

Next Generation Sequencing: Overview and Challenges

Nina Norgren, NBIS Göteborg, May 2019

Slides adapted from: Olga Vinnere Pettersson, PhD National Genomics Infrastructure hosted by ScilifeLab, Uppsala Node (UGC)

SciLifeLab

TECHNOLOGIES & SERVICES 🗸

RESEARCH V EDUCATION V

COLLABORATION

Find more information and search for what you need on the page for Technologies & Services

What is the difference between national and regional facilities?

National facilities

SciLifeLab

Affinity Proteomics

Biobank Profiling Cell Profiling Fluorescence Tissue Profiling PLA Proteomics Protein and Peptide Arrays Tissue Profiling

Bioimaging

Advanced Light Microscopy Fluorescence Correlation Spectroscopy

Bioinformatics

Bioinformatics Compute and Storage (UPPNEX) Bioinformatics Long-term Support (WABI) Bioinformatics Short-term Support and Infrastructure (BILS)

Chemical Biology Consortium Sweden

Laboratories for Chemical Biology Umeå (LCBU) The Laboratories for Chemical Biology at Karolinska Institutet (LCBKI) Uppsala Drug Optimization and Pharmaceutical Profiling (UDOPP)

Clinical Diagnostics

Clinical Biomarkers Clinical Genomics Clinical Sequencing

Drug Discovery and Development

ADME (Absorption Distribution, Metabolism Excretion) of Therapeutics (UDOPP)

Q Search for Technologies & Services

Biochemical and Cellular Screening Biophysical Screening and Characterization

Human Antibody Therapeutics In Vitro and Systems Pharmacology Medicinal Chemistry – Hit2Lead Medicinal Chemistry – Lead Identification

Protein Expression and Characterization

Functional Genomics

Karolinska High Throughput Center (KHTC)

National Genomics Infrastructure

NGI Stockholm (Genomics Applications) NGI Stockholm (Genomics Production) NGI Uppsala (SNP&SEQ Technology Platform) NGI Uppsala (Uppsala Genome Center)

Structural Biology

Protein Science Facility



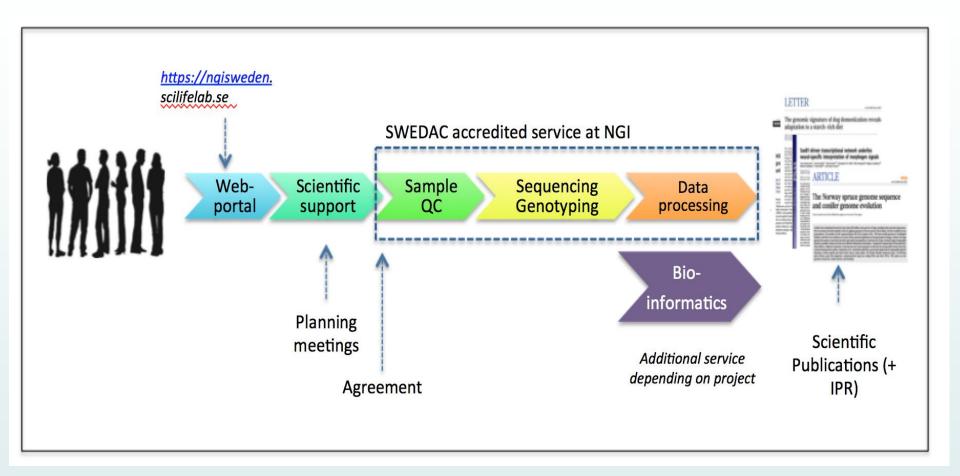








Project handling at NGI







How does a project go? Project request

√ = ×		
To: olga.pettersso	on@scilifelab.uu.se ~	
Cc:		
Bcc:		
Subject: project		
Subject. project		
Dear Olga,		
I would like to se	quence my samples on PacBio.	
How much does i	it cost? Is your queue long?	
Sincerely,)	
User		
	Admin - Contact About us Sequencing of eukaryotic genome of 50	son@igp.uu.se
	sequencing of eukaryotic genome of 50	
Order: De novo Mb in size	sequencing of eukaryotic genome of 50	O JSON
⊜ Order: De novo Mb in size Request a meel	sequencing of eukaryotic genome of 50	O JSCN (2) Edit
Order: De novo Mb in size Request a meet terester vacasst	sequencing of eukaryotic genome of 50	O JSON C' Edit ■ Delete
Order: De novo Mb in size Request a meel secretion v000311	sequencing of eukaryotic genome of 50	O JSON C' Edit B Delete OK
Admin	sequencing of eukaryotic genome of 50 ting	© JSON C'Eat B Delete © ✓ OK ✓ OK
Admin cipal Investigator name isection in the second of the sec	sequencing of eukaryotic genome of 50 ting	© JSCN C'Eat 2 Devise © ✓ OK ✓ OK ♥ no value O ✓ OK
Aarm https://www.sigator.name@ https://www.sig	sequencing of eukaryotic genome of 50 ting Doctor User doctor.user@university.se	 JSCN C'Eat Drive CK OK OK OK OK OK OK OK
Admin cipal Investigator name isection in the second of the sec	Sequencing of eukaryotic genome of 50	© JSCN C'Eat 2 Devise © ✓ OK ✓ OK ♥ no value O ✓ OK
Admin scipal Investigator name scipal Investigator name scipal Investigator phone number scipal Investigator phone number scipal Investigator phone number scipal Investigator phone number	Sequencing of eukaryotic genome of 50	C i Edit E Davie © → OK → OK
Admin Admin https://westigator.name https:/	Sequencing of eukaryotic genome of 50	© JSON © Est ■ Devia © → OK → OK
Admin scipal Investigator name scipal Investigator name scipal Investigator name scipal Investigator phone number scipal Investigator pho	Sequencing of eukaryotic genome of 50	C i Edit E Davie © → OK → OK
Admin tripal twestigator name () tripal twestigator name () tripal twestigator name () tripal twestigator enail () tripal twestigator phone number () jet tunding tring type pe ID I contact person () Project Information Project description ()	Sequencing of eukaryotic genome of 50 ting Doctor User doctocuser@university.se . Funding secured Skype Ivouid like to sequence an eukaryotic genome of 50 Mb in size. The genome highly heterosopous. We are interested in creating an elference sequence. DA has not been estimated yet, your input about choice of protocol is appreciated.	© 1504 C° Edt Bowe © ✓ OK ✓ OK ✓ OK ✓ OK ✓ OK ♥ no value O ♥ o K ♥ no value O ♥ o K
Admin Admin https://www.sigator.name attopal.lwwestigator.name attopal.	sequencing of eukaryotic genome of 50 ting Doctor User doctor.user@university.se - Funding secured Skype	JUDON C Exten C Exten C Exten C Exten C Exten C X C X C X C X O
Order: De novo Mb in size Request a meel Meetter Holdstil Aamm Copal Investigator name Copact Information Project Information Project description Copact description	sequencing of eukaryotic genome of 50 ting Doctor User doctocuser@university.se - Funding secured Skype - I would like to sequence an eukaryotic genome of 50 Mb in size. The genome highly heteroxygous. We are interested in creating a reference sequence. DNA has not been extracted yet, your input about choice of protocol is appreciated. de novo 1	 JSON C'r rati C'wata <lic'wata< li=""> C'wata</lic'wata<>
Order: De novo Mb in size Request a meel sector vacant orden	sequencing of eukaryotic genome of 50 ting Doctor User doctocuser@university.se - Funding secured Skype - I would like to sequence an sukaryotic genome of 50 Mb in size. The genome highly heterosogous. We are interested in creating a reference sequence. DNA has not been extracted yet, your input about choice of protocol is appreciated. I DNA	JICN Cred Cred
Order: De novo Mb in size Request a meel Merrier Macanti Marine Marine	sequencing of eukaryotic genome of 50 ting Doctor User doctocuse@University.se . Funding secured Skype Iwould like to sequence an eukaryotic genome of 50 Mb in size. The genome highly heterospopus. We are interested in creating a reference sequence. DAh has not been extracted yet, your input about choice of protocol is appreciated. de novo 1 DNA Not ready	J.BON C' Edit Bown C' Bown C'
Order: De novo Mb in size Request a meel terester vacast Admin Admin Admin Active Active Admin Active Admin Active Admin Active Admin Active Active Admin Admin Admin Active Admin Admin Active Admin Admi	sequencing of eukaryotic genome of 50 ting Doctor User doctocuser@university.se - Funding secured Skype	JUDON C) Edit C) Edit C) Edit C) Edit C) CK C)
Order: De novo Mb in size Request a meel Meetter HODSHI Armm Contact person O Project Information Project description Application Number of samples Application Number of samples Application Species other	sequencing of eukaryotic genome of 50 ting Doctor User doctocuse@University.se . Funding secured Skype Iwould like to sequence an eukaryotic genome of 50 Mb in size. The genome highly heterospopus. We are interested in creating a reference sequence. DAh has not been extracted yet, your input about choice of protocol is appreciated. de novo 1 DNA Not ready	J.BON C' Edit Bown C' Bown C'
Order: De novo Mb in size Request a meel terester vacast Admin Admin Admin Active Active Admin Active Admin Active Admin Active Admin Active Active Admin Admin Admin Active Admin Admin Active Admin Admi	sequencing of eukaryotic genome of 50 ting Doctor User doctocuser@university.se - Funding secured: Skype - I would like to sequence an eukaryotic genome of 50 Mb in size. The genome highly heteropopus. We are interested in creating a reference sequence. DNA has not been extracted yet, your input about choice of protocol is appreciated. de novo 1 DNA Not mady Other Neefungua spectaculus	JUSON Critical Control Contro Control Control Control Control Co





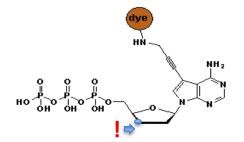
Short History of NGS

Once upon a time...

Fredrik Sanger and Alan Coulson
 Chain Termination Sequencing (1977)
 Nobel prize 1980

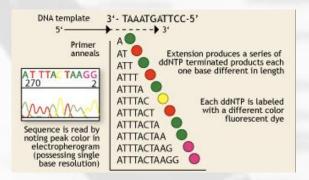
Principle:

SYNTHESIS of DNA is randomly **TERMINATED** at different points Separation of fragments that are 1 nucleotide different in size



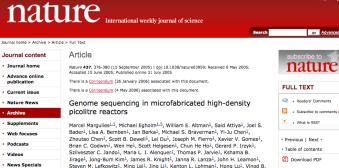
Lack of OH-group at 3' position of deoxyribose





2006: NGS was born





Journal information George T. Roth1, Gary J. Sarkis1, Jan Fredrik Simons1, John W. Simpson1, Maithreyan Srinivasan¹, Karrie R. Tartaro¹, Alexander Tomasz³, Kari A. Vogt¹ + About the journal Greg A. Volkmer¹, Shally H. Wang¹, Yong Wang¹, Michael P. Weiner⁴, Pengguang + For authors Yu1, Richard F. Begley1 & Jonathan M. Rothberg1

+ Table of contents Download PDF Steven M. Lefkowitz¹, Ming Lei¹, Jing Li¹, Kenton L. Lohman¹, Hong Lu¹, Vinod B. View interactive PDF in Makhijani¹, Keith E, McDade¹, Michael P, McKenna¹, Eugene W, Myers², Elizabeth ReadCube Nickerson¹, John R. Nobile¹, Ramona Plant¹, Bernard P. Puc¹, Michael T. Ronan¹,

Share this article CrossRef lists 376 articles

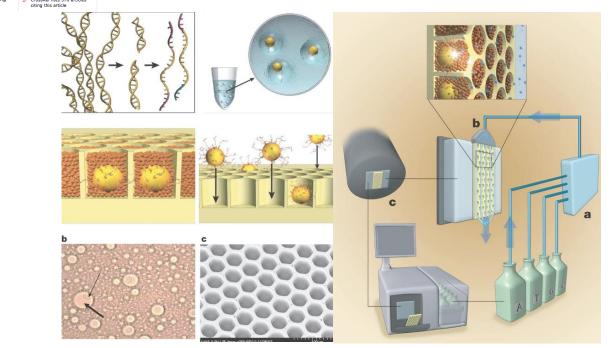
() Login 1/ Carl

go Advanced sea

in parallel

1 mln reads sequenced per run

Thousands of molecules sequenced





Roche 454 GS FLX

Since the beginning of Genomics:



First genome: virus ϕ X 174 - 5 368 bp (1977)



First organism: Haemophilus influenzae - 1.5 Mb (1995)



First eukaryote: Saccharomyces cerevisiae - 12.4 Mb (1996)



First multicellular organism: Cenorhabditis elegans - 100 MB (1998-2002)



First plant: Arabidopsis thaliana - 157 Mb (2000)

... prices go down

Human genome sequencing:

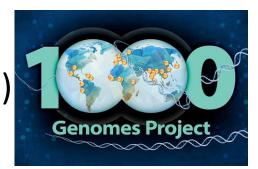
2004: Genome of Craig Wenter costs 70 mln \$

• Sanger's sequencing

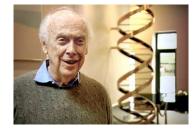
2007: Genome of James Watson costs 2 mln \$

• 454 pyrosequencing

2014: Ultimate goal: 1000 \$ / individual 2016: Illumina Xten: Almost there! (1200 \$) 2017: NovaSeq: "Hold my beer..." (100 \$)



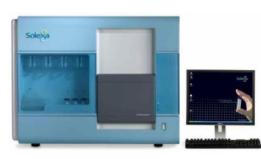






... paradigm changes

- From single genes to complete genomes
- From single transcripts to whole transcriptomes
- From single organisms to complex metagenomic pools
- From model organisms to the species you are studying
- Personal genome = personalized medicine

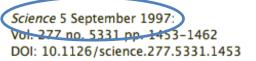








... scientific value diminishes



ARTICLES

IF 31.6

< Prev | Table of Contents | Next >

The Complete Genome Sequence of Escherichia coli K-12

Frederick R. Blattner^{*}, Guy Plunkett III^{*}, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau and Ying Shao

Journal of Biotechnology Article in Press, Corrected Proof - Note to users

doi:10.1016/j.jbiotec.2010.12.018 | How to Cite or Link Using DOI

Permissions & Reprints

IF 2.9

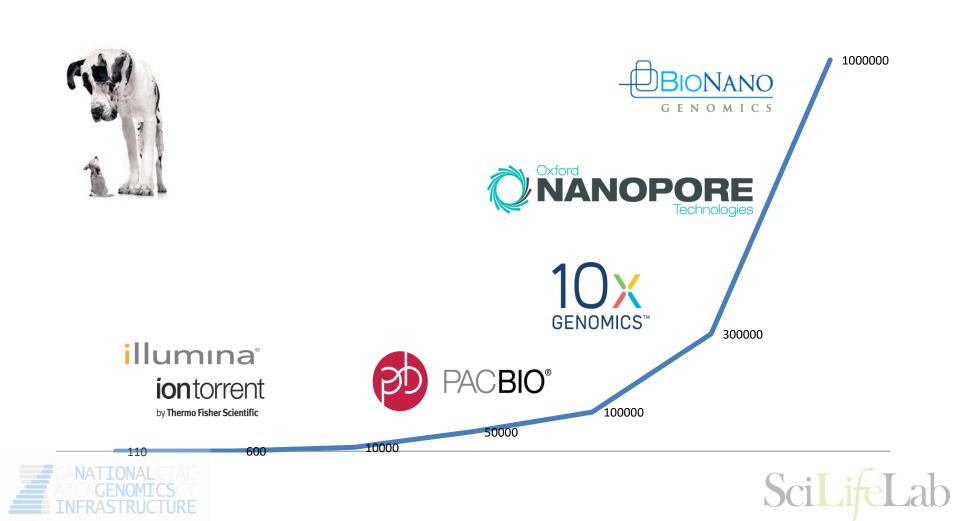
The complete genome sequence of the dominant Sinorhizobium meliloti field isolate SM11 extends the S. meliloti pan-genome

Susanne Schneiker-Bekel^a, Daniel Wibberg^a, Thomas Bekel^b, Jochen Blom^b, Burkhard Linke^b, Heiko Neuweger^b, Michael Stiens^{a, c},

Frank-Jörg Vorhölter^a, Stefan Weidner^a, Alexander Goesmann^b, Alfred Pühler^a and Andreas Schlüter^{a, 🍐 🛸}

Current Technologies

Read length



Illumina

Instrument	Yield and run time	Read Length	Error rate	Error type
HiSeq2500	120 Gb – 600 Gb 27h or standard run	110x110 (250x250)	0.1%	Subst
MiSeq	540 Mb – 15 Gb (4 – 48 hours)	up to 350x350	0.1%	Subst
HiSeqXten	800 Gb - 1.8 Tb (3 days)	150x150	66	66
NovaSeq 6000	250 Gb – 3 Tb	150x150	66	66

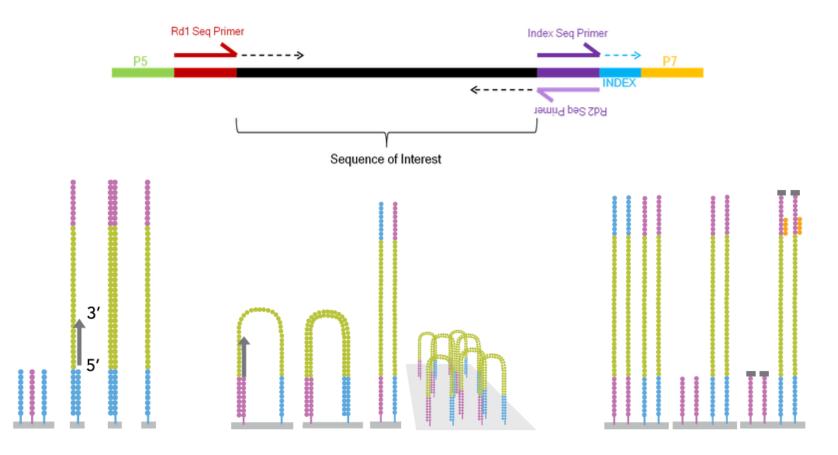
Main applications

- Whole genome, exome and targeted reseq
- Transcriptome analyses
- Methylome and ChiPSeq
- Rapid targeted resequencing (MiSeq)
- Human genome seq (Xten)





Illumina: bridge amplification



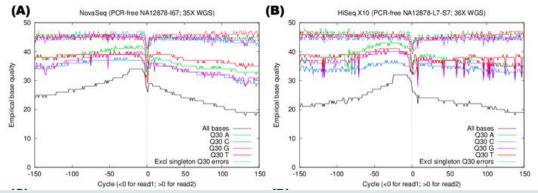
https://www.youtube.com/watch?v=fCd6B5HRaZ8



NovaSeq 6000

- NGI has five instruments
- Flexible and scalable using multiple flow cell types
- Quick and easy operation using RFID labeled reagent cassettes
- Onboard clustering and automatic washing minimises hands on time during runs
- 2 color chemistry T=Green

C=**Red** A=**Green/Red** G=no signal

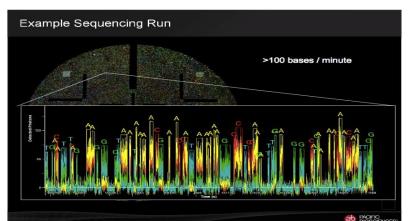




PacBio

Instrument	Yield/cell and run time	Read Length	Error rate	Error type
RSII	250 Mb – 1.8 Gb 30 - 600 min	250 bp – 60 kb <i>(78 kb)</i>	15 % (single pass) 0.0001% (circular consensus)	Indels, random
SEQUEL	2-14 Gb 30-2400 min	250 bp – 80 kb (160 kb)	as RSII	Indels, random

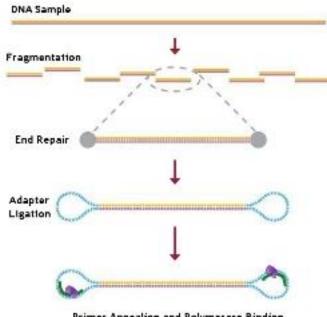
Single-Molecule, Real-Time DNA sequencing



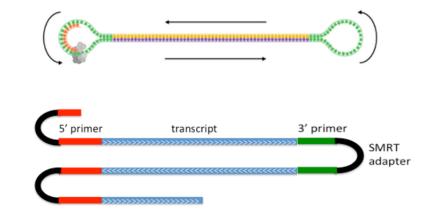


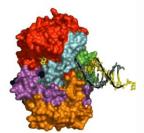
PacBio: SMRT - technology





Primer Annealing and Polymerase Binding to SMRTbell Template





Time



SMRT = Single Molecule Real Time

SMRT sequencing: common misconceptions

High error rate?

Irrelevant, because errors are random

Depending on coverage

Examples:

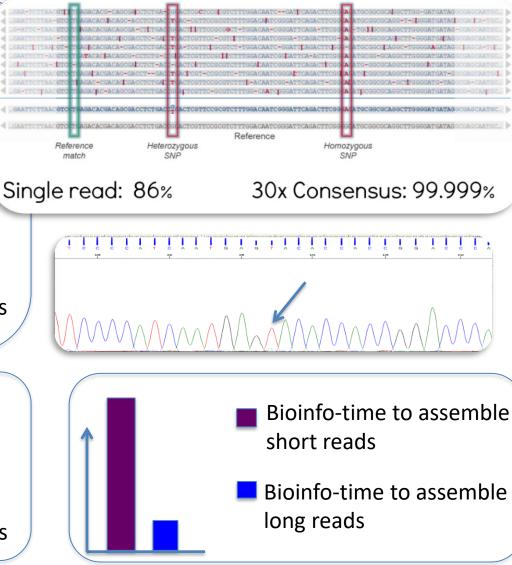
- 8 Mb genome, 8 SNPs detected
- 65 kb construct: 100% correct sequence
- Detection of low frequency mutations

Single-molecule reads without PCR-bias

High price?

Not for small genomes

Better assembly quality



Oxford	Nanopore	DNA DOUB	A flow of ions through the pore creates a current.
Flow Cells run in parallel	Yield - run time	One protein unzips the DNA helix into two strands.	Each base blocks the flow to a different degree, altering the current.
MinION (1)	1 – 10 Gb / cell	A second protein creates a pore in the membrane and holds	[Լտ]]տլե
GridION (5)	5 – 50 Gb / 5 cells	an *adapter* molecule.	O The adapter molecule keeps bases in place long enough for them to be
PromethION (12 - 24 - 48)	20 – 100 Gb / cell	MEMBRANE	identified electronically.



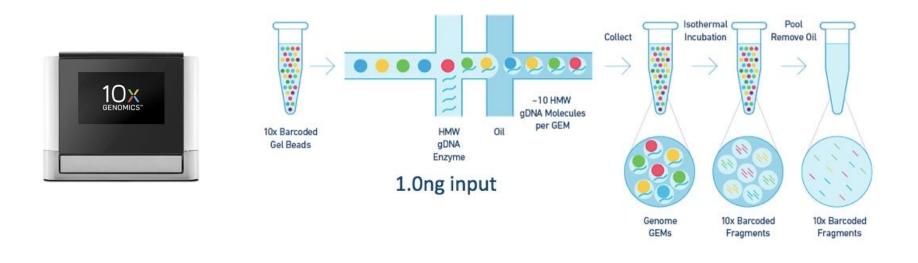
Reads up to 6-8 Gb 10-15% error rate Life time 5 days

Longest reads: beyond 1 Mb



10x Genomics (Chromium)







Fragment length: 50 kb – 100+ Kb



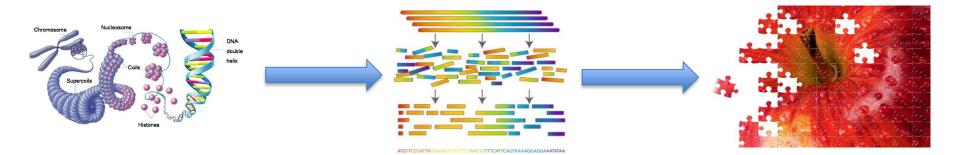
NGS Applications

NGS/MPS applications

- Whole genome sequencing:
 - De novo sequencing
 - Re-sequencing
- Transcriptome sequencing:
 - mRNA-seq
 - miRNA
 - Isoform discovery
- Target re-sequencing
 - Exome
 - Large portions of a genome
 - Gene panels
 - Amplicons

Whole genome sequencing: de novo

De novo: used to assemble a genome without previous reference



Conventional strategy (Golden Standard):

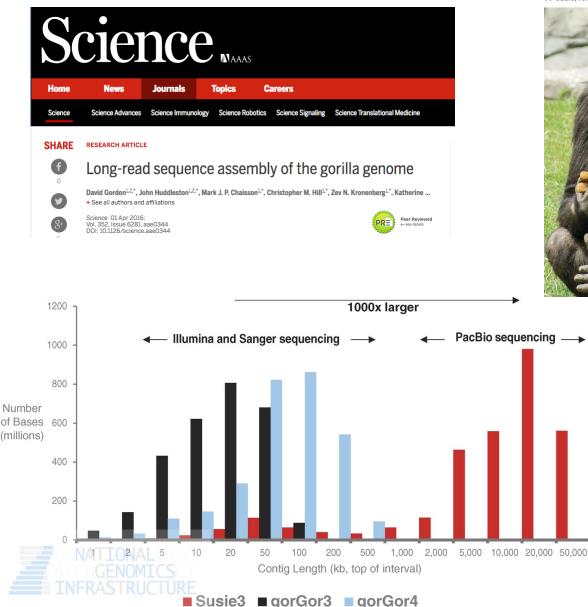
Illumina 50x sequencing on HiSeqX or NovaSeq, several insert sizes (+ Mate Pairs)

Current recommendation* (Platinum genome):

100x PacBio (ONT) only + Hi-C (coverage depends on heterozygocity) *Plus RNA-seq data for annotation*

* 2019-02-05

De novo – do it with long reads!

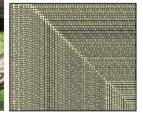


A Susie, reference sample



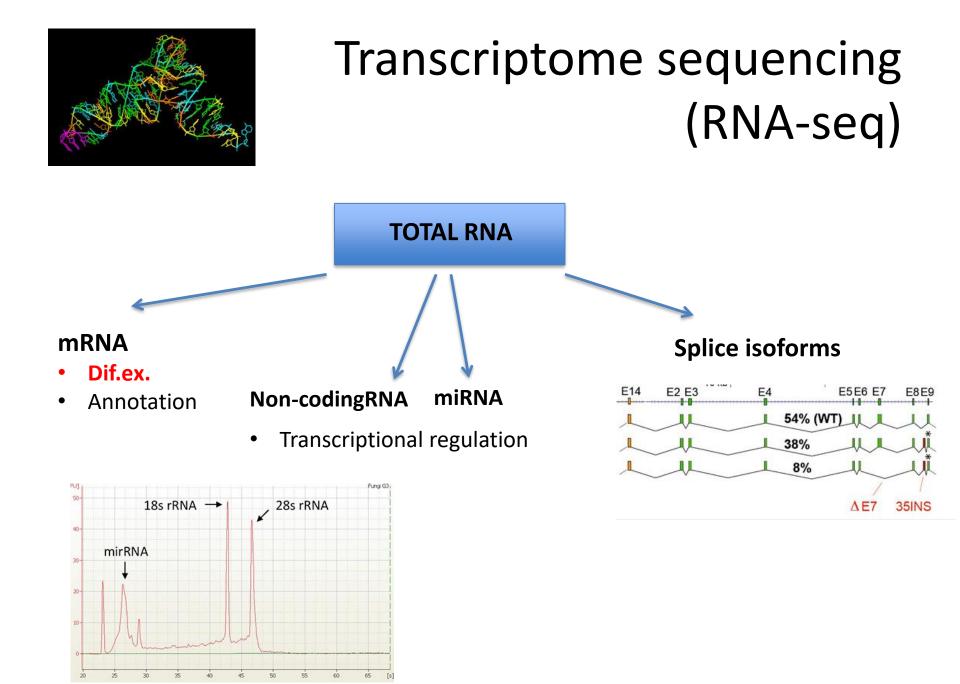
0 5 10 15 20 25 30 35 Contig size (Mbp)

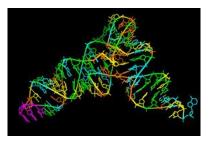
C Short-read assembly (gorGor3)



Beware: up to 80% of novel structural variants can be missing from short-read data.

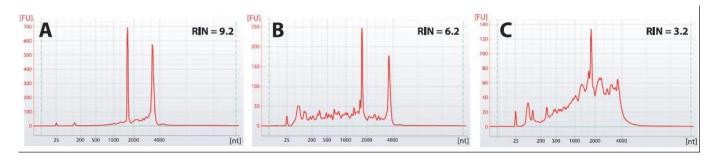
\Rightarrow Sequence fewer genomes, but with long reads



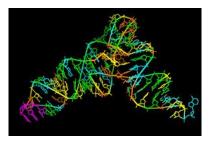


RNA-seq experimental setup

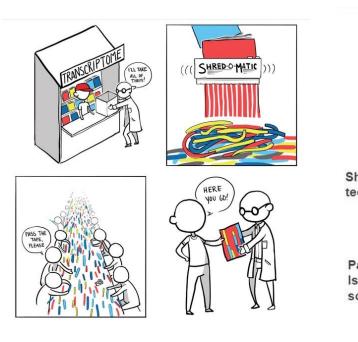
- mRNA only: any kit
- mRNA and miRNA: only specialized kits
- Always use DNase!
- RIN value above 8.

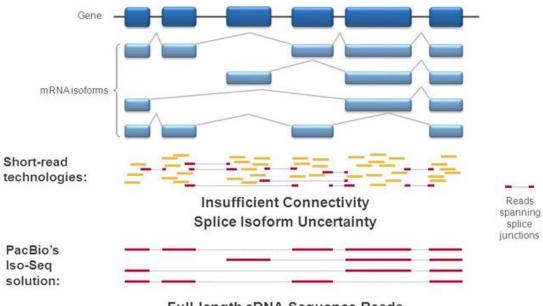


- CONTROL vs experimental conditions
- Biological replicates: 4 strongly recommended



RNA-seq with long reads





Full-length cDNA Sequence Reads Splice Isoform Certainty – <u>No Assembly Required</u>

NATURE METHODS | NEWS AND VIEWS

-< 🖶

Genomics: the state of the art in RNA-seq analysis

lan Korf

Nature Methods **10**, 1165–1166 (2013) | doi:10.1038/nmeth.2735 Published online 26 November 2013 PacBio Iso-seq: full-length transcriptome seq

Coming soon: direct RNA-seq on ONT

Main types of equipment & applications





Illumina HiSeq NextSeq, HiSeqX10, MiSeq, MiniSeq, NovaSeq

lon S5 XL



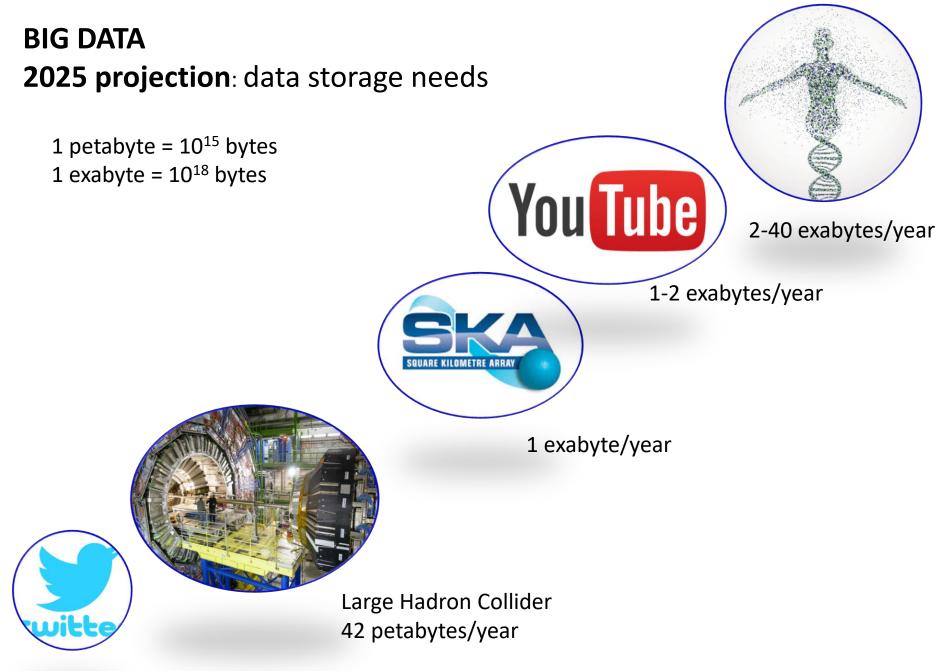
PacBio RSII SEQUEL

Short paired reads HIGH throughput

Human WGS Re-sequencing 30x mRNA and miRNA De novo transcriptome Exome ChIP-seq Short amplicons Methylation Short single-end reads FAST throughput

mRNA and miRNA Exome ChIP-seq Short amplicons Gene panels Clinical samples Ultra-long reads FAST throughput

Long amplicons Re-sequencing De novo sequencing Novel isoform discovery Fusion transcript analysis Resolving haplotypes Clinical samples



1-17 petabytes/year



Thanks for listening! Questions?

support@ngisweden.se



