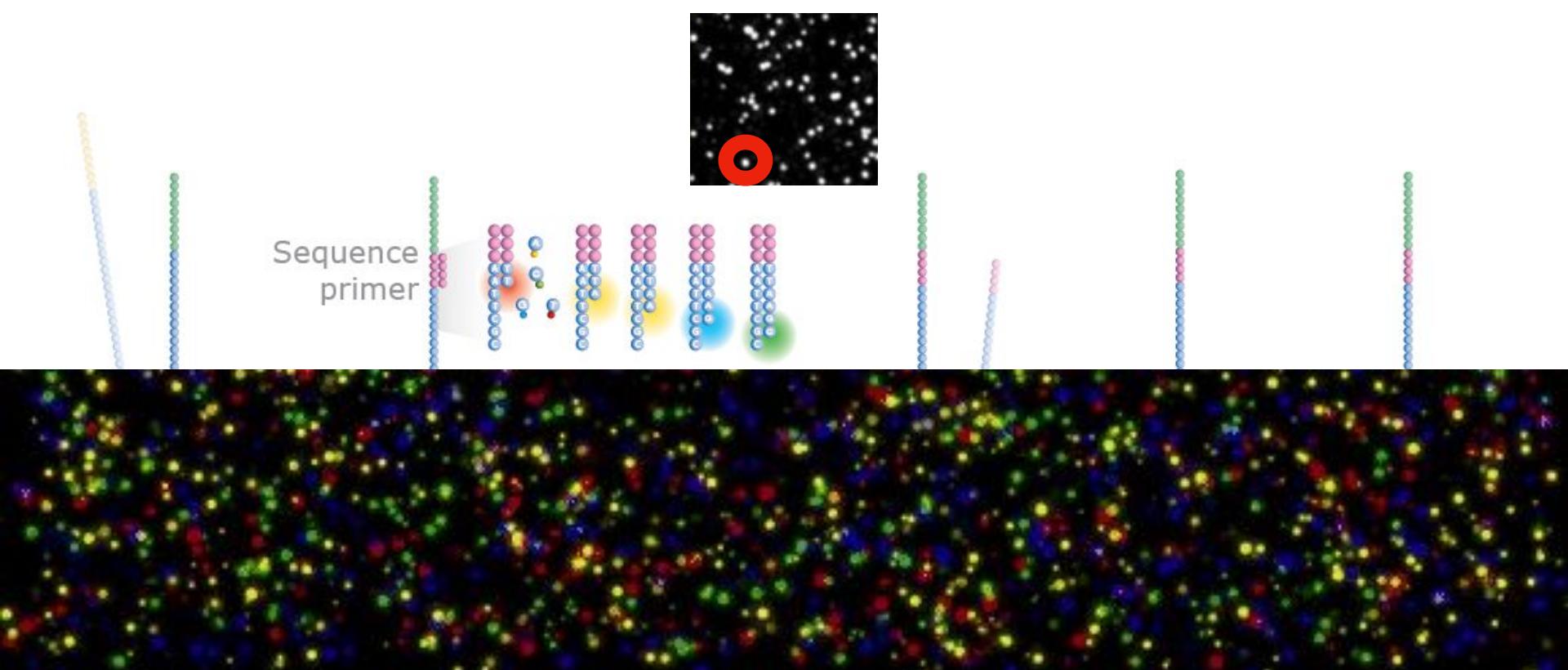
The reverse
strand is
cleaved and
washed away

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

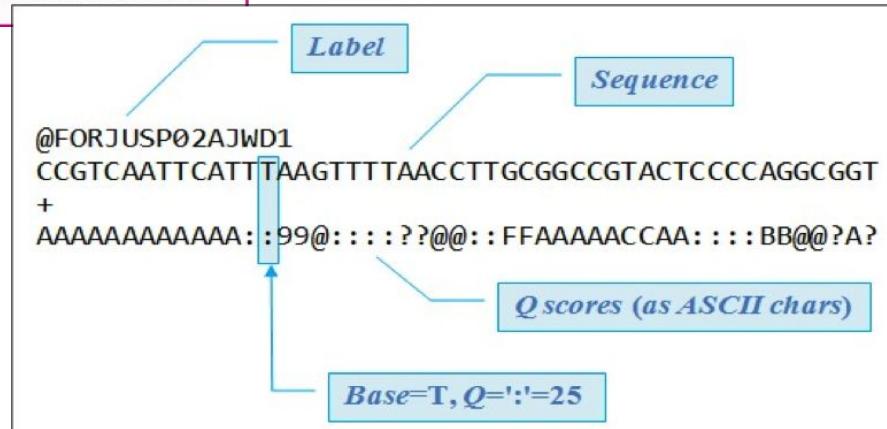
Fasta

Header
Sequence
Header
Sequence
Header
Sequence

- >VIT_201s0011g03530.1
- AATTAAGCATAAATACTCACTTACCCCTTATTTCTTATCTCTCATCATTGGTGCAG
- GACCATGAGAACAAAGCTGCAATGGGTAGGGTTCTCGCAAGGCATGCAGCCAAGACTGCATCA
- >VIT_201s0011g03540.1
- CAGGTAGCGTAGTTAACCCCTAGCGCTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
- AGCCTCTGAGACACCACTCAAACCTTCACTTAAATACACATCCCTCACACCCTTCAATT
- >VIT_201s0011g03550.1
- CATGCAAAGCTGAACCGCATGCTGATTGGTGGTAAGTGGTAGTTGAGTAATTTGACAGTGAA
- GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTCATCACGTGGGCCA

Fastq

*These are just
text files*



Costs associated with bulk* sequencing

Library Preparation



Sequencing



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

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

24 samples

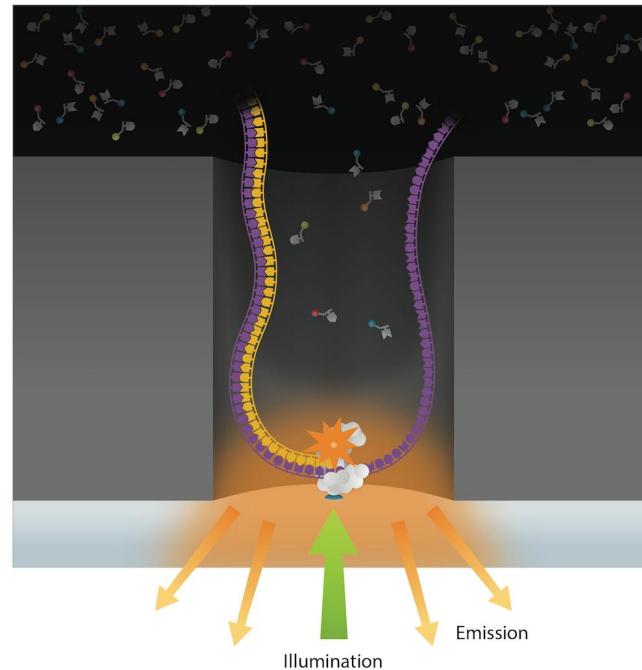
library prep = \$51/sample
sequencing = \$58/sample
data output = 15M reads/sample

Long read sequencing

PacBio Sequel 2



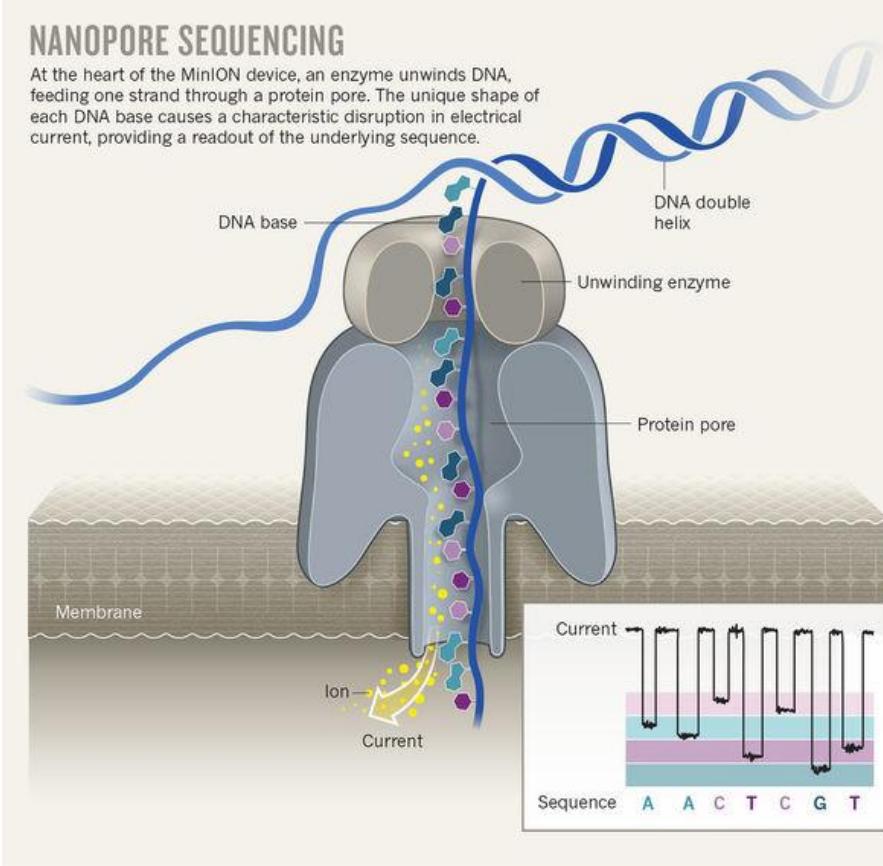
SMRT sequencing



Oxford Nanopore



Oxford Nanopore



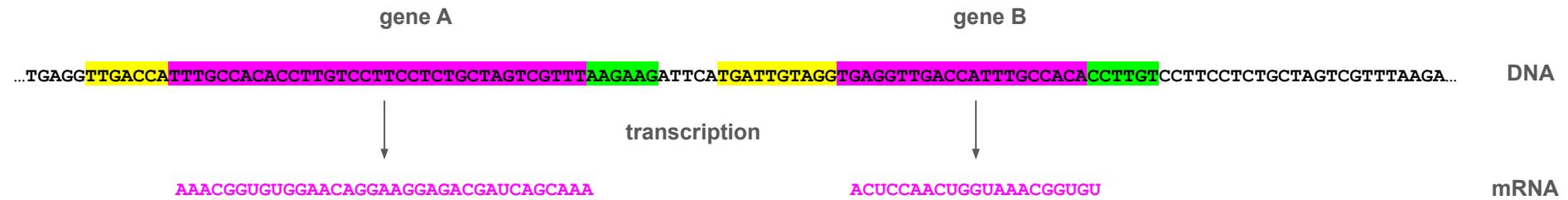
Yaaay!

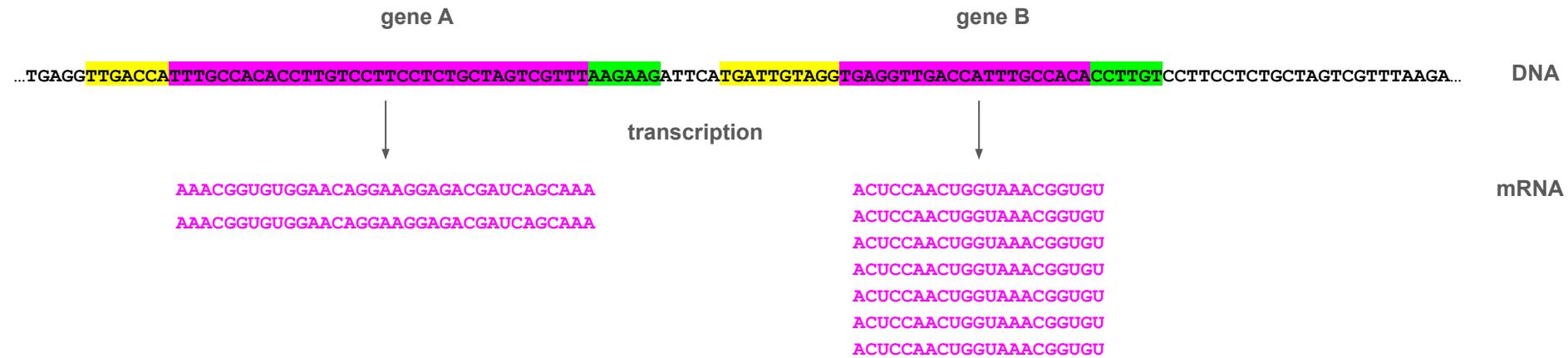
...TGAGGTTGACCATTGCCACACCTTGTCCCTCCTGCTAGTCGTTAAGAAGATTCATGATTGTAGGTGAGGTTGACCATTGCCACACCTTGTCCCTCCTGCTAGTCGTTAAGA... DNA

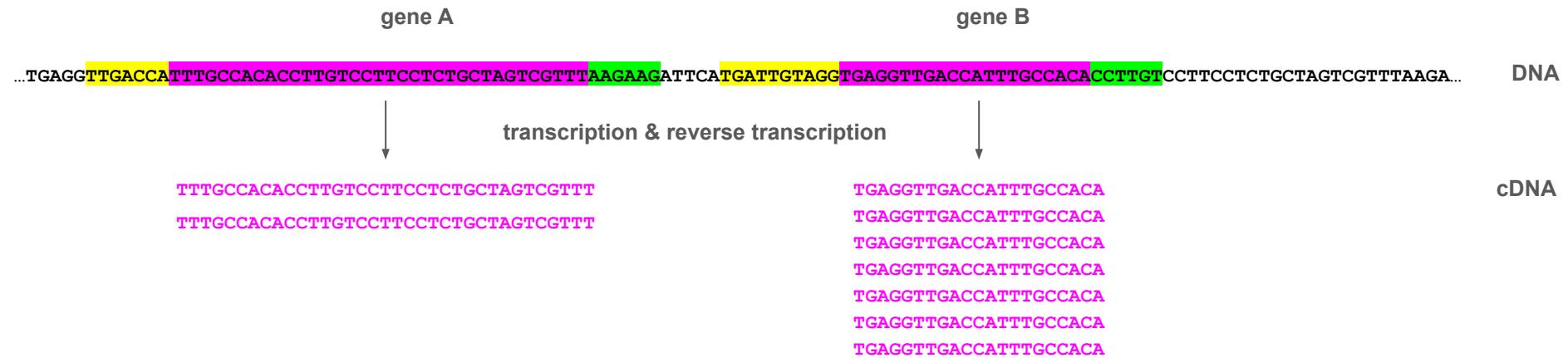
gene A

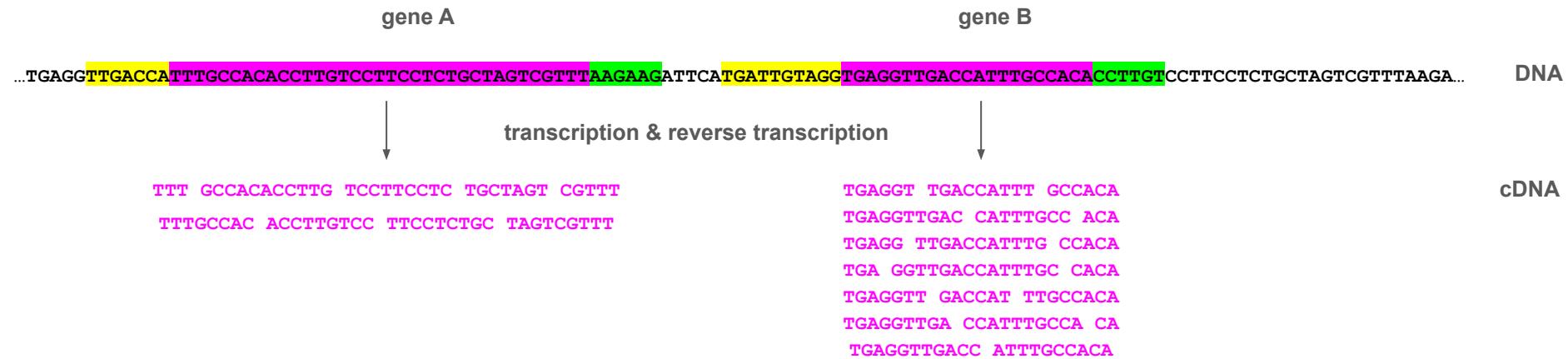
gene B

...TGAGGTTGACCATTTGCCACACCTGTCCTTCCTCTGCTAGTCGTTAAGAAGATTCATGATTGATTGAGGTGACCATTTGCCACACCGTGCCTCCTCTGCTAGTCGTTAAGA... DNA









gene A

gene B

...TGAGGT TGACCA TTTGCCACACCTGTCCTTCCTCTGCTAGTCGTTAAGAAAGATTCATGATTGTAGGTGAGGTTGACCATTTGCCACACCGT CCTTCCTCTGCTAGTCGTTAAGA... DNA

DNA

cDNA

A bag of sequenced reads

TGAGGT TGACCATTT GCCACA TGAGGTTGA CCATTGCCA CA
TGAGGTTGACC ATTTGCCACA
TGAGG TTGACCATTTG CCACA TGAGGTTGAC CATTGCC ACA
TTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTT
TGAGGTT GACCAT TTGCCACA TGA GGTTGACCATTGC CACA
TTT GCCACACCTTG TCCTTCCTC TGCTAGT CGTTT

gene A

gene B

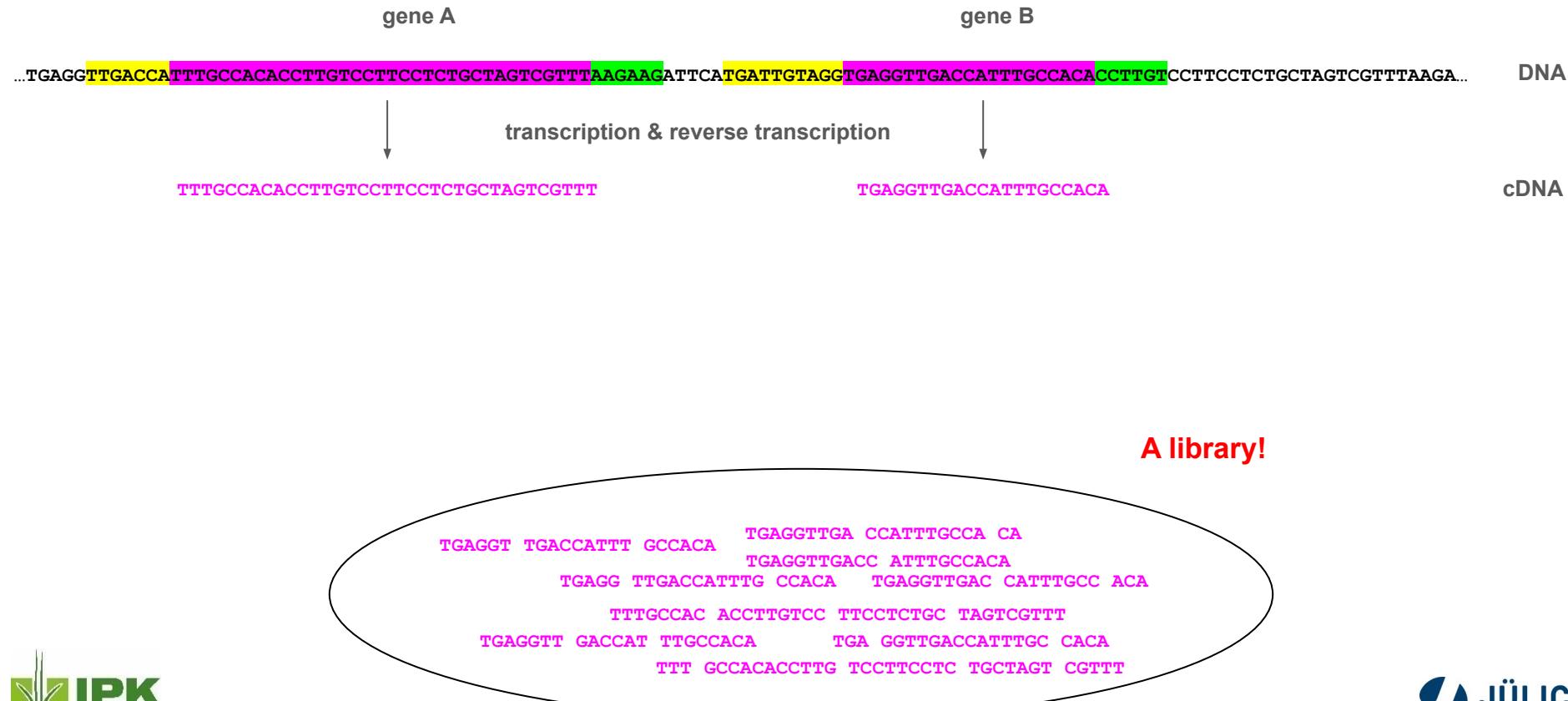
...TGAGGTGACCA TTTGCCACACCTGTCCTTCCTCTGCTAGTCGTTAAGAAAGATTCATGATTGATTGAGGTGAGGTTGACCATTTGCCACACCGT CTTCCCTCTGCTAGTCGTTAAGA... DNA

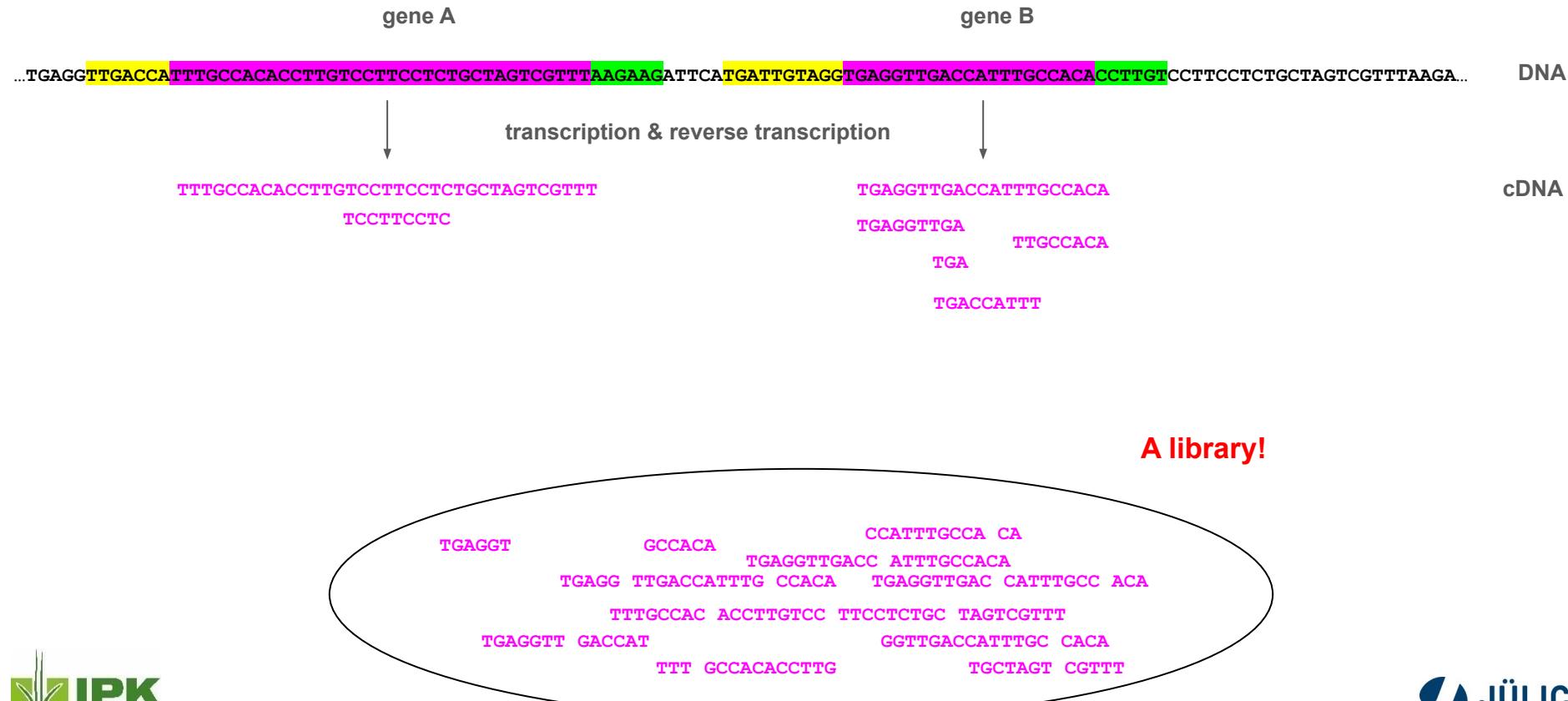
DNA

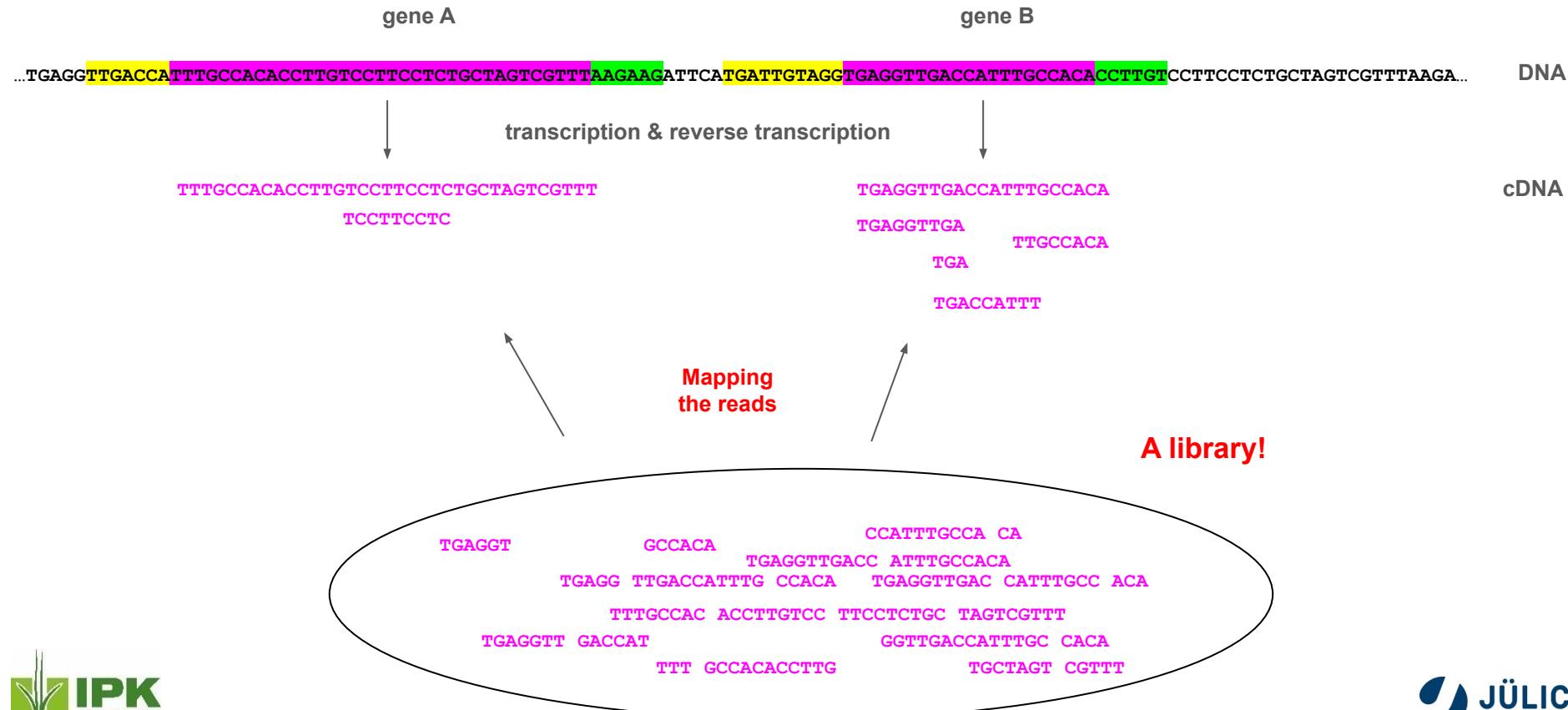
cDNA

A library!

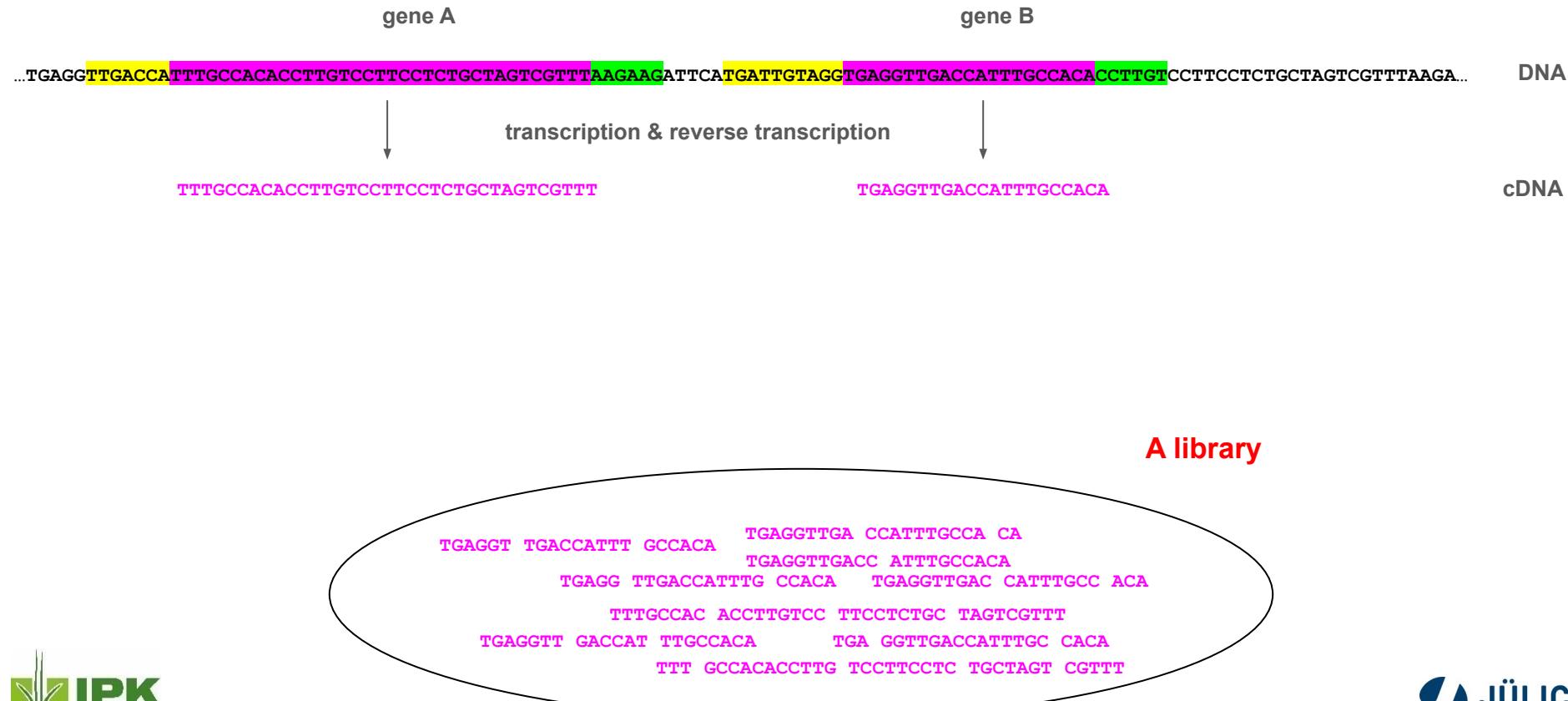
TGAGGT TGACCATTT GCCACA TGAGGGTGA CCATTGCCA CA
TGAGGTTGACC ATTTGCCACA
TGAGG TTGACCATTTG CCACA TGAGGGTGCACATTTGCC ACA
TTGCCAC ACCTTGTCC TTCCTCTGC TAGTCGTT
TGAGGTT GACCAT TTGCCACA TGA GGTTGACCATTTGC CACA
TTT GCCACACCTTG TCCTTCCTC TGCTAGT CGTTT



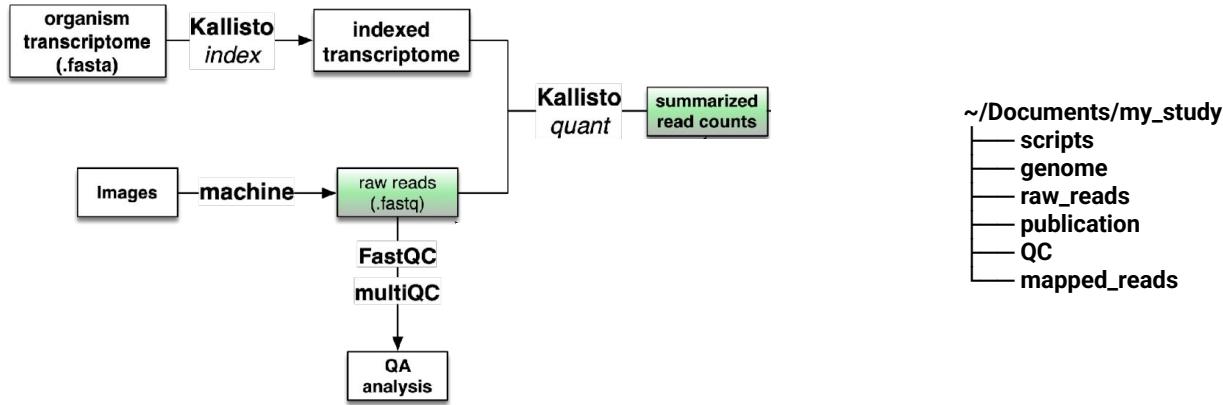




What data do we need to start an RNAseq project?

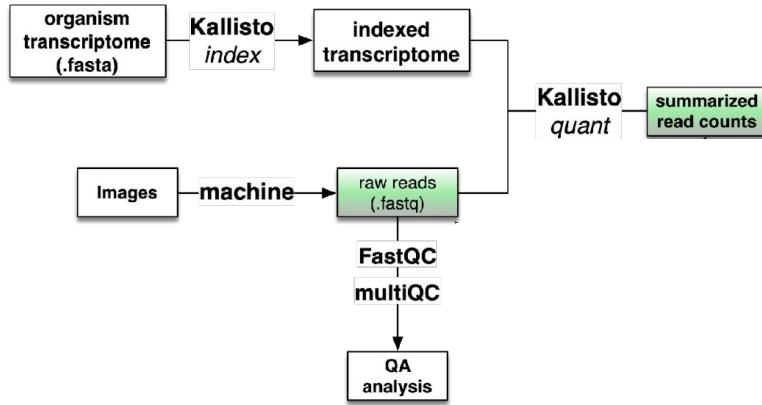


RNAseq analysis pipeline



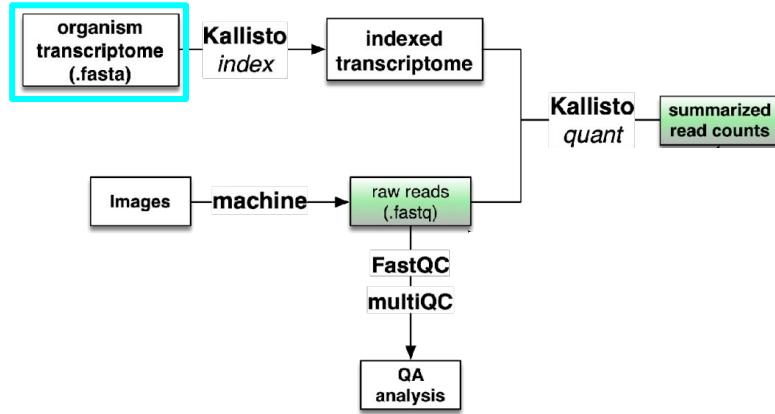
```
~/Documents/my_study
├── scripts
├── genome
├── raw_reads
├── publication
├── QC
└── mapped_reads
```

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

e! EnsemblPlants • HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

Search: for
e.g. [Carboxy*](#) or [chx28](#)

All genomes

-- Select a species --

[View full list of all species](#)

Favourite genomes

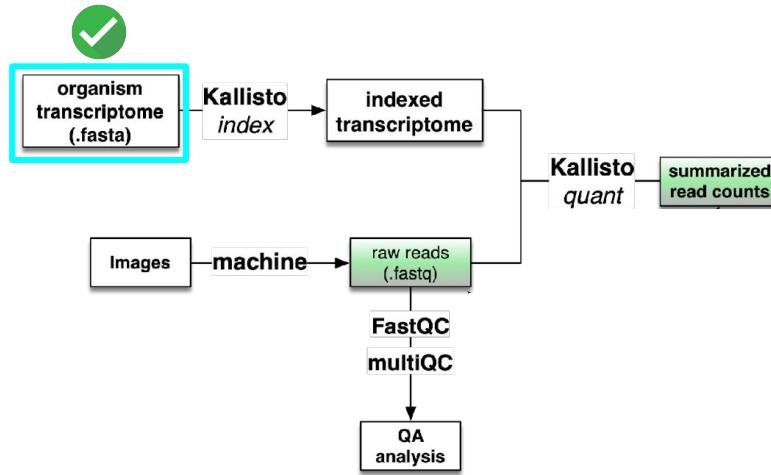
	Arabidopsis thaliana TAIR10
	Oryza sativa Japonica Group IRGSP-1.0
	Triticum aestivum IWGSC
	Hordeum vulgare MorexV3_pseudomolecules_assembly
	Zea mays Zm-B73-REFERENCE-NAM-5.0
	Physcomitrium patens Phypa_V3

Get the reference genome

The screenshot shows the EnsemblPlants homepage. At the top, there is a navigation bar with links to HMMER, BLAST, BioMart, Tools, Downloads, Help & Docs, and Blog. Below the navigation bar is a search bar with a dropdown menu set to "All species" and a text input field for searching. A red arrow points upwards from the search bar towards the navigation bar. Below the search bar is an example search term "e.g. Carboxy* or chx28". The main content area is divided into two sections: "All genomes" on the left and "Favourite genomes" on the right. The "All genomes" section has a dropdown menu "Select a species" and a link "View full list of all species". The "Favourite genomes" section lists several species with their assembly names and small images:

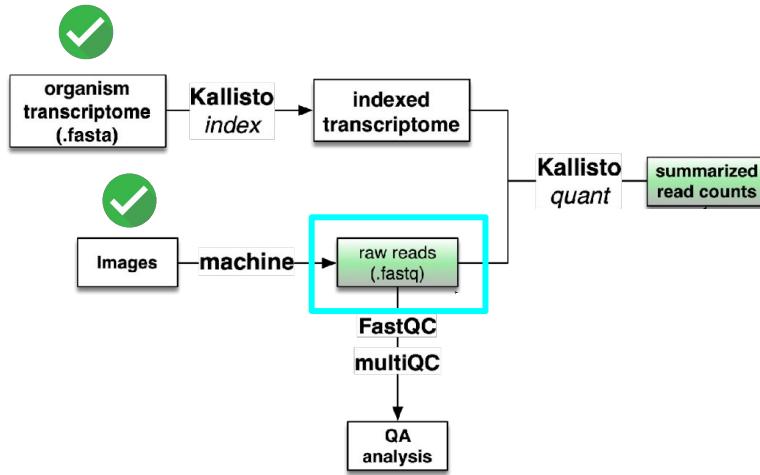
- Arabidopsis thaliana* TAIR10
- Oryza sativa Japonica Group* IRGSP-1.0
- Triticum aestivum* IWGSC
- Hordeum vulgare* MorexV3_pseudomolecules_assembly
- Zea mays* Zm-B73-REFERENCE-NAM-5.0
- Physcomitrium patens* Phyta_V3

RNAseq analysis pipeline



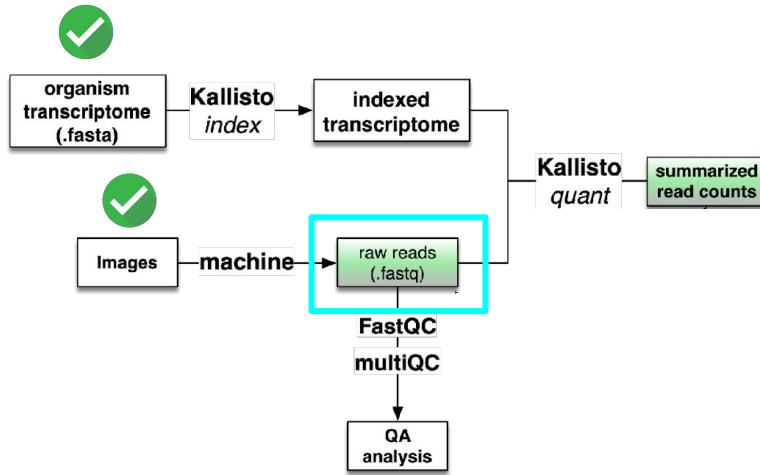
```
~/Documents/my_study
├── scripts
├── genome
│   └── Solanum_lycopersicum.SL3.0.cdna.all.fa.gz
├── raw_reads
├── publication
│   └── manuscript.pdf
└── QC
    └── mapped_reads
```

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

National Library of Medicine
National Center for Biotechnology Information

SRA

Log in

SRA Advanced Search Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

- [Documentation](#)
- [How to submit](#)
- [How to search and download](#)
- [How to use SRA in the cloud](#)
- [Submit to SRA](#)

Tools and Software

- [Download SRA Toolkit](#)
- [SRA Toolkit Documentation](#)
- [SRA-BLAST](#)
- [SRA Run Browser](#)
- [SRA Run Selector](#)

Related Resources

- [Submission Portal](#)
- [dbGaP Home](#)
- [BioProject](#)
- [BioSample](#)

FOLLOW NCBI

X f in

Connect with NLM

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

Web Policies
FOIA
HHS Vulnerability Disclosure

Help
Accessibility
Careers

NLM | NIH | HHS | USA.gov

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

From SRA search

SRA Advanced Search Builder

(("solanum lycopersicum"[Organism]) AND "illumina"[Platform] AND "strategy rna seq"[Properties])

[Edit](#) [Clear](#)

Builder

Organism [Show index list](#)

AND Platform [Show index list](#)

AND Properties [Hide index list](#)

[Previous 200](#) [Next 200](#) [Refresh index](#)

strategy rip seq (20962)
strategy rna seq (4704915)
strategy selex (39382)
strategy ssrna seq (4703)
strategy synthetic long read (11812)
strategy targeted capture (391761)
strategy tethered chromatin conformation capture (654)
strategy tn seq (12547)
strategy validation (99)
strategy wcs (17036)

AND All Fields [Show index list](#)

From SRA search

SRA Advanced Search Builder

(("solanum lycopersicum"[Organism]) AND "illumina"[Platform]) AND "strategy rna seq"[Properties]

[Edit](#)

[Clear](#)

Builder

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
```

strategy validation (99)
stratenv.wcs (17036)

AND ▾ All Fields ▾

[Show index list](#)

[Search](#) or [Add to history](#)

[Refresh index](#)

From ENA search

ENATM European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

We recommend that you subscribe to the ENA-announce mailing list for updates on ENA services.

Effective September 1st, 2023, our data retrieval APIs will implement enhanced performance measures. Each IP Address will be subject to a rate limit of 50 requests per second, ensuring optimized and efficient access to our APIs.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

Submit | Search | Rulespace | Support

Please take our brief (9 minutes) survey about how ENA data is used : <https://forms.gle/hZRpEnjGhAqVKA>

Latest ENA news

ENA: Improving spatio-temporal annotations **Dec 1, 2021, 1:00:00 AM**
The European Nucleotide Archive, along with its partners in the International Nucleotide Sequen
[Read more >](#)

Retirement of old ENA Browser on 5th August 2020 **Jul 16, 2020, 2:00:00 AM**
The new ENA Browser (<https://www.ebi.ac.uk/ena/browser/home>) has been running in parallel to our old Browser (<https://www.ebi.ac.uk/ena>) since mid 2019.
[Read more >](#)

[See all news](#)

The European Nucleotide Archive (ENA) is part of the ELIXIR infrastructure
The ENA is an ELIXIR Core Data Resource. [Learn more](#)

The European Nucleotide Archive (ENA) is a Global Core Biodata Resource
The ENA is a GCB Global Core Biodata Resource. [Learn more](#)

EMBL-EBI is the home for big data in biology.
We help scientists exploit complex information to make discoveries that benefit humankind.

www.ipk-gatersleben.de

1 Data Type

2 Query

3 Inclusion/Exclusion

4 Fields

5 Data Filters

Reset

Type to filter query params

Query:

Build Query ?

AND OR + Add rule ○ Add group X Delete

NCBI Taxonomy = solanum lycopersicum

NCBI Taxonomic classification

Solanum lycopersicum L., 1753

4081

Solanum lycopersicum var. cerasiforme (Alef.) Fosberg, 1955 192583

Solanum lycopersicum var. cerasiforme x Solanum pimpinellifolium 2928697

Solanum lycopersicum x Solanum habrochaites 357543

Solanum lycopersicum x Solanum pennellii 872925

Solanum lycopersicum x Solanum peruvianum 95449

Solanum lycopersicum x Solanum pimpinellifolium 286550

Solanum lycopersicum x Solanum tuberosum 298325

Solanum

Update query

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 Mishra², Bettina Tschiersch³, 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 Mishra², Bettina Tschiersch³, 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.,

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 accession numbers 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 Tissues¹[OPEN]

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

al 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.

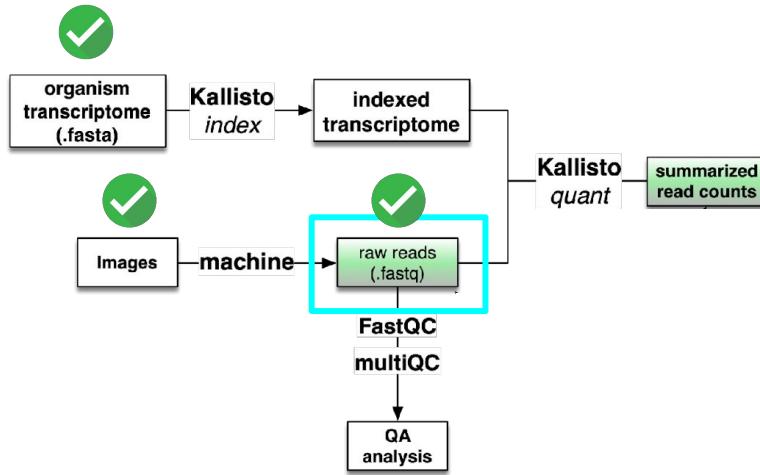
Get the files

```
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.

:om/p/phys/air

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

```
1 @SRR10056916.5114340 5114340 length=101
2 GATTAACATATTTCATTTGAACTTGTAGCGTTCTACACATTGCTTACCCCTCCCTCGATAATATCCACCGAACATCCGTGG
3
4 @SRR10056916.5114341 5114341 length=101
5 CGCCGTTCCCGAGAGGGTGTTCAGAGCTGGAGTTATCCAAGAGATAACGCCGAAGATTTTTGAATCTACTCAAGAAAAGAAGAAAAGAAGTGTGG
6
7 @SRR10056916.5114342 5114342 length=101
8 GTGAAAGTACTACGCCCTGGTGGCCGGTGGAGGGCCGGCCGGAGAGGTAATGGTGTGGATATGGTGAATCAGCAGTGTTCAGCCCCATCC
9
10 @SRR10056916.5114343 5114343 length=101
11 CTAGGAGATTTTGATCTCATCAGAGTTATACGTCAATAGAAAAGTAGAGATGACCTAAGAGTAAGAATCTAATATTTGAATATTGAAATGAGGG
12
13 @SRR10056916.5114344 5114344 length=100
14 AGACAGAGCTGTTTGCCTGCAGAACTGCAACGGGATAGTGAGGGAGGCCACCGAGGAGGTTGATGGATGCTATGGCTAGGCAGACTCCACTGGG
15
16 @SRR10056916.5114345 5114345 length=99
17 AGTATGTTAACCTATCGCGAACATTCAAAACAAAAGGGAAAGAGGTTAGCAAGGCTCTAAGATCCAGAGGCTGGAACACCATTGACTCTC
18
19 @SRR10056916.5114346 5114346 length=100
20 ACTAGAACATCATCACCCACTTACATCTCAGCAAAACCATCATGTTAACAGTGATGGTGCCTCTGGAGAGCGTGGATCAAATGCACGAACAGTG
21
22 @SRR10056916.5114347 5114347 length=101
23 GCTTAAAGAGGTGAAGAGCAACGGTCAGATGCTGGCTTAAATGAGATTTTGAGAAGAACCAACAACATCAAGAATTATGGTATTTGGCTCGTT
24
25 @SRR10056916.5114348 5114348 length=101
26 TGAGTGGCGTCTGGATTCTGCCACACTCTCGATTCCCCGGGTCATCTGAACATATAATGTGGAGAAATGCCATGAGGAAGGAGGCCAGTATT
```

.fastq

reference

```
ACGTTCTGGAGTGGCTGCAGCTACCCATGATGCGTAACAGGCTGGCCATCCAAG
CCATGCAGGATCACTCAGGATTTCAGTTCACCTCTATTCCAAGCATTACCTCAA
AGGACCCAGCAGCTACACCCCTACAGGCTTCAAGGCCACCTCATAGTCATGCTCTCC
CATTTACCCCTACCCATCCTGATCGGTATGCCCTAGCCTGACCCCTTAGATAAGCAA
TGAGGTAGGAAGAACAAACCCCTTGGCTTCCCTGGCATGTTGGAGAAAGTGCCTGCCTGG
TCCGAGCCGCCCTGGTCTGAAGCAGGTGCTCTGCTTACCTTGCTTAGGCTGCTGCA
GAAGCACCTGCCGGTGCAGTCAAGCACCCTTGTGCTAGAGCCCTCATCACCTCAG
CTGTCACCATGGCCAGGAACCAACAGCAGTGGTTACTGCTGGGGTAAACT
AACTCAGTGGAAATGGGCTGTTACTTGGGCTGCTCAACTCATAAAGTTGGCTGATT
TGAAAAAAAGCTCATAAATAAAAGGATGTTGATGGCTGGTCCC
>ENST00000397163.8 cdna chromosome:GRCh38:15:42359501:42412317:1 gene:ENSG0000092529.25
gene_biotype:protein_coding transcript_transcript_biotype:protein_coding gene_symbol:CAPN3 description:calpain 3
[Source:HGN Symbol;Acc:HGN;C:1480]
ACTCTTTCTCTCCCTCTGGCATGCTGCTGGAGAGCCCAAAGTCAACAT
TGCTTCAGAAATCTTAACTGACTCATTTCTCAGGAGAACTTATGGCTTCAAGTCACAGC
TCGTTTTAAAGATGGACATAACCTGAGACCTCTGATGGCTTCAACTTTGAACTG
GATGTGGACACTTTCTCAGATGACAGAATTACTCCAACCTCCCTTGCAGTTGCTT
CCTTCTTAAAGCTTGTATCTTCTTAAAGCTTTCTCAGAAAGC
TTGCCATGCCAGCTTACAGGCTTACCCGGGAGAGCAAGGCCACTGAGGCTGGGGTGGAAACC
CCCCAGGGCCAGTCTCAGGCTTACCCGGGAGAGCAAGGCCACTGAGGCTGGGGTGGAAACC
CAAGTGGCATATTACGGCATCATGCCAAATTCTTATTATCGGAGTAAAGAGA
AGACATTGGAGCAACCTCACAAGAAATGCTAGAAAAGAAGTTCTTATGTGGACCTG
AGTCCCCACGGATGAGACCTCTCTTTATAGCCAGAAGTTCCCATCAGTTGCT
GGAAAGACCTCCGGAAATTGGAGAATCCCCGATTATCATGATGGAGCAACAGAA
CTGACATCTGTCAGGGAGGCTAGGGGACTGCTGGTTCTGGAGCCATTGCTGCTGA
CCCTGAACCAAGCACCTTTCAGTCATACCCATGATCAAAGTTCATGAAACAT
```

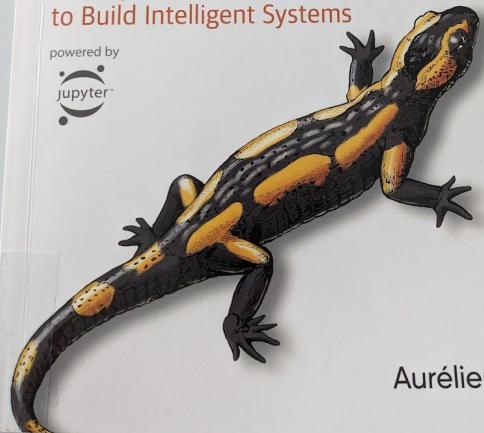
.fasta

O'REILLY®

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques
to Build Intelligent Systems

powered by
 jupyter™



Aurélien Géron

2nd Edition
Updated for
TensorFlow 2

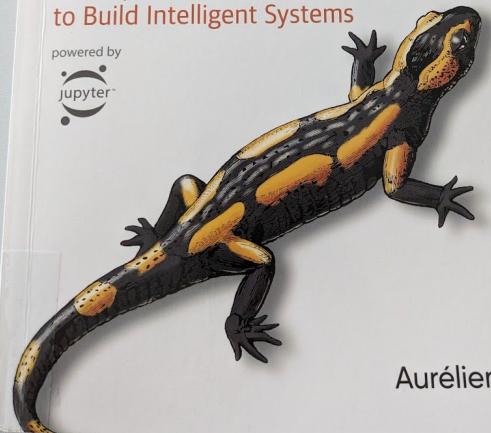
“tensor arrays”

O'REILLY®

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques
to Build Intelligent Systems

powered by
 jupyter™



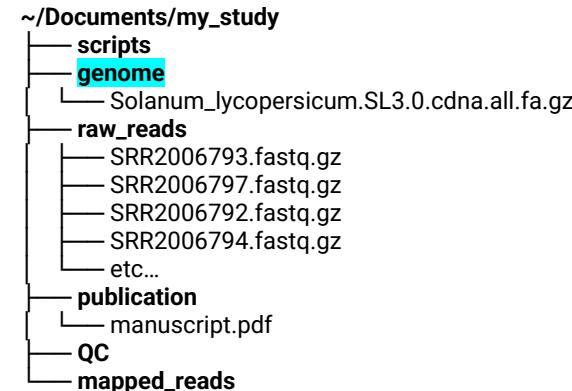
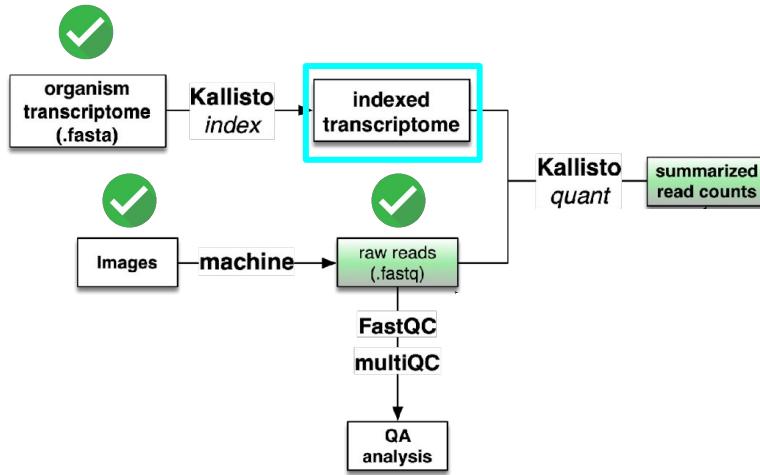
Aurélien Géron

2nd Edition
Updated for
TensorFlow 2

T

- t-Distributed Stochastic Neighbor Embedding (t-SNE), 233
- tail-heavy histograms, 51
- Talos library, 322
- target model, 639
- TD error, 630
- TD target, 630
- temperature
 - in Boltzmann machines, 775
 - in text generation, 531
- Temporal Difference Learning (TD Learning), 629
- tensor arrays, 383, 786
- TensorBoard, 317

RNAseq analysis pipeline



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’*

Building an index

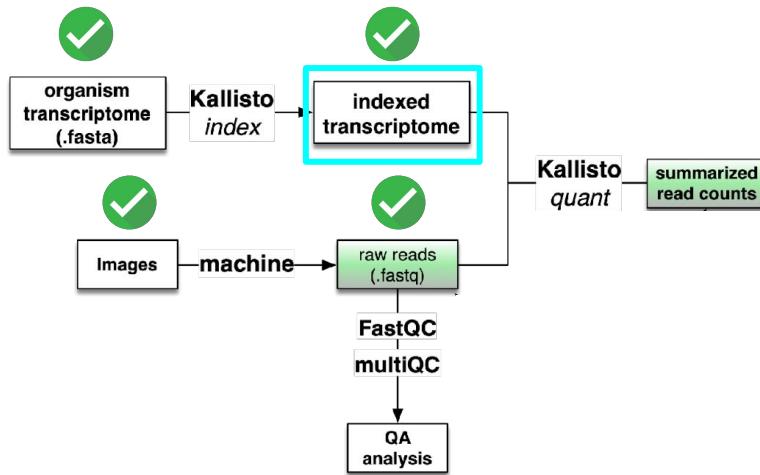
Use Kallisto to build index from reference fasta

```
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
```

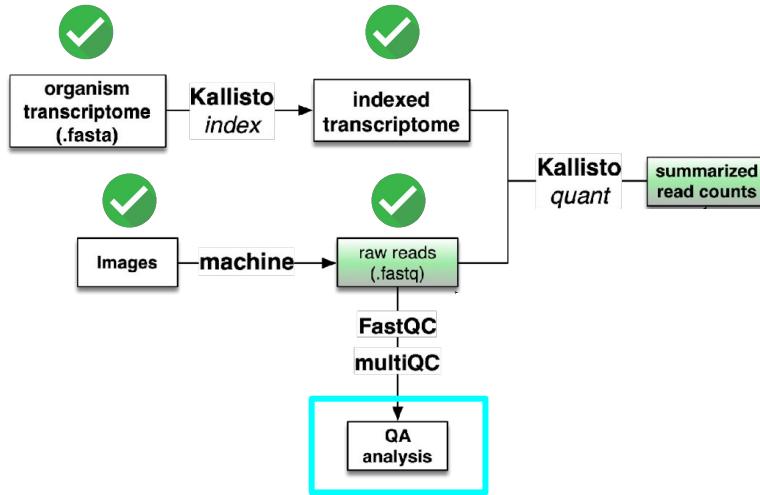
No spaces. `my_mc_name`, or `my_mcName`, or `my.mc.name` ...just not, `my mc name`

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

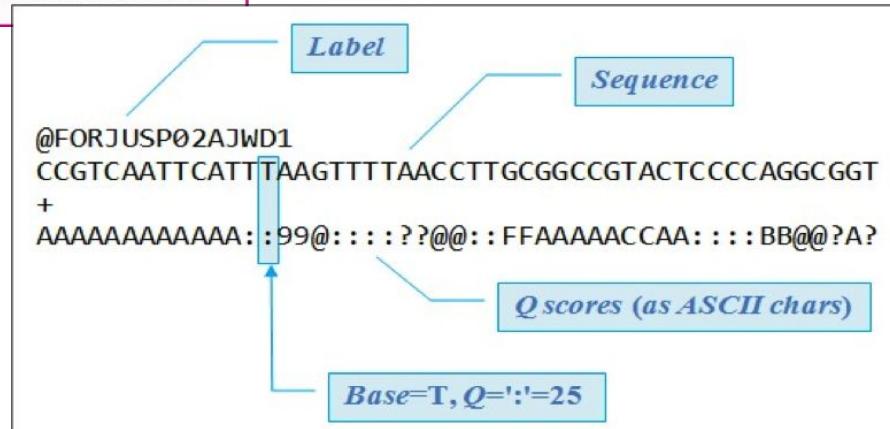
Fasta

Header
Sequence
Header
Sequence
Header
Sequence

- >VIT_201s0011g03530.1
- AATTAAGCATAAATACTCACTTACCCCTTATTTCTTATCTCTCATCATTGGTGCAG
- GACCATGAGAACAAAGCTGCAATGGGTAGGGTTCTCGCAAGGCATGCAGCCAAGACTGCATCA
- >VIT_201s0011g03540.1
- CAGGTAGCGTAGTTAACCCCTAGCGCTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
- AGCCTCTGAGACACCACTCAAACCTTCACTTAAATACACATCCCTCACACCCTTCAATT
- >VIT_201s0011g03550.1
- CATGCAAAGCTGAACCGCATGCTGATTGGTGGTAAGTGGTAGTTGAGTAATTTGACAGTGAA
- GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTCATCACGTGGGCCA

Fastq

*These are just
text files*



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

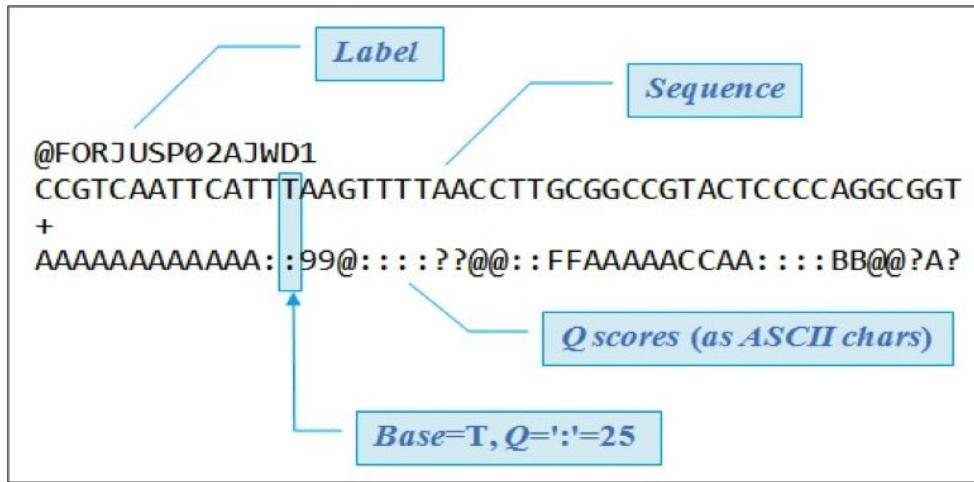
1/10000

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

Fastq

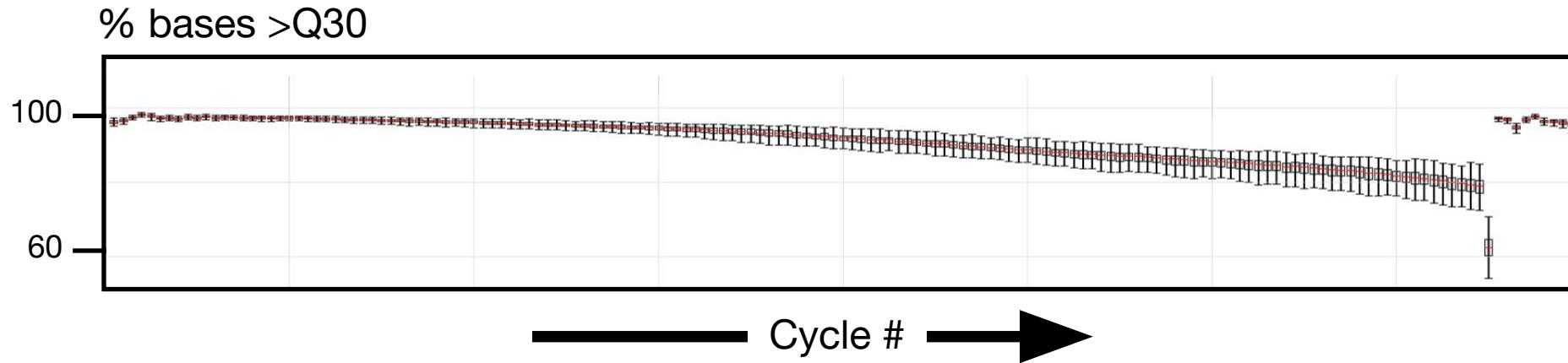
Quality score



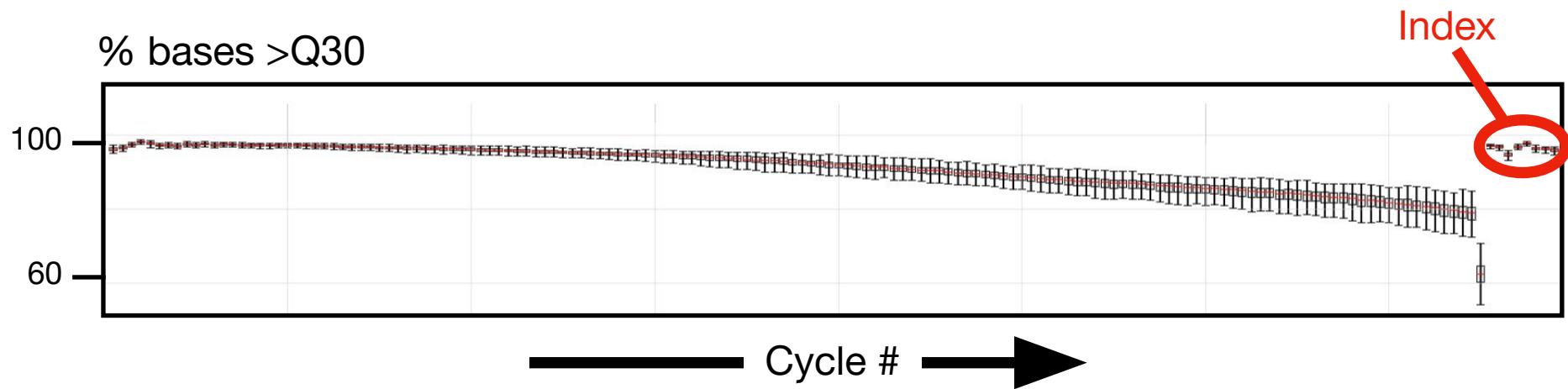
1/1000

1/10000

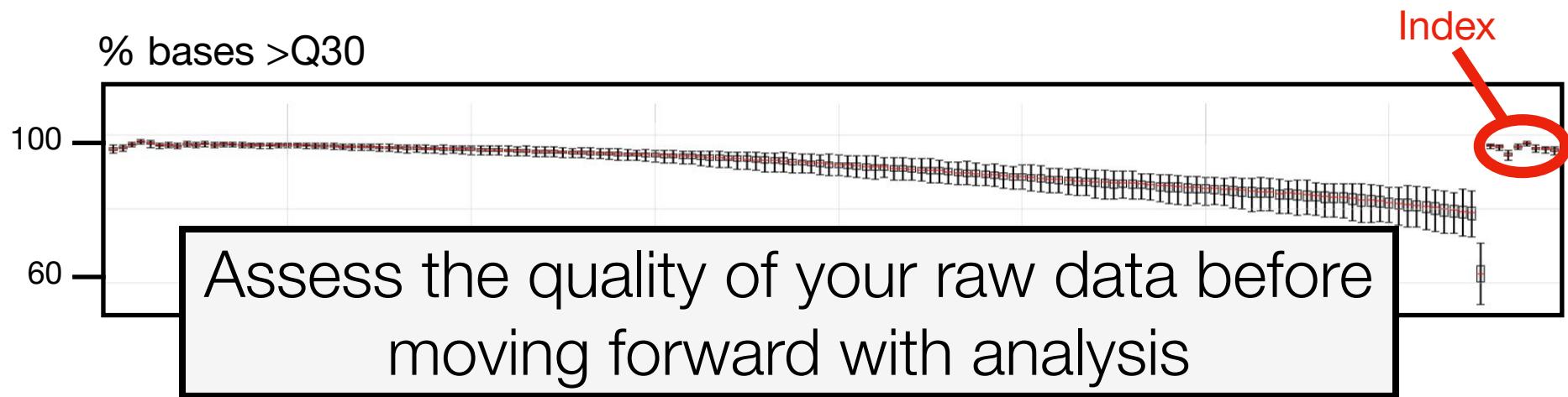
Illumina sequences are short for a reason



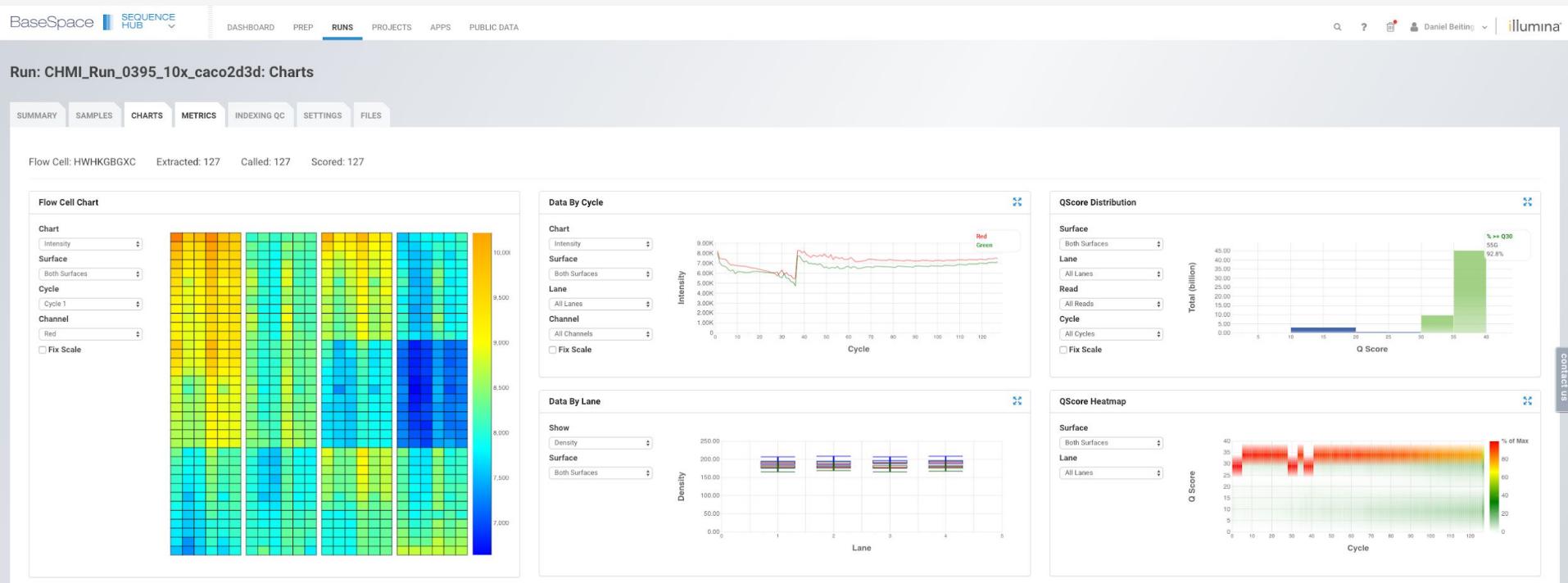
Illumina sequences are short for a reason



Illumina sequences are short for a reason



Quality assessment via Illumina's BaseSpace



Quality assessment of fastq files using fastqc



Babraham Bioinformatics



About | People | Services | Projects | Training | Publications

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Quality assessment of fastq files using fastqc

Babraham Bioinformatics

About | People | Services | Projects |

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download).
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)

About



Ubuntu

Device Name	NAM-10 >
Hardware Model	System manufacturer System Product Name
Memory	64,0 GiB
Processor	Intel® Core™ i7-6700 CPU @ 3.40GHz × 8
Graphics	NVIDIA Corporation
Disk Capacity	8,5 TB
OS Name	Ubuntu 22.04.5 LTS
OS Type	64-bit
GNOME Version	42.9
Windowing System	X11
Software Updates	>

Quality assessment of fastq files using fastqc

The screenshot shows the Babraham Bioinformatics website. In the top left, there's a logo with a stylized 'B' and the text 'Babraham Institute'. The main header has 'Babraham Bioinformatics' in large white letters. Below the header, there's a navigation bar with links for 'About | People | Services | Projects |'. A sidebar on the left is titled 'FastQC' and contains two rows of information: 'Function' (A quality control tool for high throughput sequence data) and 'Language' (Java). To the right of the sidebar is a window titled 'About' with the Ubuntu logo and the word 'Ubuntu'. Inside the window, there's a 'Device Name' field containing 'NAM-10' with a dropdown arrow, and a 'Software Updates' section.

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

OS Type	64-bit
GNOME Version	42.9
Windowing System	X11
Software Updates	>

Quality assessment of fastq files using fastqc

Babraham Bioinformatics

About | People | Services | Projects |

FastQC

Function	A quality control tool for high throughput sequence data
Language	Java

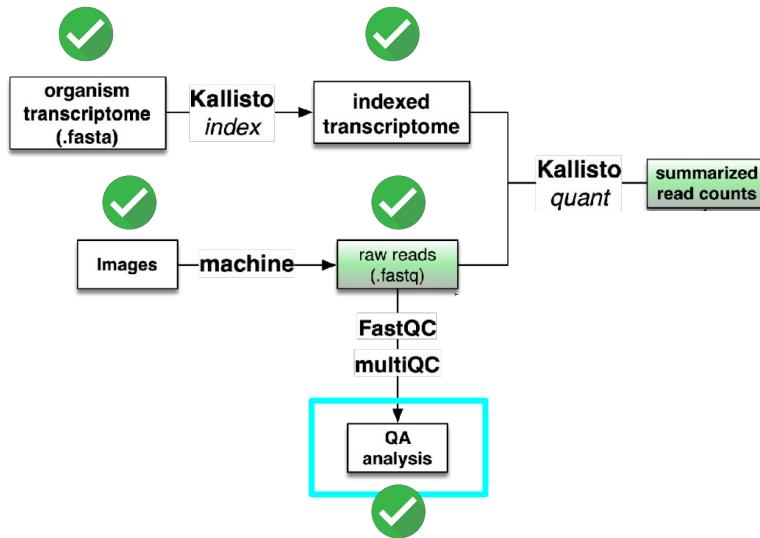
```
fastqc *.gz -t 8
```

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



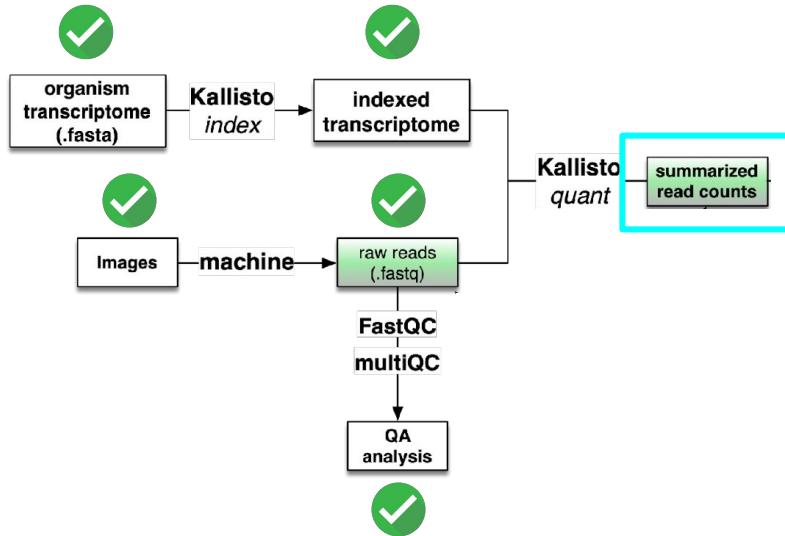
OS Type	64-bit
GNOME Version	42.9
Windowing System	X11
Software Updates	>

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
```

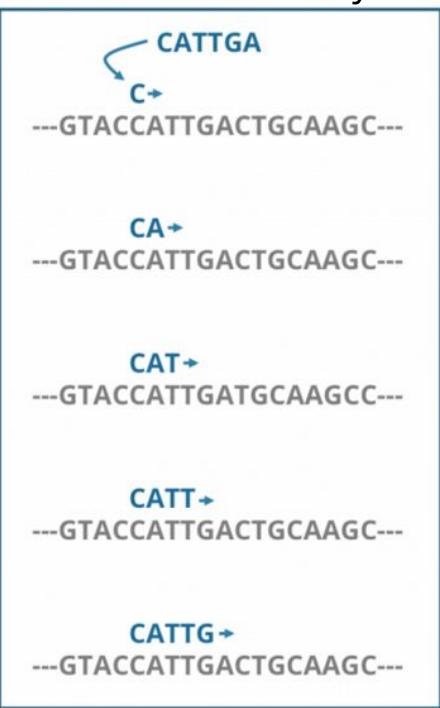
RNAseq analysis pipeline



```
~/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

'Seed and Extend' alignments

extension only

CATTG
---GTACCAATTGACTGCAAGC---

CA+
---GTACCAATTGACTGCAAGC---

CAT+
---GTACCAATTGATGCAAGCC---

CATT+
---GTACCAATTGACTGCAAGC---

CATTG+
---GTACCAATTGACTGCAAGC---

seed -> extend

VS

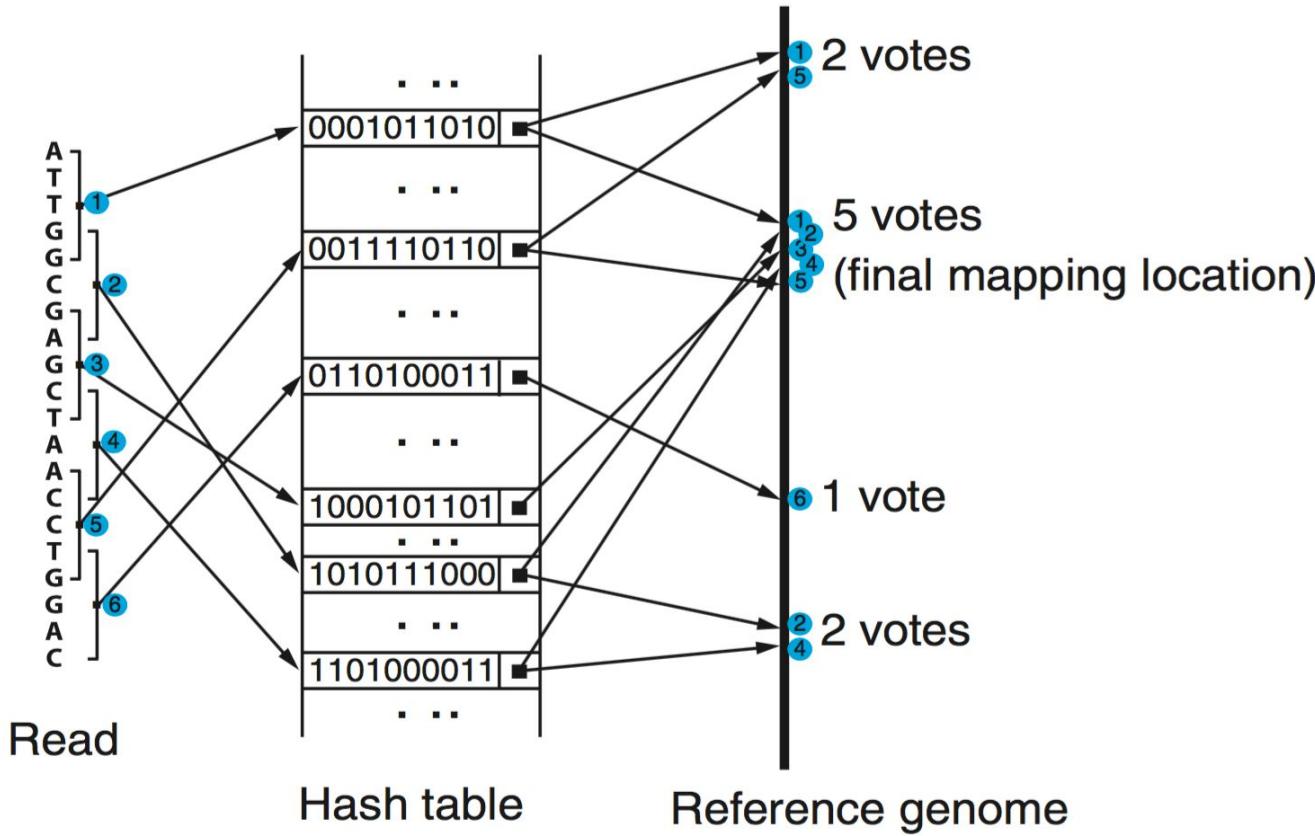
CATTG
---GTACCAATTGACTGCAAGC---

CATTG+
---GTACCAATTGACTGCAAGC---

brute force, but slow

fast(er)

R Subread aligner



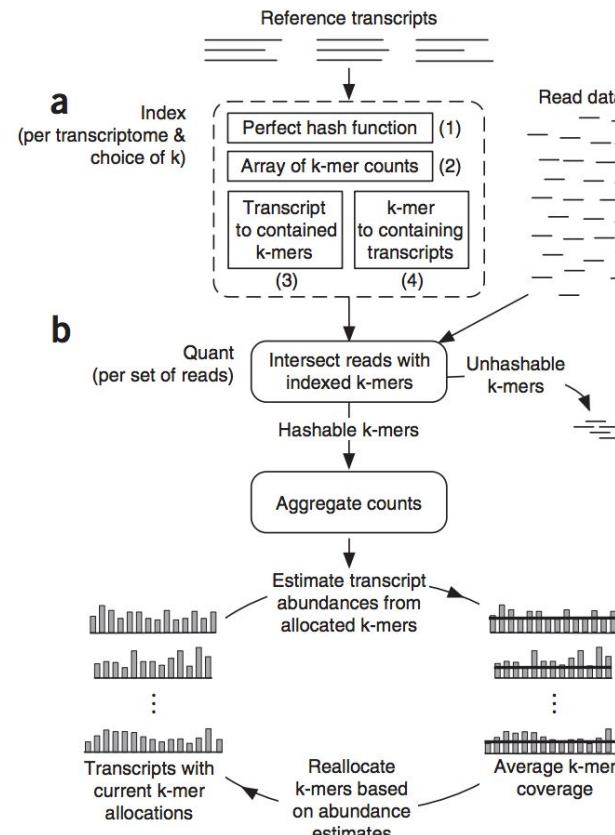
Reads are hard to align, but that's a function of their length. Why not approach the problem by shredding reads into Kmers - enter “Sailfish”

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

Rob Patro¹, Stephen M Mount^{2,3} & Carl Kingsford¹

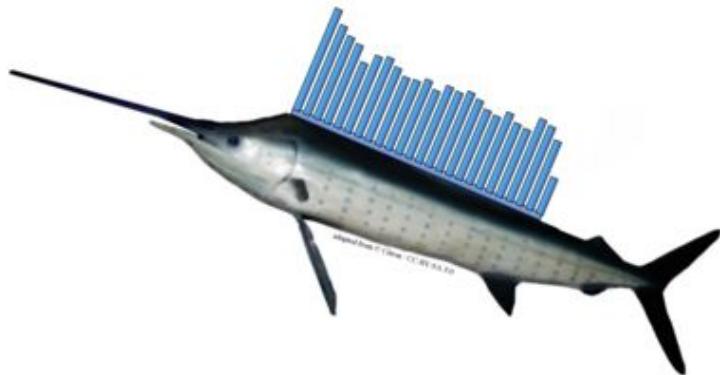
Nature Biotechnology **32**, 462–464 (2014)

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



Sailfish

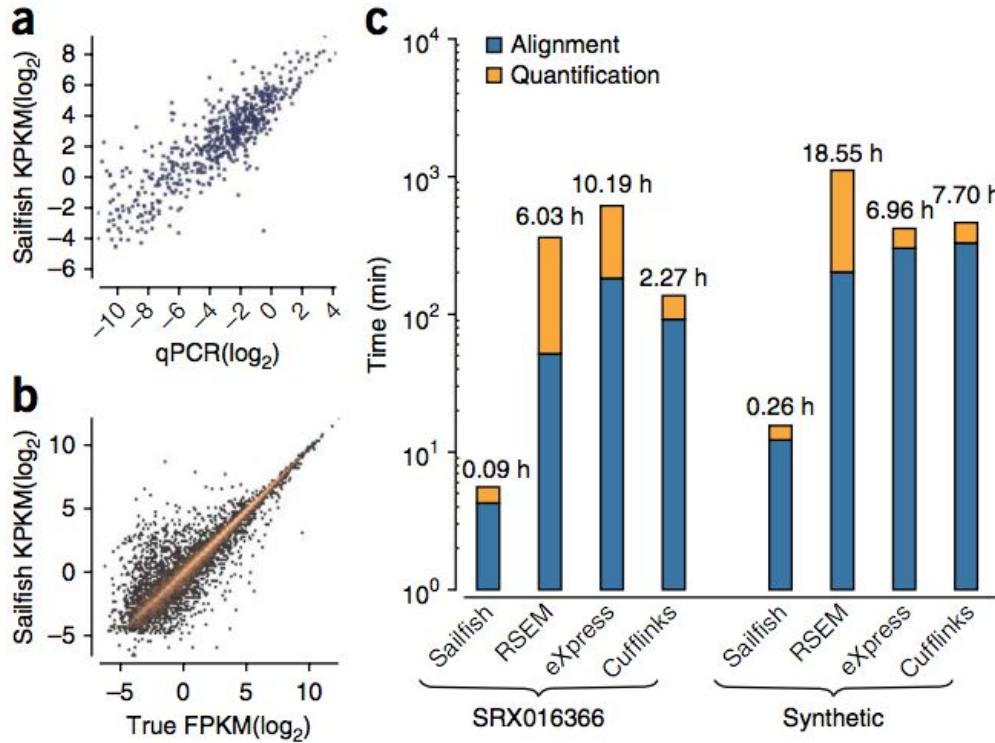
'lightweight read alignment'



“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.”

Sailfish

'lightweight read alignment'



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

BRIEF COMMUNICATIONS

nature
biotechnology

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

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray¹, Harold Pimentel², Páll Melsted³
& Lior Pachter^{2,4,5}

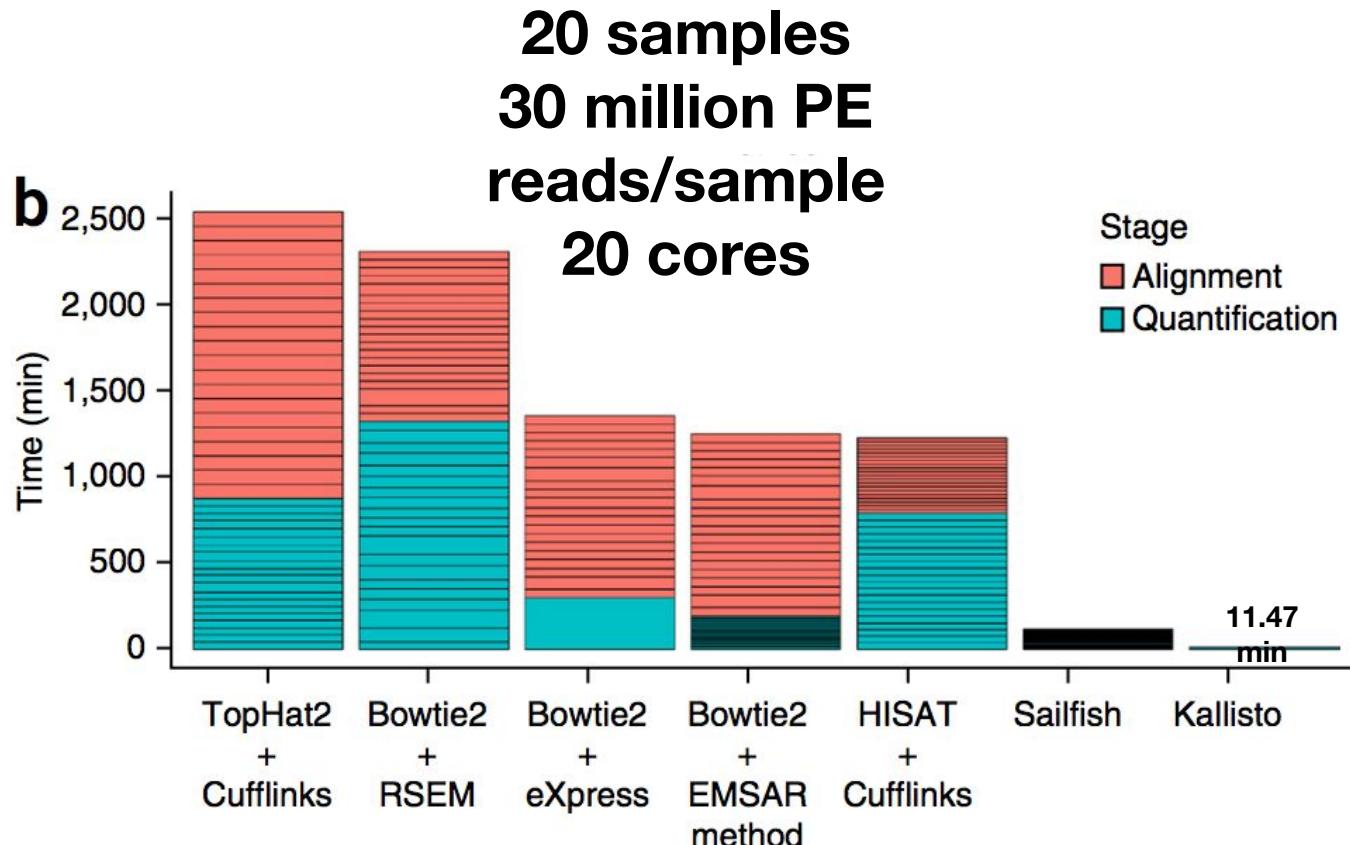
We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. 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 assembly⁸, 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 class for each vertex as the union of the sets of the k -compatibility classes containing the read path.

Kallisto - it's fast



Calling the program
and function

Name of
output folder

Single vs paired-end
and fragment size

Name and
path of index

Number of
threads to use

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 of
output folder

Single vs paired-end
and fragment size

Name and
path of index

Number of
threads to use

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.fa.index -o mapped -t 4  
--single -l 50 -s 30 SRR2006793.fastq.gz
```

Calling the program
and function

Name of
output folder

Single vs paired-end
and fragment size

Name and
path of index

Number of
threads to use

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.fa.index -o  
mapped_reads/control1 -t 4 --single -l 50 -s 30 raw_reads/SRR2006793.fastq.gz
```

Calling the program
and function

Name of
output folder

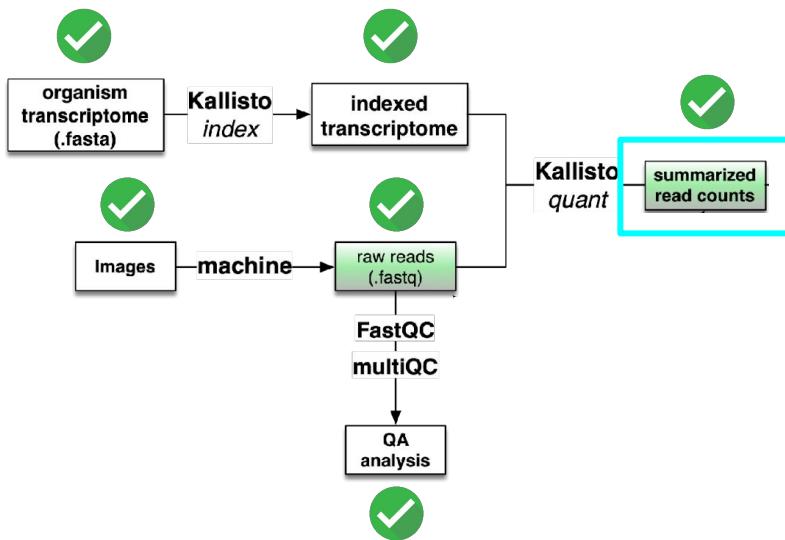
Name of fastq files to
be mapped

Name and
path of index

Number of
threads to use

```
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
    └── control2
        ├── abundance.h5
        └── abundance.tsv
    └── etc...
        ├── abundance.h5
        └── abundance.tsv
    └── run_info.json
```

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

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

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)

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

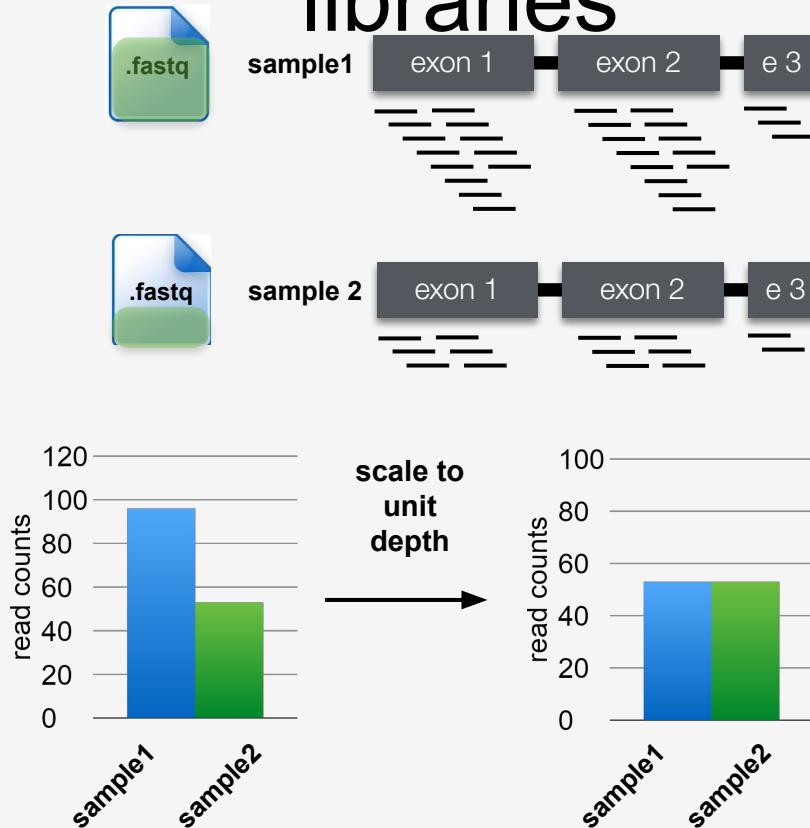
The length of a transcript after adjusting for the total number of possible positions a fragment of size X could originate from

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

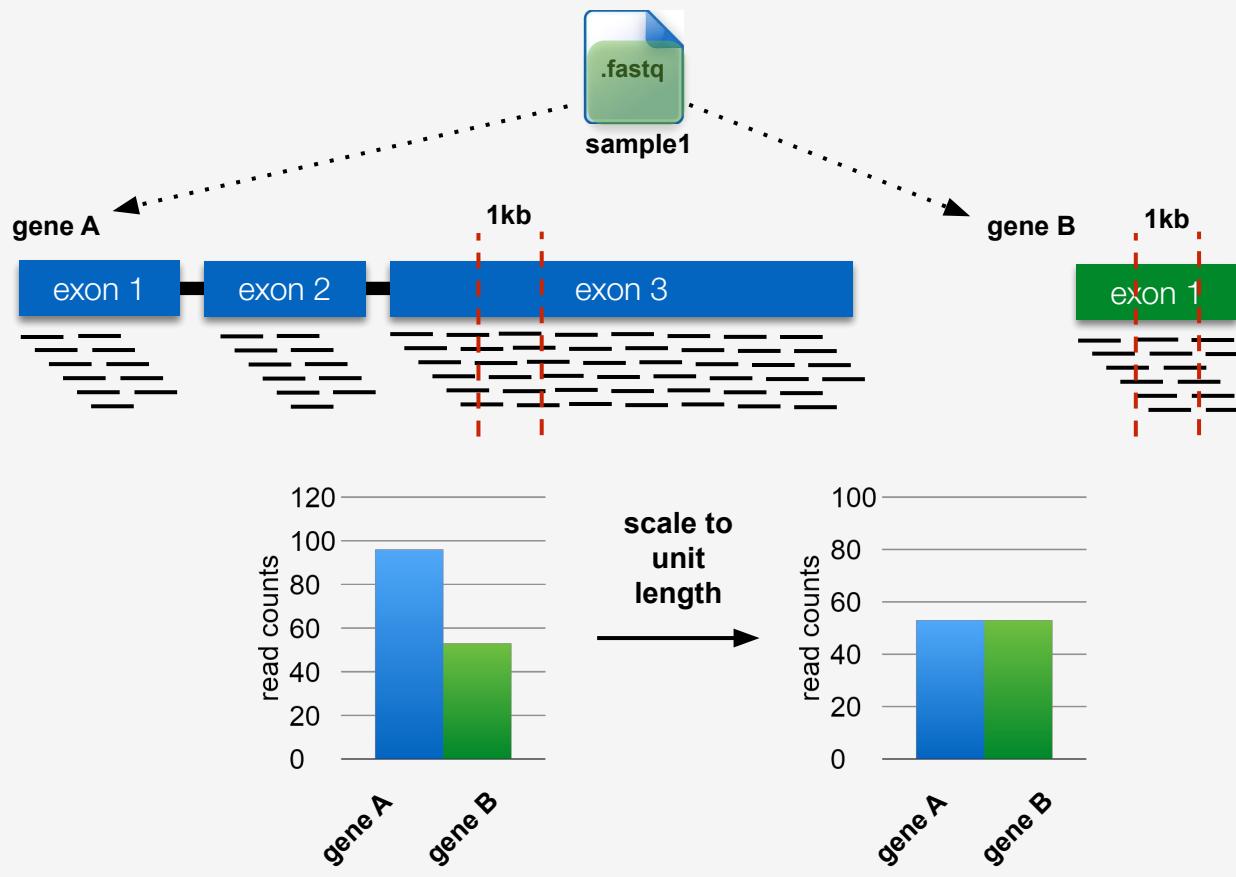
<i>transcript</i>	ATGCGTAACATG	$L_{\text{actual}} = 12$	
<i>fragment</i>	NNN	$L_{\text{fragment}} = 3$	$L_{\text{effective}} = 10$



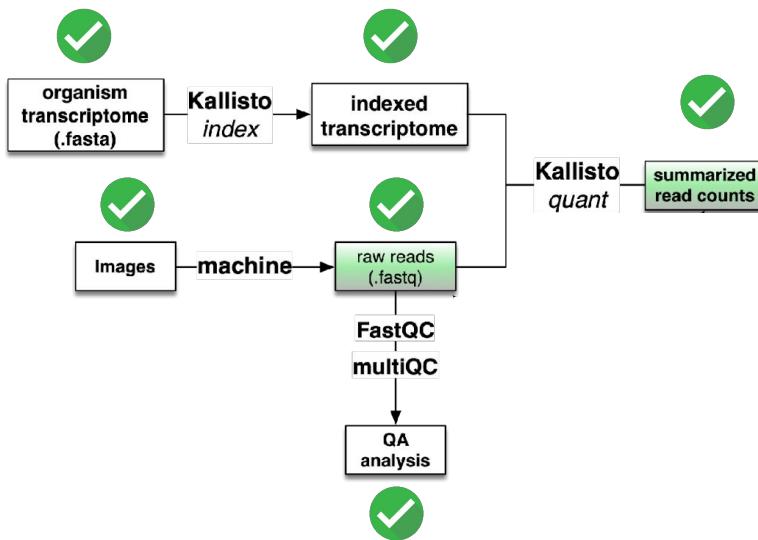
scaling units for between libraries



scaling units for within-sample comparisons



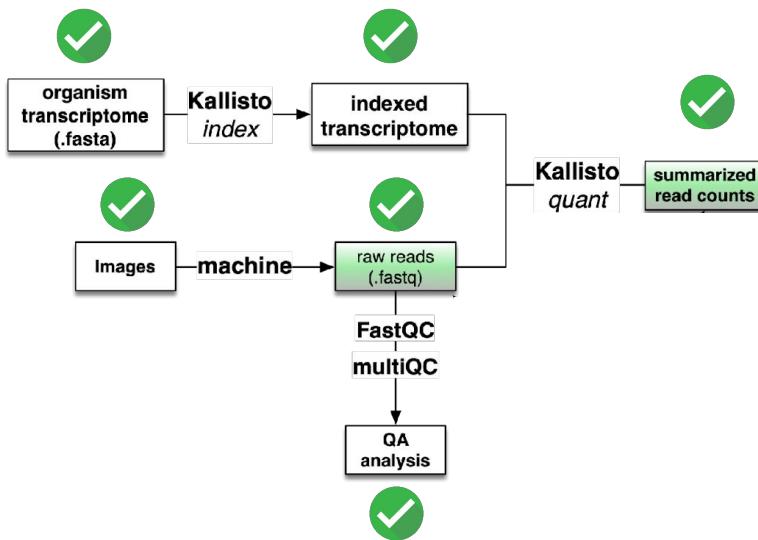
RNAseq analysis pipeline



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

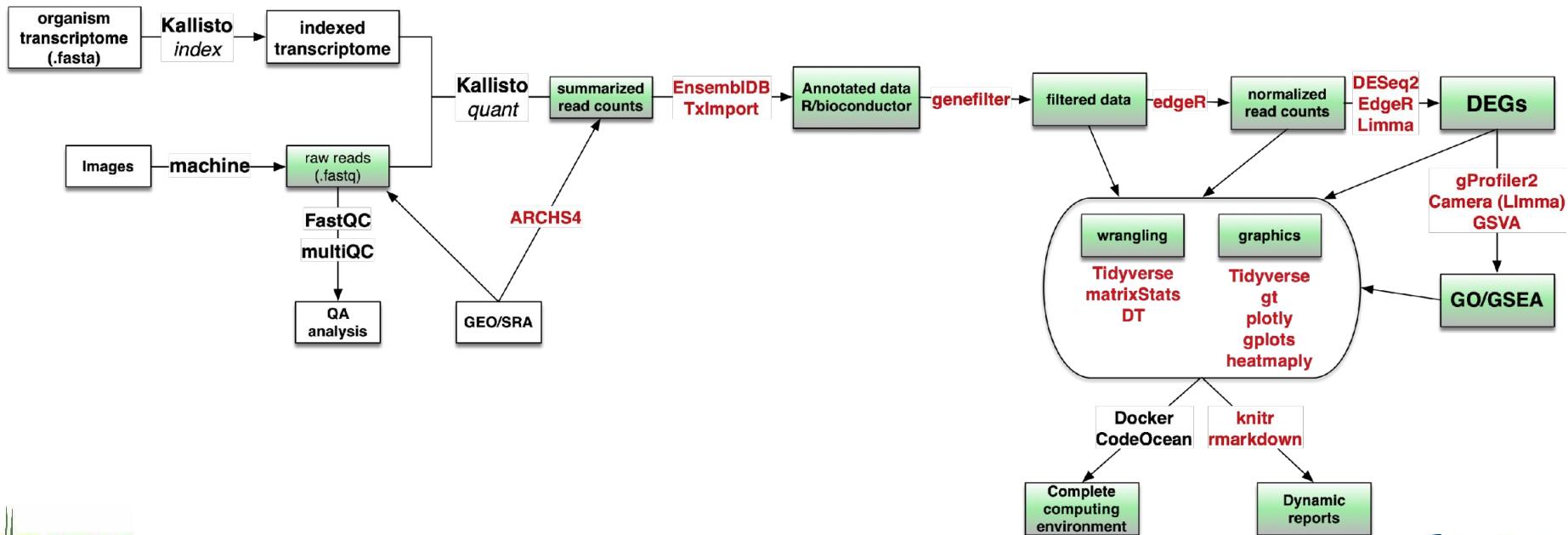


abundance - Notepad				
File	Edit	Format	View	Help
ENST00000371754.8	5235	4986	1208.55	4.16607
ENST00000469991.1	640	391	14.5506	0.639615
ENST00000371752.5	7212	6963	6866.66	16.9498
ENST00000396105.6	7209	6960	12876.7	31.7988
ENST00000371744.5	2964	2715	1972.76	12.4887
ENST00000455070.1	2122	1873	34.6566	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)

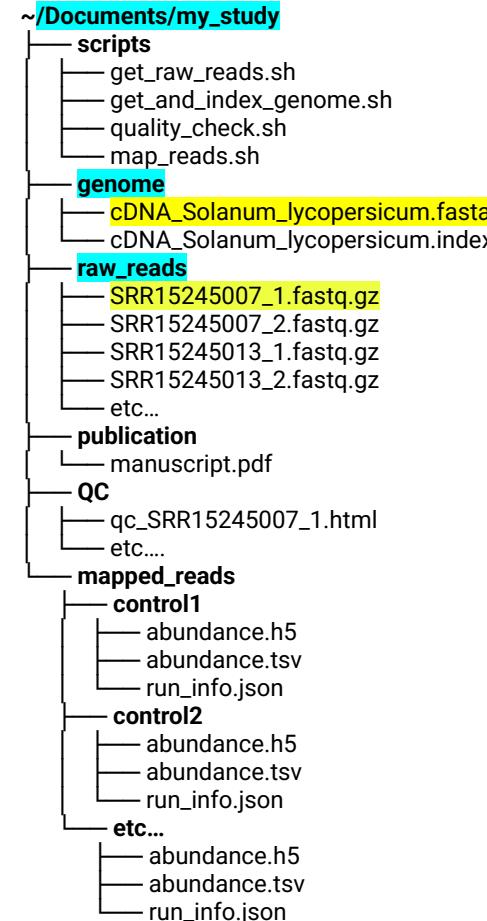
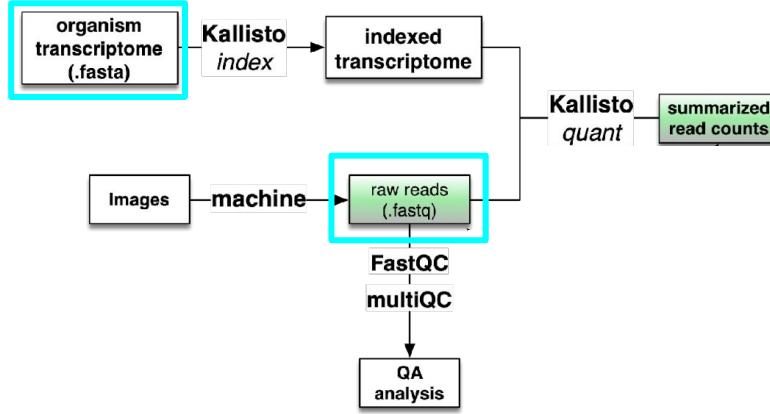
Yaaay!

The RNAseq analysis pipeline

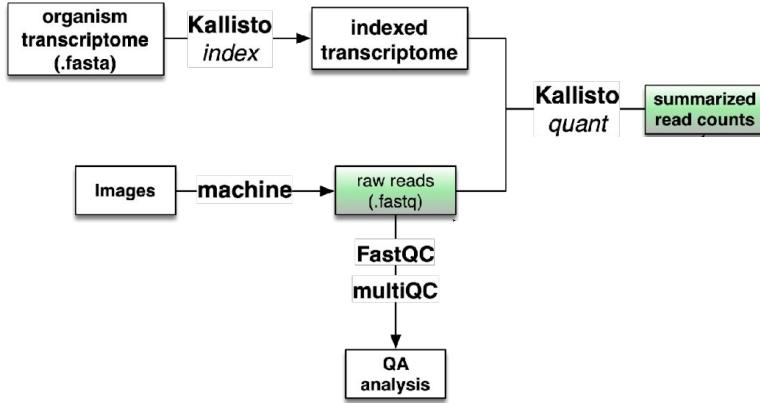


```
~/Documents/my_study
├── scripts
│   ├── get_raw_reads.sh
│   ├── get_and_index_genome.sh
│   ├── quality_check.sh
│   └── map_reads.sh
├── genome
│   ├── cDNA_Solanum_lycopersicum.fasta
│   └── cDNA_Solanum_lycopersicum.index
└── raw_reads
    ├── SRR15245007_1.fastq.gz
    ├── SRR15245007_2.fastq.gz
    ├── 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



The RNAseq analysis pipeline



The End

