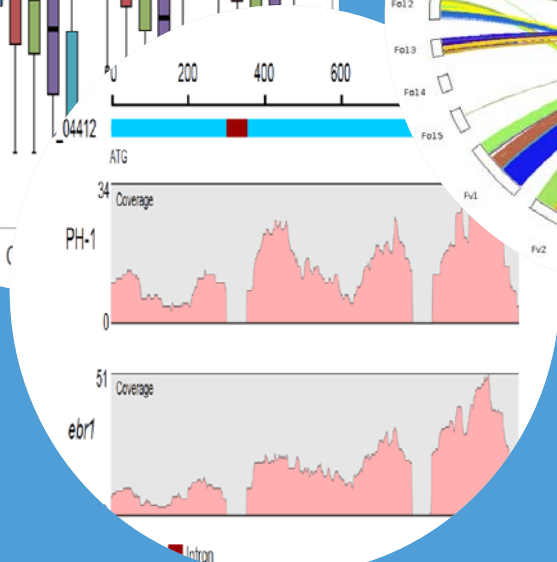
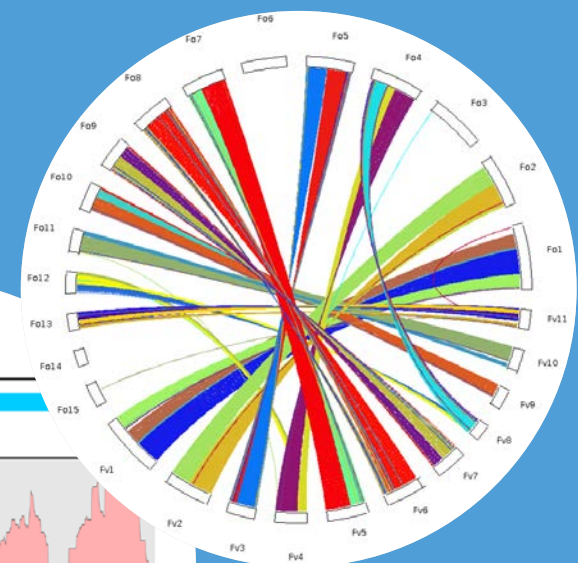
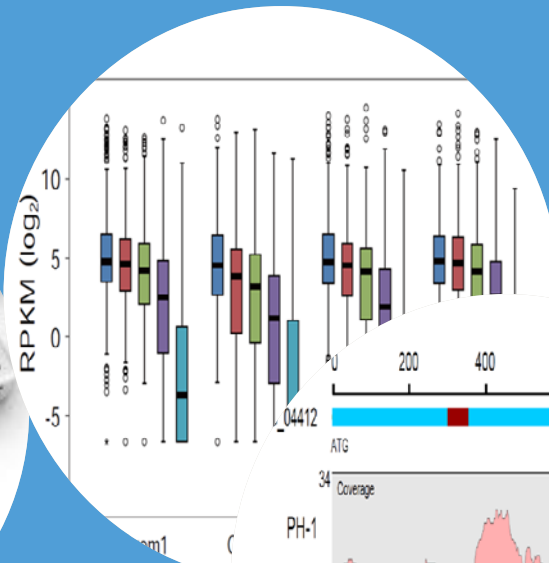


What's in the pipeline?

Understanding the basics of NGS

Theo van der Lee

Workshop on the use of NGS technologies for plant pest diagnostics,
Bari, 22th November 2017



Introduction

Theo van der Lee

Senior Scientist Biointeractions

Genetics, molecular diagnostics

Population and functional genetics

Involved in genomics research since 2000



Citations	19603
h-index	35
i10-index	59
>1000/year per year	Google Scholar

The Plant Sciences Group



Research at Biointeractions

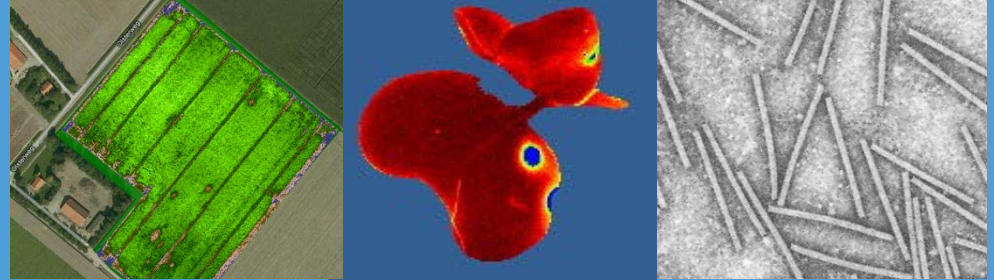
- Plant parts/Seeds/Bulbs
- Soil/water/air
- Rhizosphere
- Endophytes
- Biocontrol

- Detection
- Microbiology
- Bioassays
- Bioinformatics



Detection and quantification:

- imaging



- On-site: LAMP



- Real-time PCR

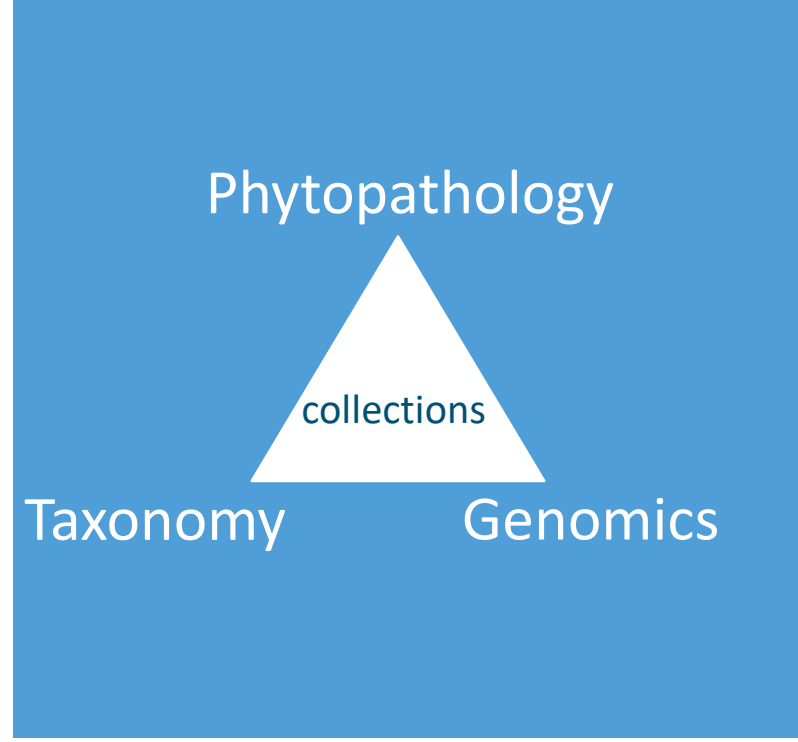
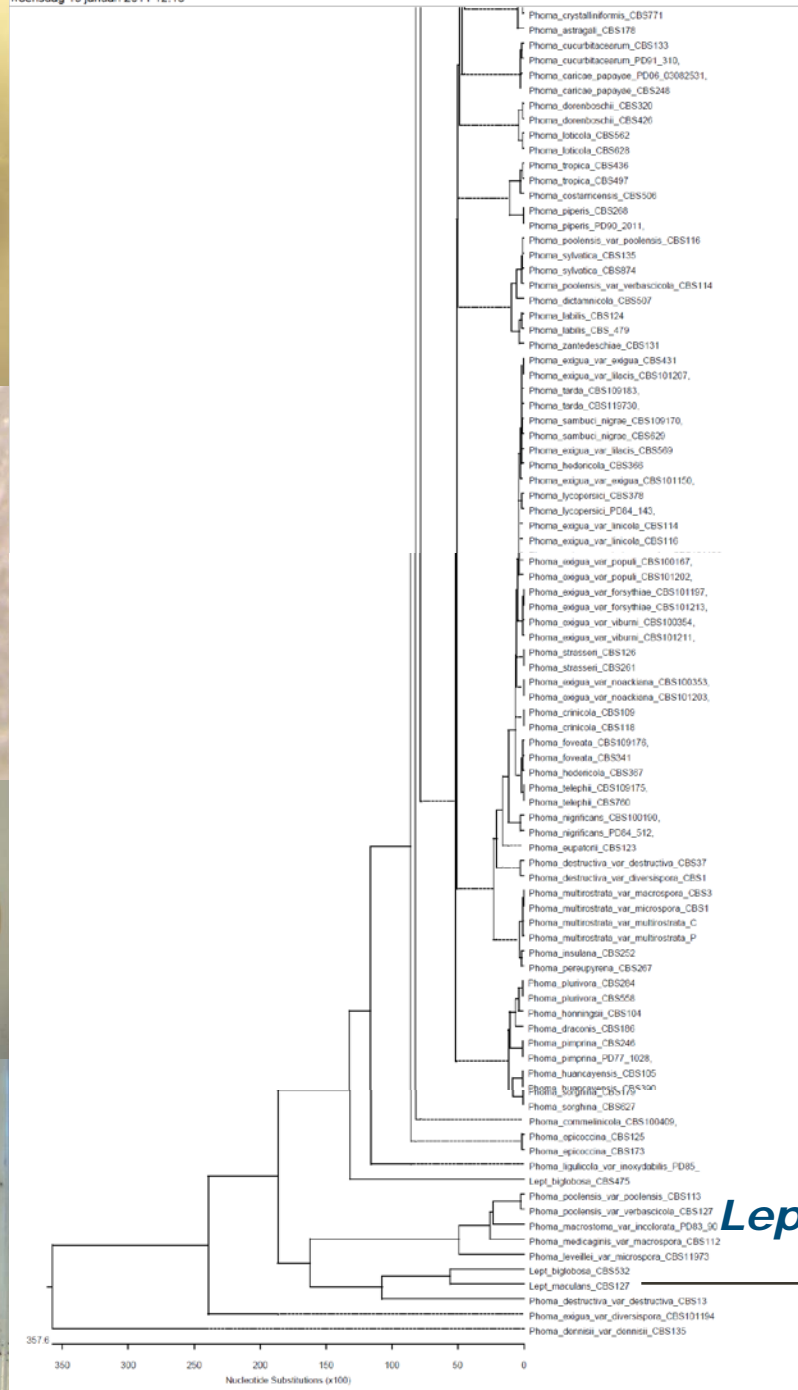
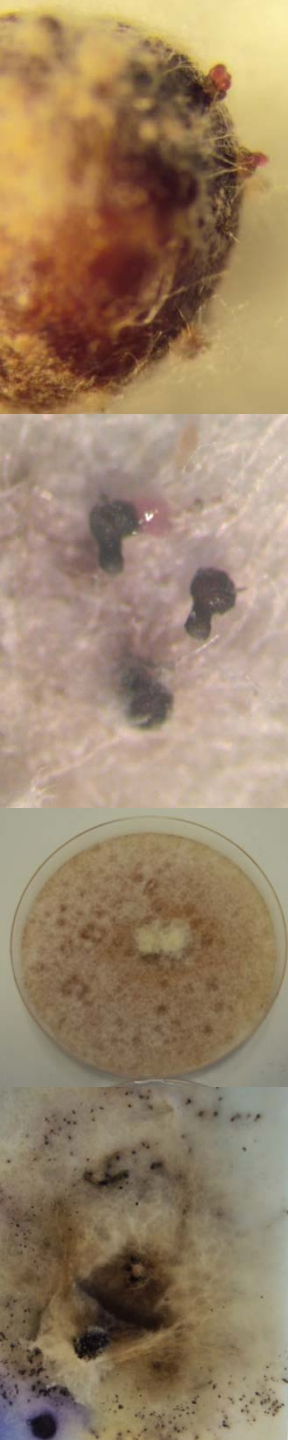


- Luminex

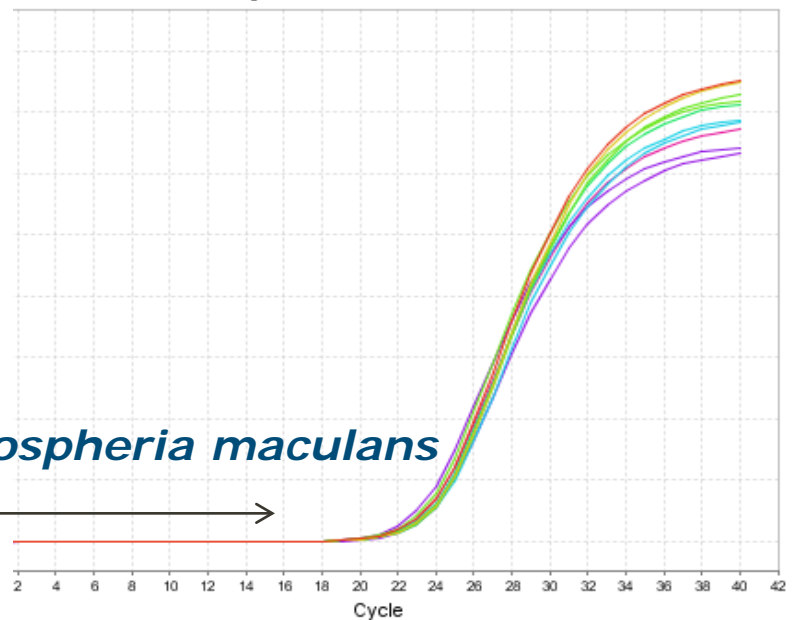


- Next Generation Sequencing





Amplification Plot



Leptosphaeria maculans

What is in the pipeline?



What's in the pipeline?

- Next Generation Sequencing (NGS) platforms
- NGS workflow
- NGS data analysis
- Examples
- Conclusions
- Outlook

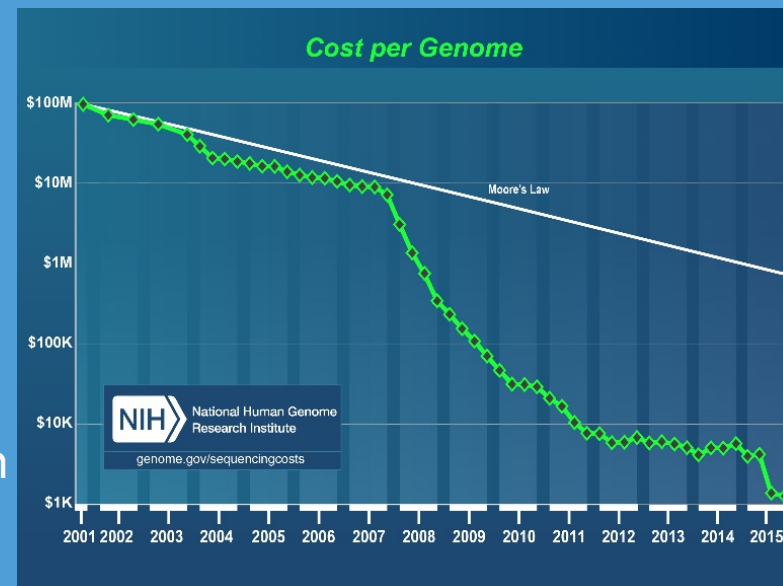
Genomics is essential for biological research

Applications

- Barcoding, MLA, diagnostic markers, phylogeny
- Generation of markers for population studies
- Genome -content, -expression, -evolution
- Understanding pathogenesis/communities
- Diagnostics and detection

Genomics is booming

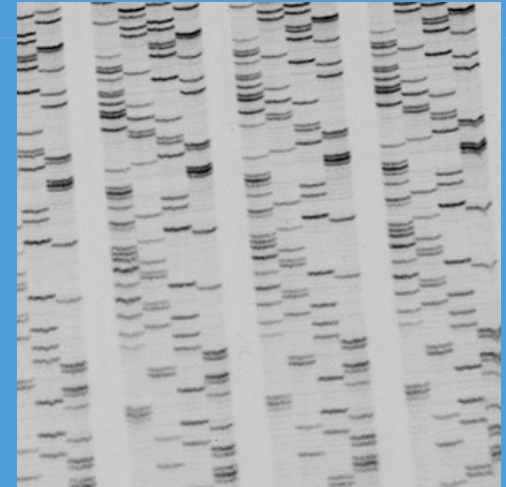
- Affordable
 - More data
 - More tools
- Genomics will guide biological research



Generations of sequencing

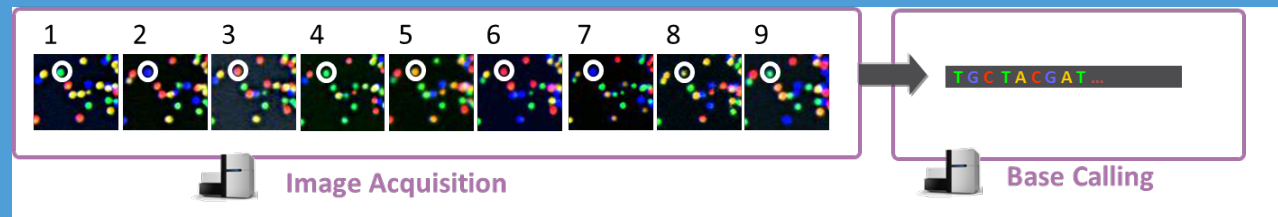
■ 1st generation Sanger sequencing

- >900 bp
- 96 lanes



■ 2nd generation Illumina (Roche 454)

- 125-300 bp
- 300 Gb/run

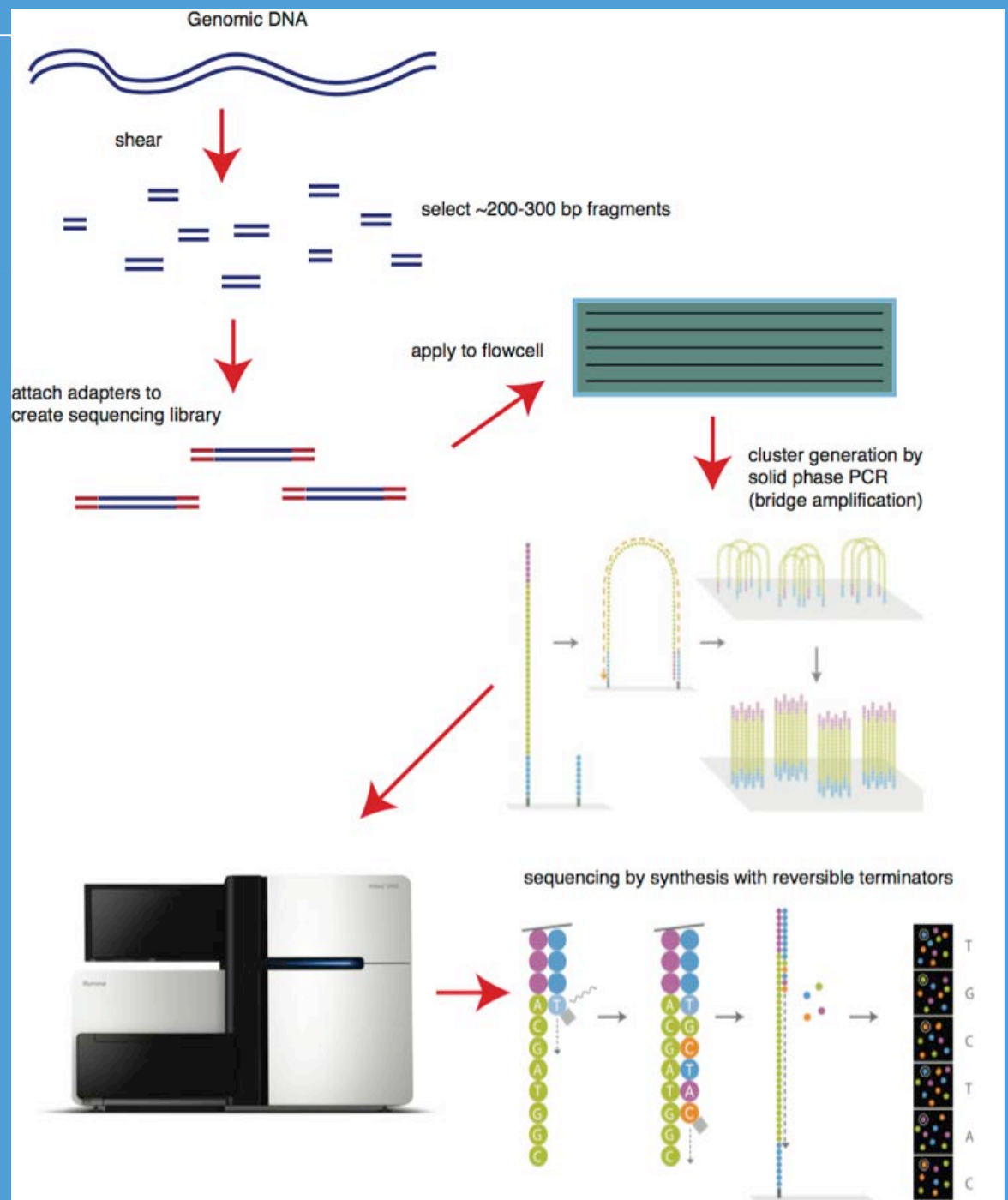


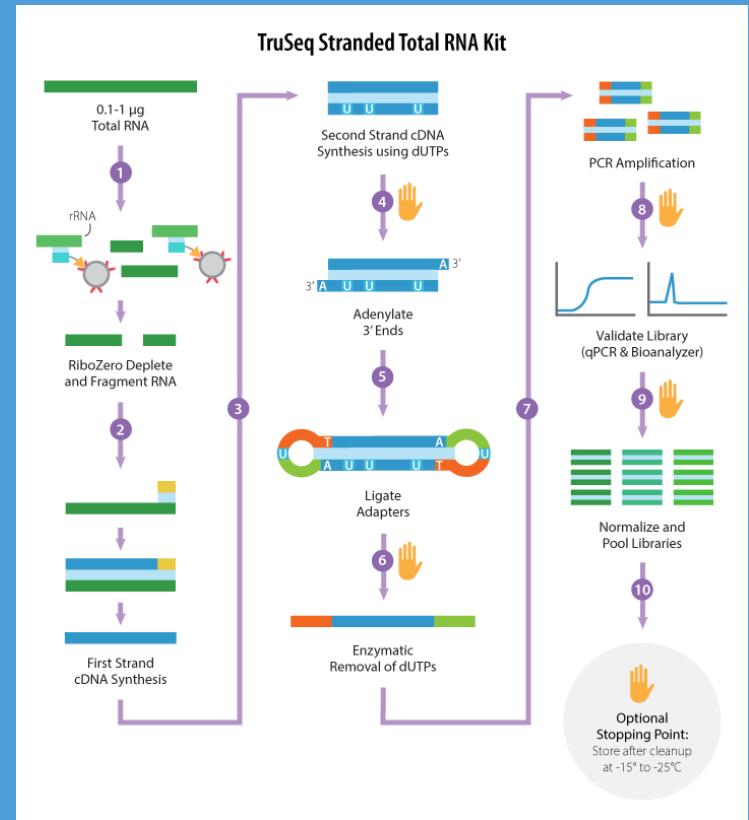
■ 3rd generation sequencing

- Single molecule sequencing
- Long reads >80 Kb



Illumina sequencing





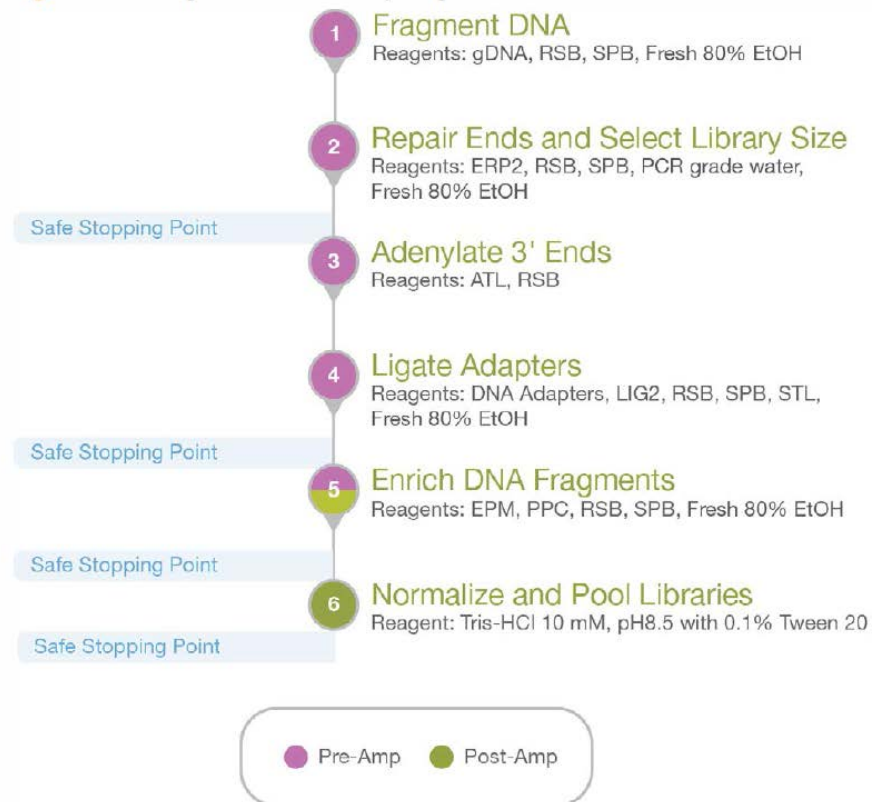
<https://www.bing.com/images/search?view=detailV2&ccid=yfgJOv0o&id=7D2226AA316C15844A95419CD54A1BDE5A0CA956&thid=OIP.yfgJOv0o8e3WXD6kV1yS2AELEs&q=ngs+sequencing&simid=608027187575850251&selectedIndex=22&ajaxhist=0>

TruSeq® Nano DNA Library Prep

Low sample (LS) protocol

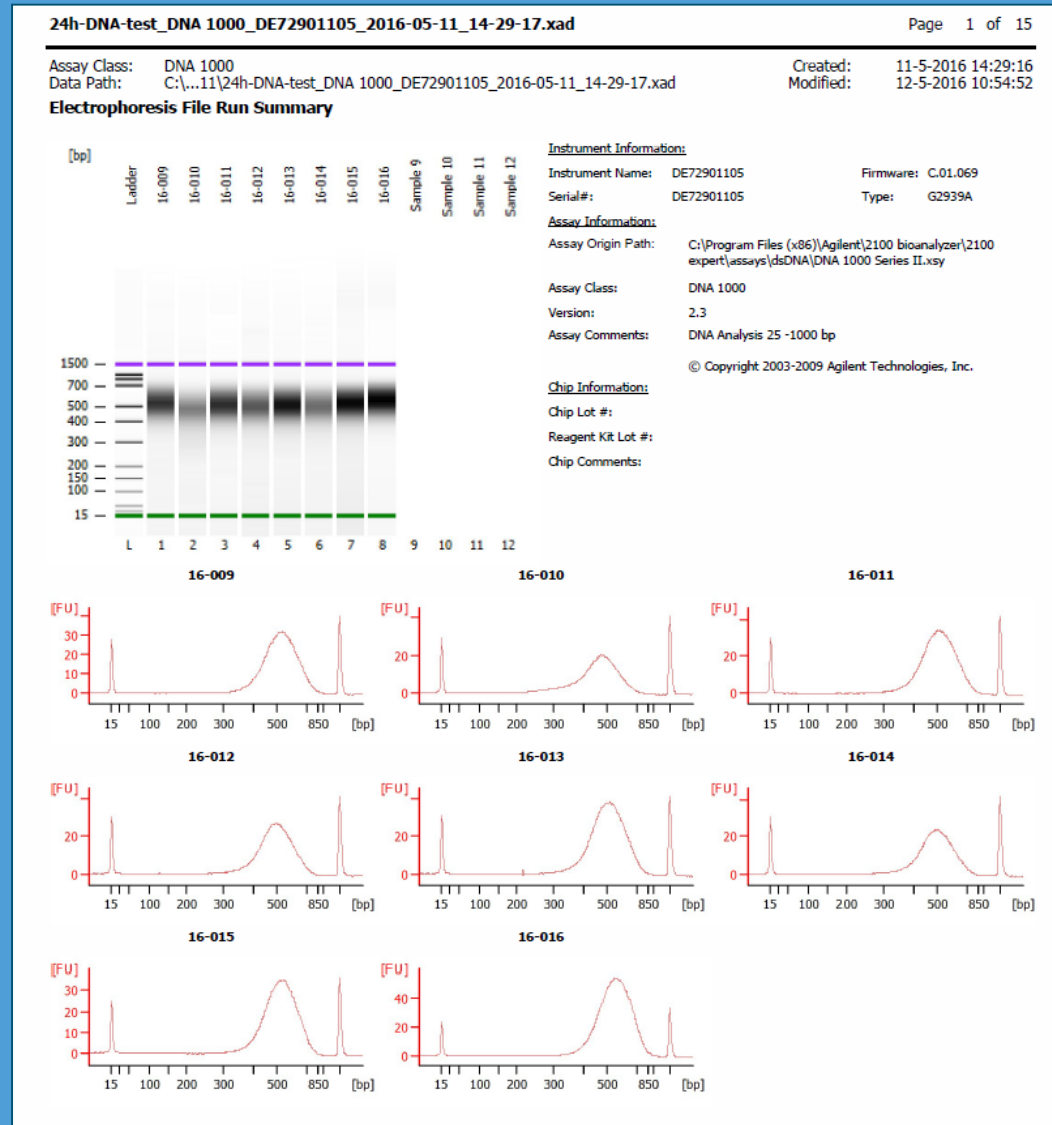
Library Prep Workflow

Figure 1 TruSeq Nano DNA Library Prep Workflow



Fragmentation of DNA

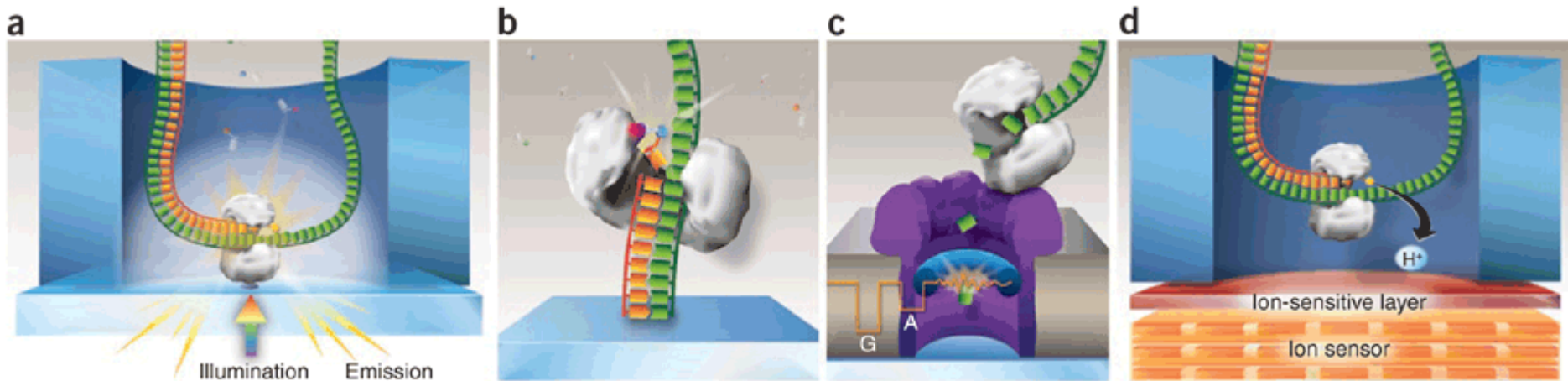
- Sonification: Covaris E210 Sonicator
- Expected size : 100-1000 nts
- Purification of required size
- DNA fragment + adapter seq (2*63 nts)



MiSeq equipment



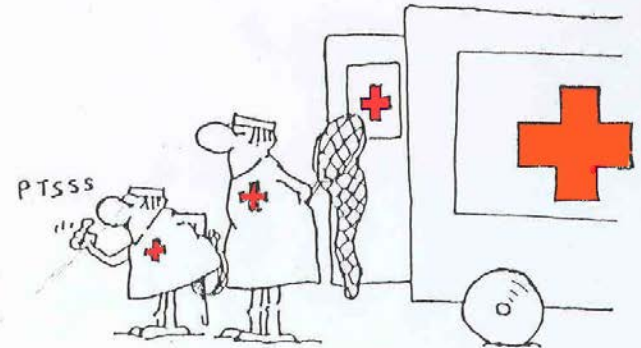
3rd generation: Single Molecule Sequencing strategies



- (a) Pacific Biosciences SMRT (single-molecule real-time) DNA sequencing method.
- (b) Life Technologies FRET sequencing platform uses base fluorescent labeling technology, a DNA polymerase modified with a quantum dot and DNA template molecules immobilized onto a solid surface.
- (c) The Oxford nanopore sequencing platform uses an exonuclease coupled to a modified α -hemolysin nanopore (purple, pictured in cross section) positioned within a lipid bilayer. As sequentially cleaved bases are directed through the nanopore, they are transiently bound by a cyclodextrin moiety (blue), disturbing current through the nanopore in a manner characteristic for each base.
- (d) The Ion Torrent sequencing platform uses a semiconductor-based high-density array of microwell reaction chambers positioned above an ion-sensitive layer and an ion sensor.

a DNA sequence in itself is useless

CTATTGAAAGCTTAGCAGCAAAAGCTGCAAGCAATATGGGGGTAAGTATCAAGGAAACCGTCAAAAGTGGCCGTTA
TGTATCCACTCCACACCCACATCTAGGCCGGTGAGCCCTGAAAGGTGTCTCGTATTGGACACTTAGCCAGGCTC
TGATCGAAGGAGTTTAGCCCGAATTTAGCTTGGCCCTTCGGTTAATAGTAATCTCTTATAAGCAAGCCAGAGG
CGTGGCACGGCTGAGATCCGATGTGATCCAATTCATGCTTTTAAAGGTTGTTACTCCTTAGTAGAGTGGGAG
GGCAACCAACGAACACTCGGAACAGTTCATCGTTTTGTCCAGGATAGCCCTAGAGTAGTCAGCGATAC
CTTTACGCTTGAGGGGTGTATAAGGCTACACTTGGCGATGTCTCTCTCCTTTTTATCCAAACACACTTCGTTG
AACAGGCCCTCACCGATGCTACCGGATGGTGAGTTTTTCCAAAGCGACTTCGCAACACTGGCGGCGTGTTA
TGGGTCATTTAGTGTGATGATGCTTCCACGGCATGTTCAAGGTGCTGGCTTCACCGATGAGTGTCCGGCGTAC
GGTCTACACCGATACCAAGTTACATTTGATGGCAGGTTCCACGGAGCCGCAACAGAGAGGTGACCGAAGCA
CAACCGCATGTCTTTGTCCGACACACCTACGTTTCAGCGTGCCTTTTAAACAATAGCCTCGCGTTGCCGTG
CGTTGCCGGGGAAAGGAACTGCGCGGCTCTTCCAACTTATACCTTTGTTCCGGCCAGAGAGGATTTGAG
TAGTAGTCTCCCATATGCGTACCGAAGGCCAATATTTAAATAGTTCCGCTCAGTAGGTCGTTGAT
CTTATGACTTCTCCGCGTCTGGCGATCATATAGGTTTCCGTAACCGCATAGACATGCGCTTT
AGAGGGGTCAATACGGCACCGTAAGGGCTGCGTGACGTTCCCCGTTCCGCTAATG
TCATCGGGGAAAGACGTCGACGTTGGTGGATCTCACCTCTACCACCTAG
AGGCTCTTGAATTCGGGGGCGCTAATAGCT
GGCCGCTGGTCACCATTTA
TGTACCGATGACC
ATACGAG
TGA



© Mordillo, Herz, Oehlmann 1985

Why Next generation sequencing makes a difference....

- ❖ More data
- ❖ More complex
- ❖ Unbiased

- Data analysis can be iterative, data production is done at once
- Large body of data allows for testing self-consistency

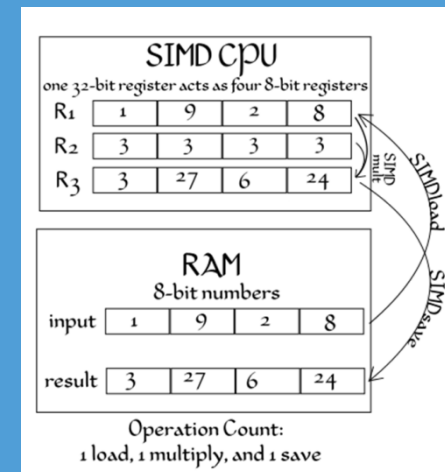
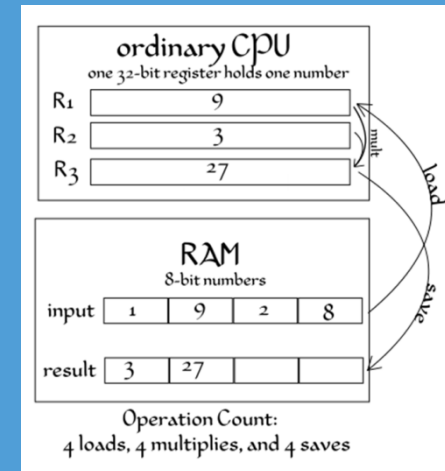
Software to analyse NGS data

- CLC bio's software
- Geneious
- Galaxy pipelines
- DNASTar
- Homebrew software

Analysis of millions of reads is a challenge

- SIMD technology

Single instruction, multiple data
multiple processing elements that
perform the same operation on
multiple data points
simultaneously.

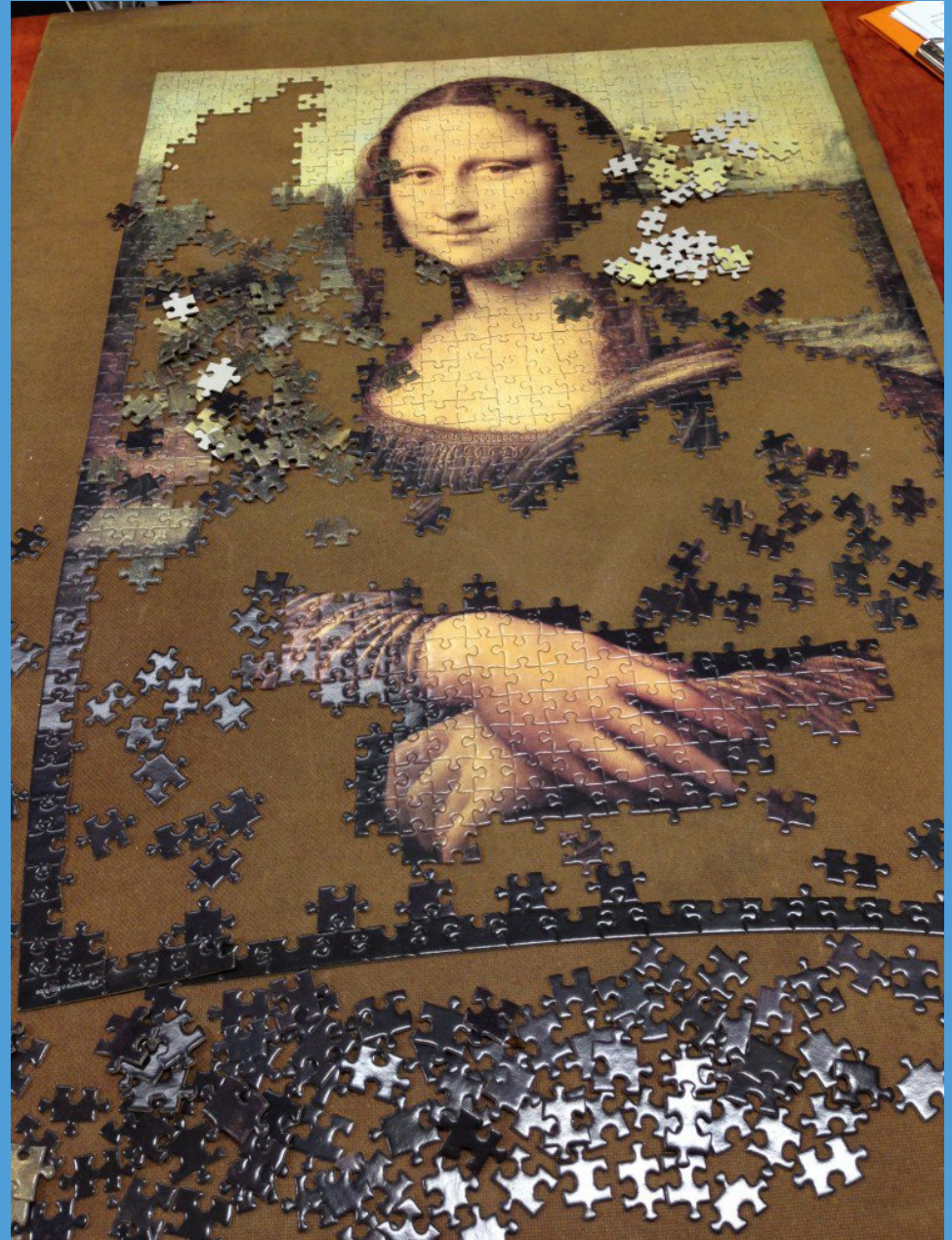


Introduction to terms

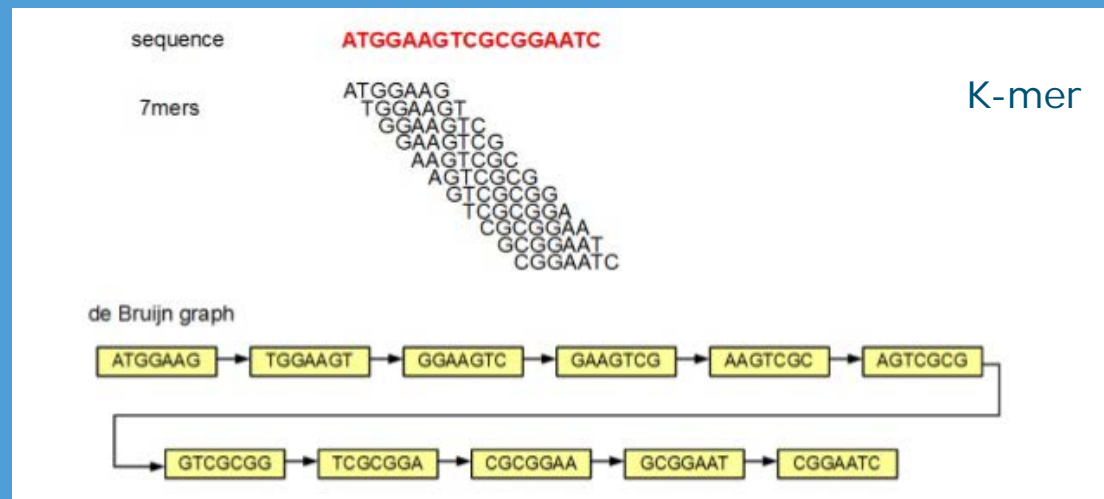
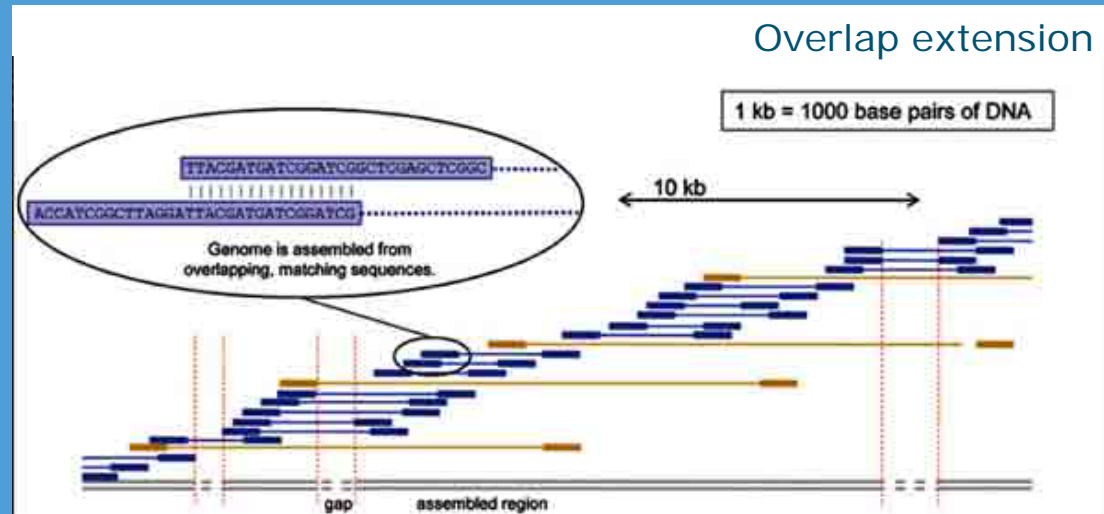
- De novo assembly
- Read mapping
 - Normalization: Reads per Kilobase per million (RPKM)
 - Genome coverage: depth versus width
 - eProbes: selective regions/random mapping
 - Best match of mapped reads (NCBI)

Two NGS analysis strategies

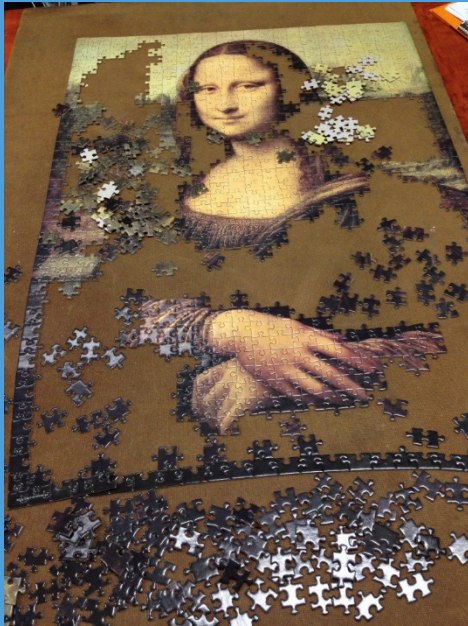
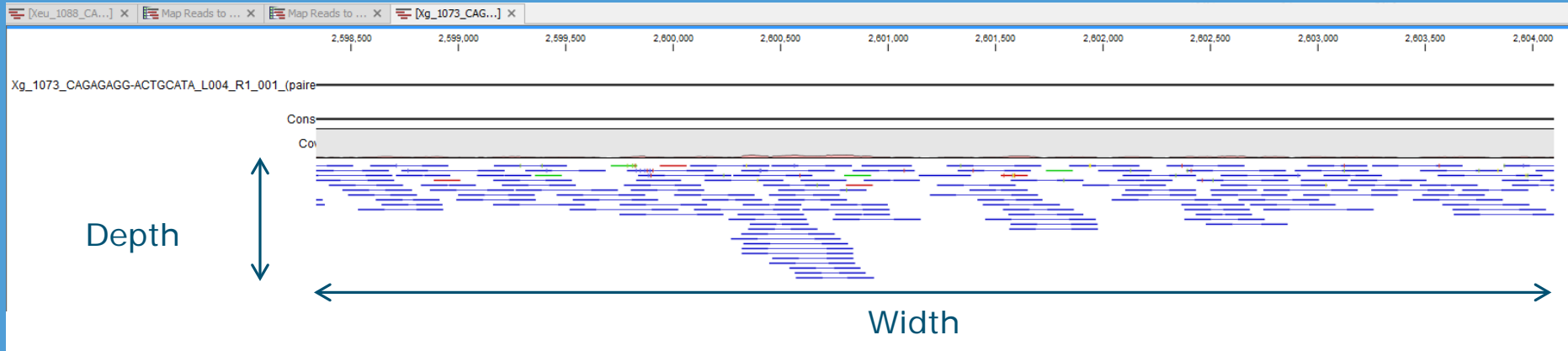
- De novo assembly
- Read mapping



From short reads to a genome by de novo assembly



Read mapping or reference based assembly



- Normalization: Reads Per Kilobase per Million (RPKM)
- Genome coverage: Width versus Depth
- Reads in pairs
- Homology

eProbes (selective regions/random mapping)

Situation	Random	E-probe
1	1	0
2	1	1
3	1	0
4	0	0

Ref A CTAGGAGGGATA**G**TAGATGATGAT
 CTAGGAGGGATA**G**TAGATGATGAT
 Ref B CTAGGAGGGATA**G**TAGATGATGAT

Ref A CTAGGAGGGATA**G**TAGATGATGAT
 CTAGGAGGGATA**G**TAGATGATGAT
 Ref B CTAGGAGGGATA**T**TAGATGATGAT

Ref A CTAGGAGGGATA**T**TAGATGATGAT
 CTAGGAGGGATA**G**TAGATGATGAT
 Ref B CTAGGAGGGATA**T**TAGATGATGAT

Ref A CTAG**A**AGGGATA**T**TAGAC**G**ATG**T**
 CTAGGAGGGATA**G**TAGATGATGAT
 Ref B **AC**AGGAGGGATA**T**TAGATGAC**G**AT

Read

CTAGGAGGGATA**G**TAGATGATGAT

NGS for detection and diagnostics strategy

Instrumentation:

Normalization, RPKM

eProbes (selective regions/random mapping)

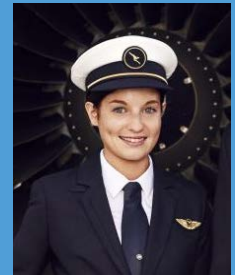


Pilot: read mapping on reference sequences

NGS TaqMan

NGS barcode/MLSA

NGS Whole genome



Co-pilots:

Genome coverage → width versus depth

Best match of mapped reads (NCBI)



Explorer:

Low homology

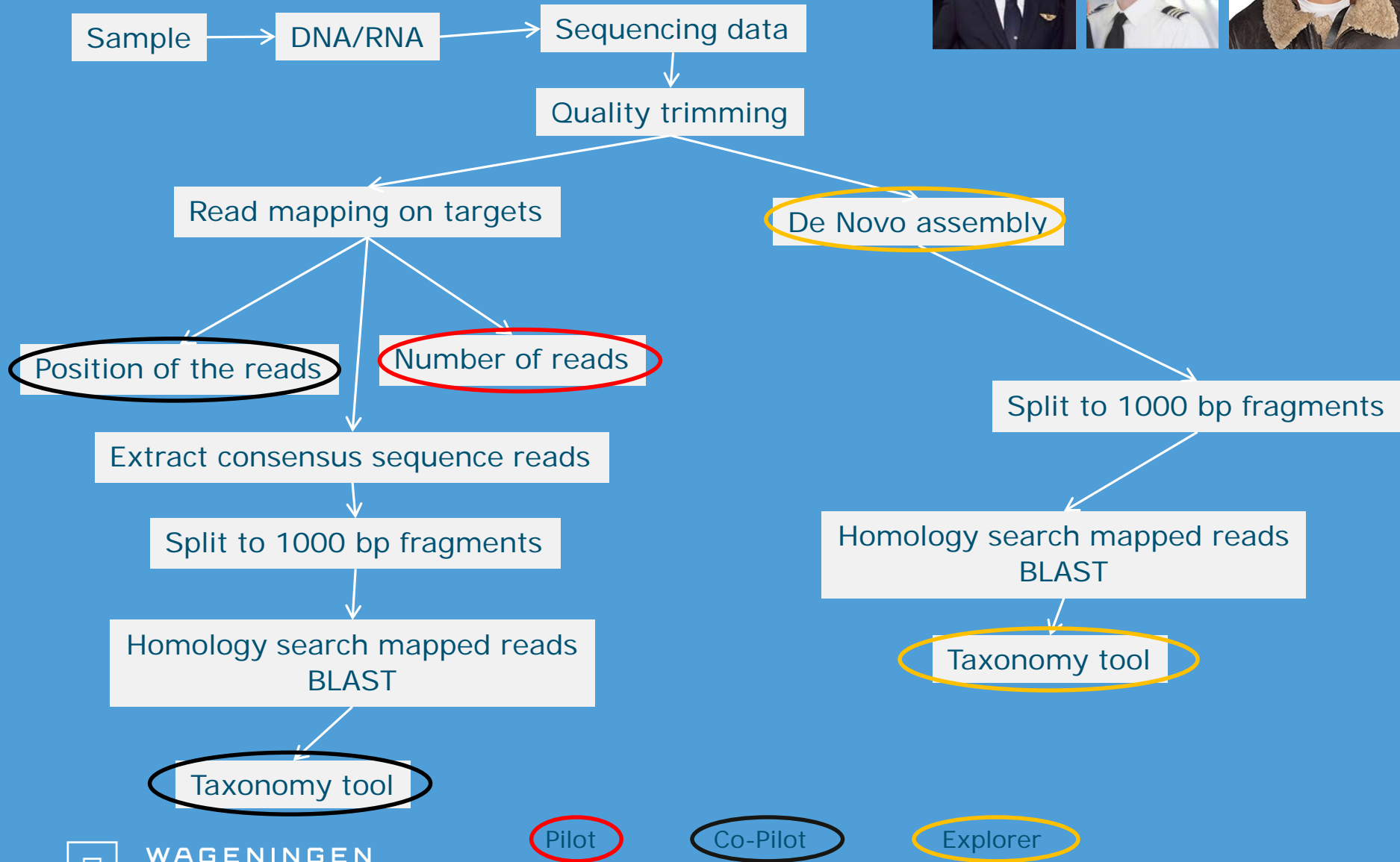
De novo assembly and characterization of derived contigs

Blastn

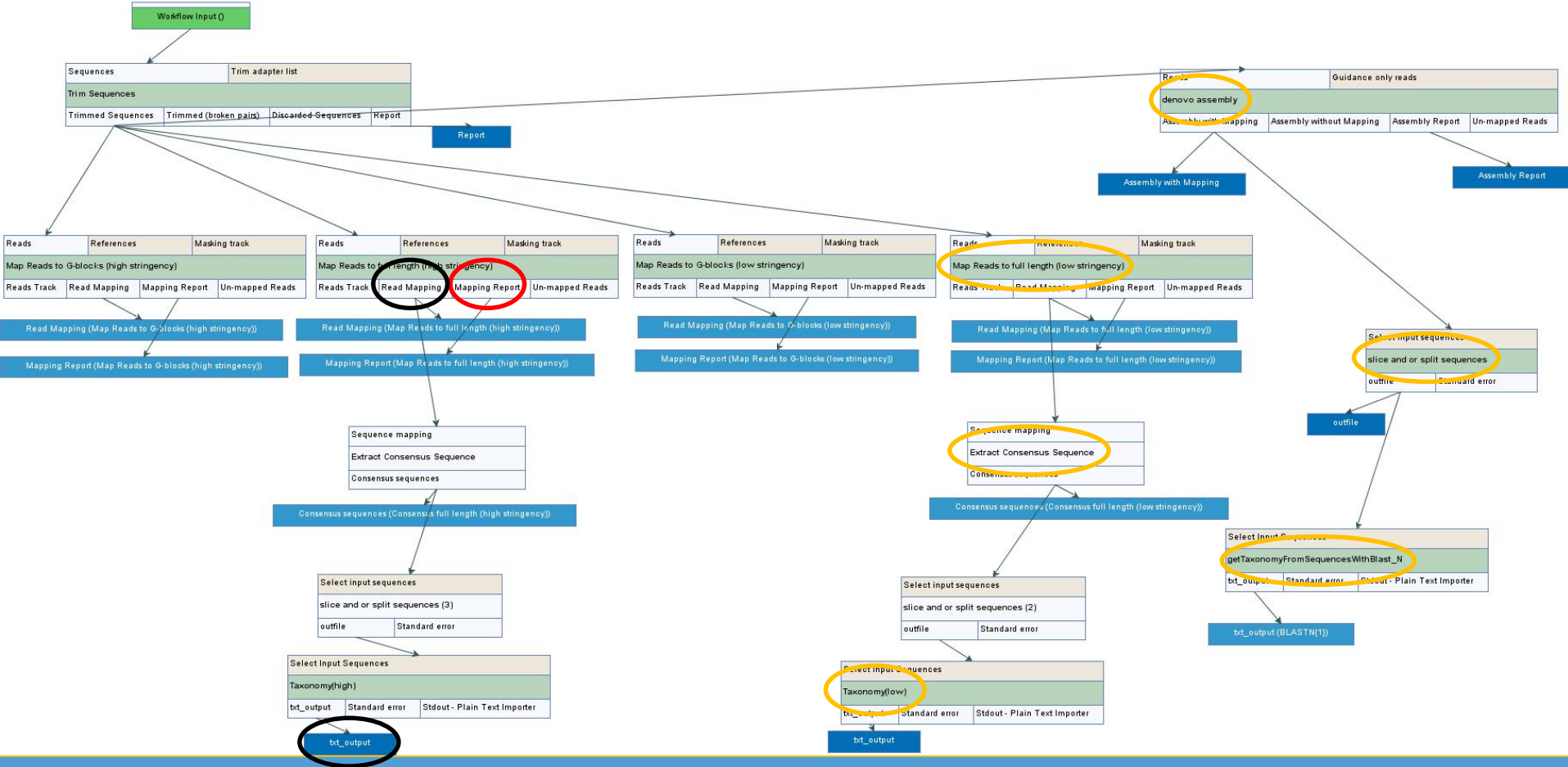
Blastx



NGS data analysis pipeline



NGS data analysis pipeline



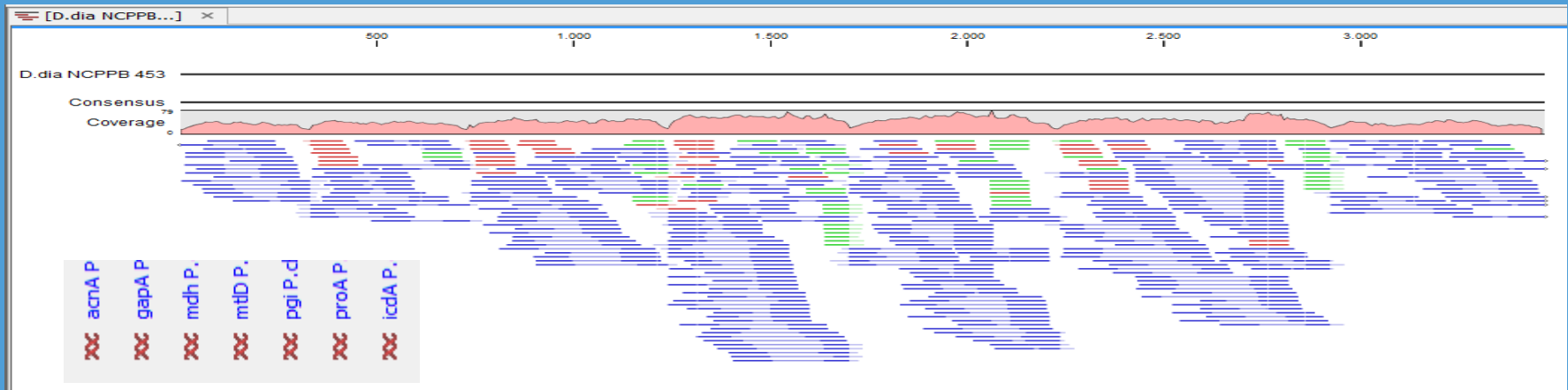
Pilot

Co-Pilot

Explorer



Species identification/confirmation based on MLSA



Rows: 13 Filter ▾

Name	Consensus length	Total read count	Reads in pairs	Average coverage	Reference sequence	Reference length
P.cbe IPO_0261	0	0	0	0,00	P.cbe IPO_0261	3473
P.cod IPO_3658	0	0	0	0,00	P.cod IPO_3658	3476
P.was Wpp163	0	0	0	0,00	P.was Wpp163	3469
P.atr SCRI1043	0	0	0	0,00	P.atr SCRI1043	3466
P.cca PC1	0	0	0	0,00	P.cca PC1	3466
D.ddi NCPPB 2976	0	0	0	0,00	D.ddi NCPPB 2976	3472
D.dia NCPPB 453 mapping	3466	1654	902	43,32	D.dia NCPPB 453	3466
D.sol IPO_2222	0	0	0	0,00	D.sol IPO_2222	3459
D.dda NCPPB 898 mapping	83	3	2	0,04	D.dda NCPPB 898	3474
D.chr NCPPB 402	0	0	0	0,00	D.chr NCPPB 402	3462
D.par NCPPB 2511	0	0	0	0,00	D.par NCPPB 2511	2927
D.zea NCPPB 2538 mapping	30	1	0	9,60E-3	D.zea NCPPB 2538	3124
P.cbr IPO_3540	0	0	0	0,00	P.cbr IPO_3540	2241

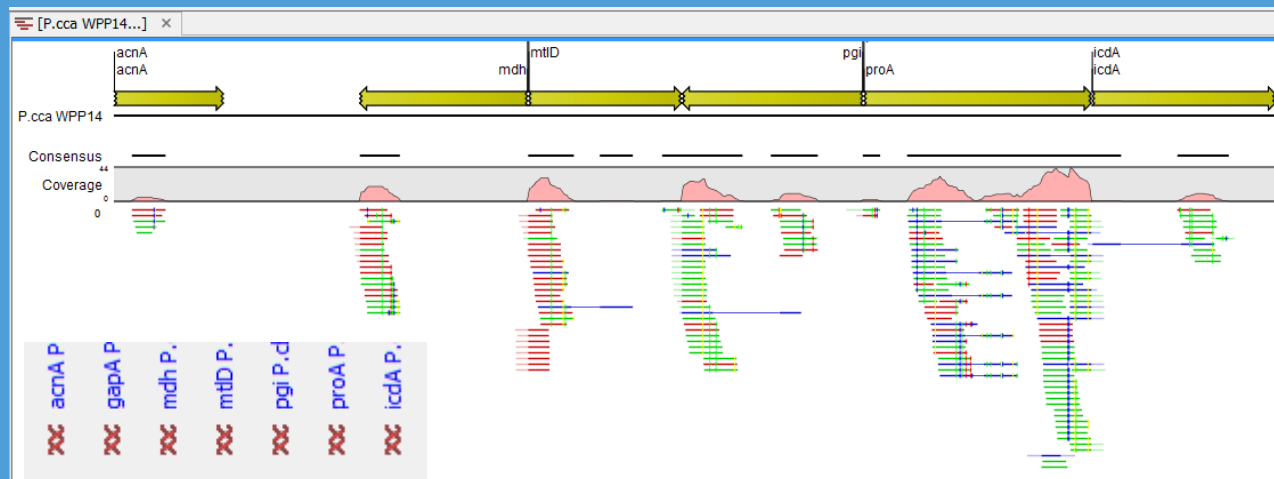
Species identification/confirmation based on MLSA

Rows: 14

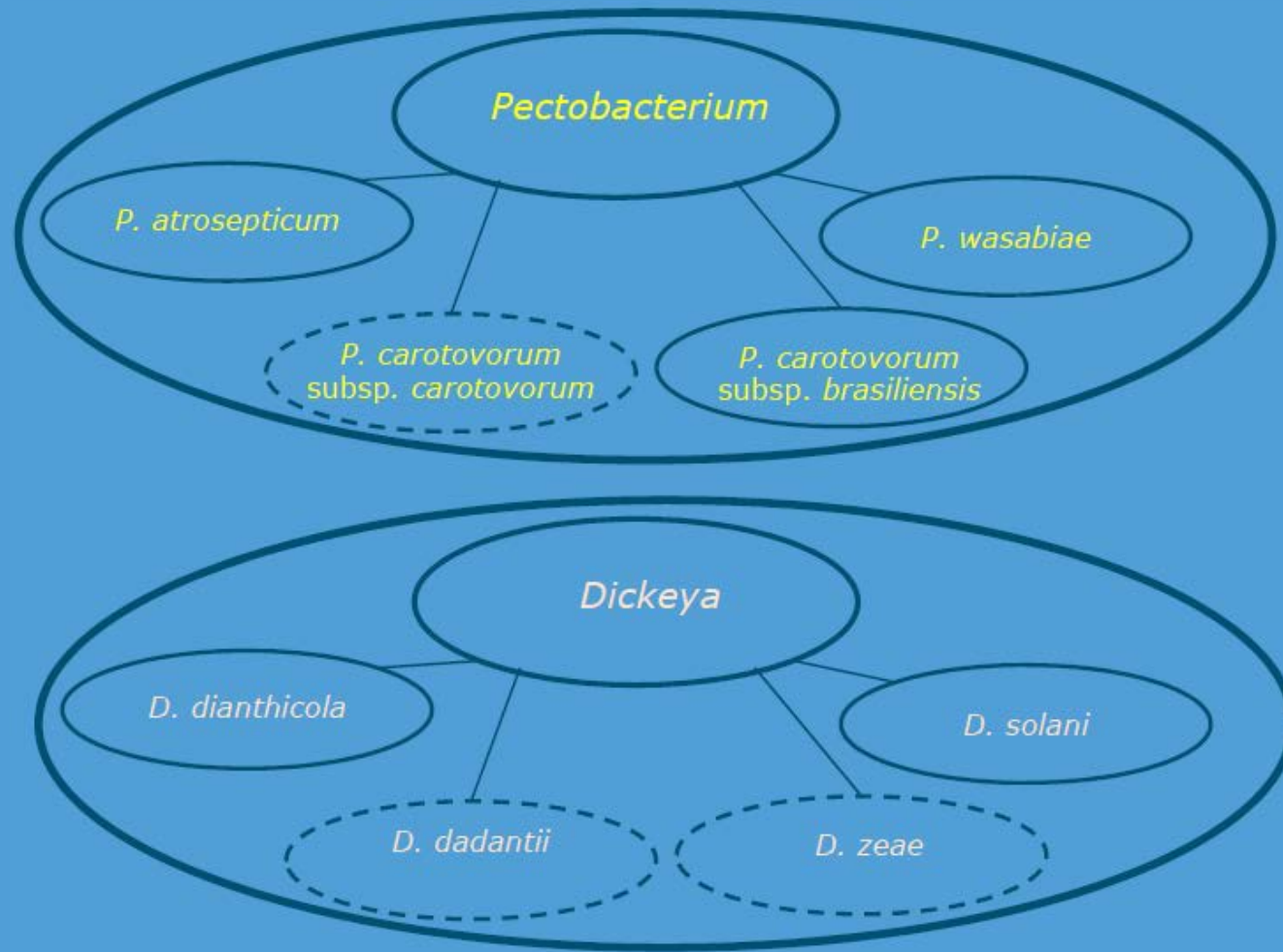
Name	Consensus l...	Total read ...	Average coverage	Reference sequence	Reference l...
P.cbe IPO_0261 mapping	374	51	1,30	P.cbe IPO_0261	3473
P.cod IPO_3658 mapping	2862	543	14,70	P.cod IPO_3658	3476
P.was Wpp163 mapping	228	9	0,25	P.was Wpp163	3469
P.atr SCRI1043 mapping	79	1	0,02	P.atr SCRI1043	3466
P.cca PC1 mapping	344	40	0,96	P.cca PC1	3466
D.ddi NCPPB 2976	0	0	0,00	D.ddi NCPPB 2976	3472
D.dia NCPPB 453	0	0	0,00	D.dia NCPPB 453	3466
D.sol IPO_2222	0	0	0,00	D.sol IPO_2222	3459
D.dda NCPPB 898	0	0	0,00	D.dda NCPPB 898	3474
D.chr NCPPB 402 mapping	25	1	7,22E-3	D.chr NCPPB 402	3462
D.par NCPPB 2511 mapping	83	3	0,05	D.par NCPPB 2511	2927
D.zea NCPPB 2538	0	0	0,00	D.zea NCPPB 2538	3124
P.cbr IPO_3540 mapping	1140	131	3,53	P.cbr IPO_3540	3469
P.cca WPP14 mapping	1679	230	5,77	P.cca WPP14	3486

Originates from:

- ✘ acnA P.cbe IPO_0261 (f
- ✘ gapA P.cbe IPO_0261 (f
- ✘ mdh P.cbe IPO_0261 (hi
- ✘ mtD P.cbe IPO_0261 (h
- ✘ pgi P.cbe IPO_0261 (his
- ✘ proA P.cbe IPO_0261 (f
- ✘ icdA P.cbe IPO_0261 (hi

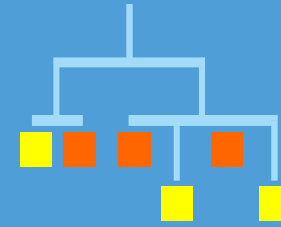


Enterobacteriaceae species involved in potato soft rot

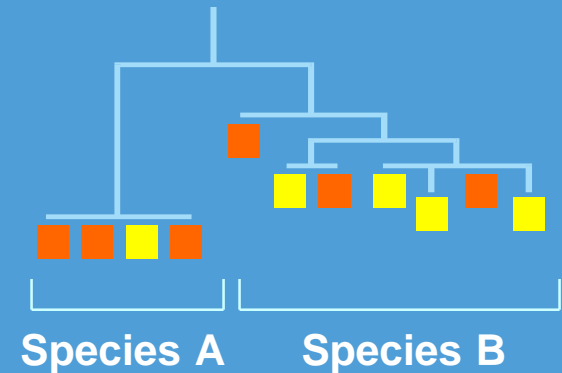


Evolutionary Fate of Neutral Polymorphism

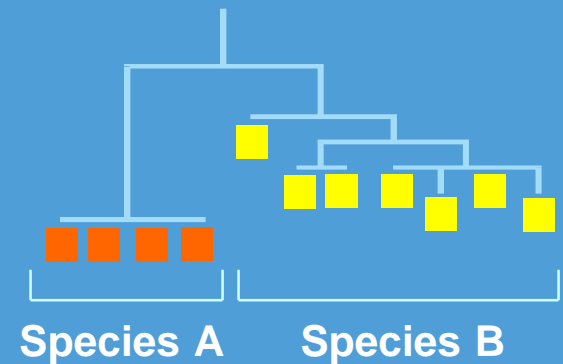
Polymorphism within a species



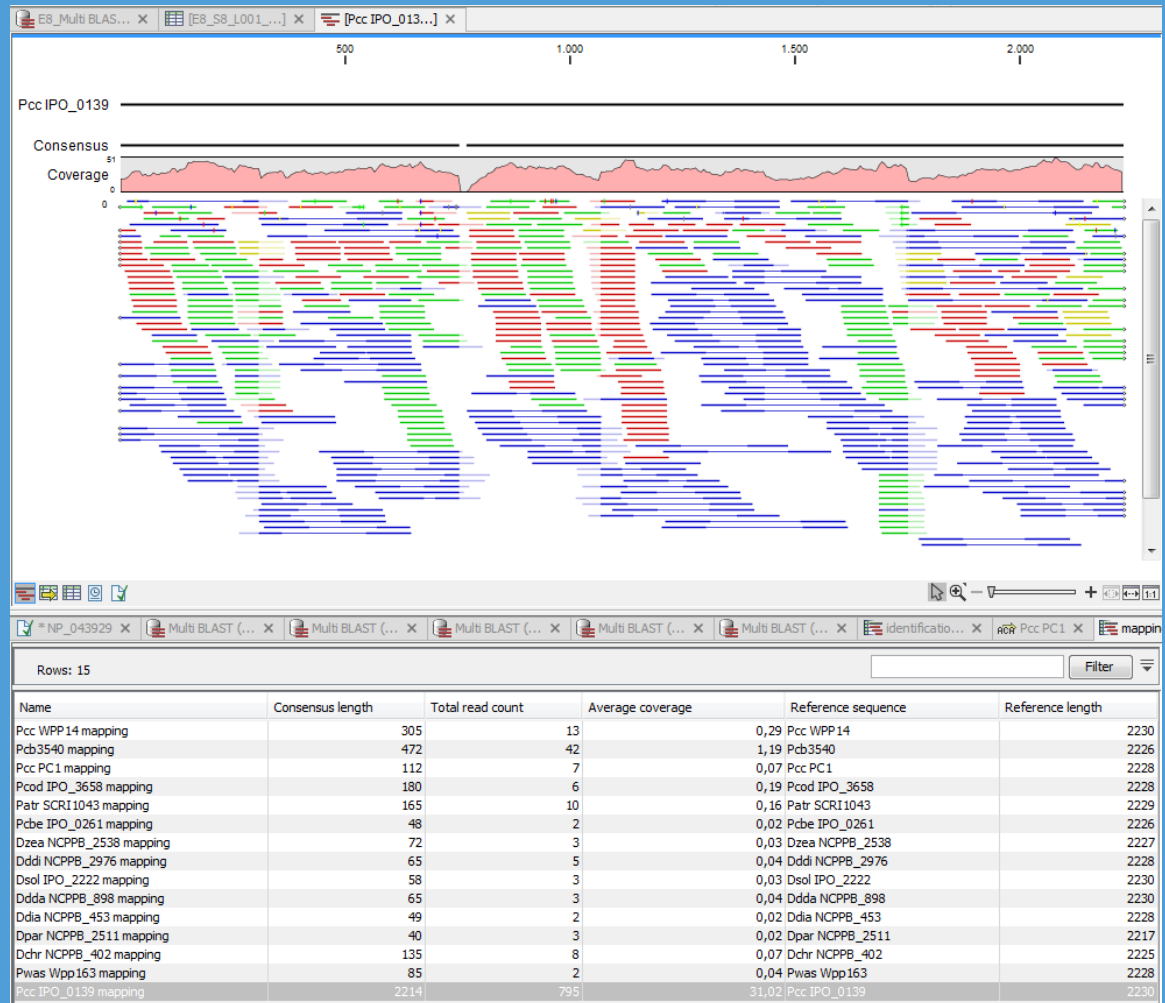
Shared polymorphism immediately following speciation



Shared polymorphism lost due to drift in the absence of significant gene flow



New barcoding genes help correct in species identification

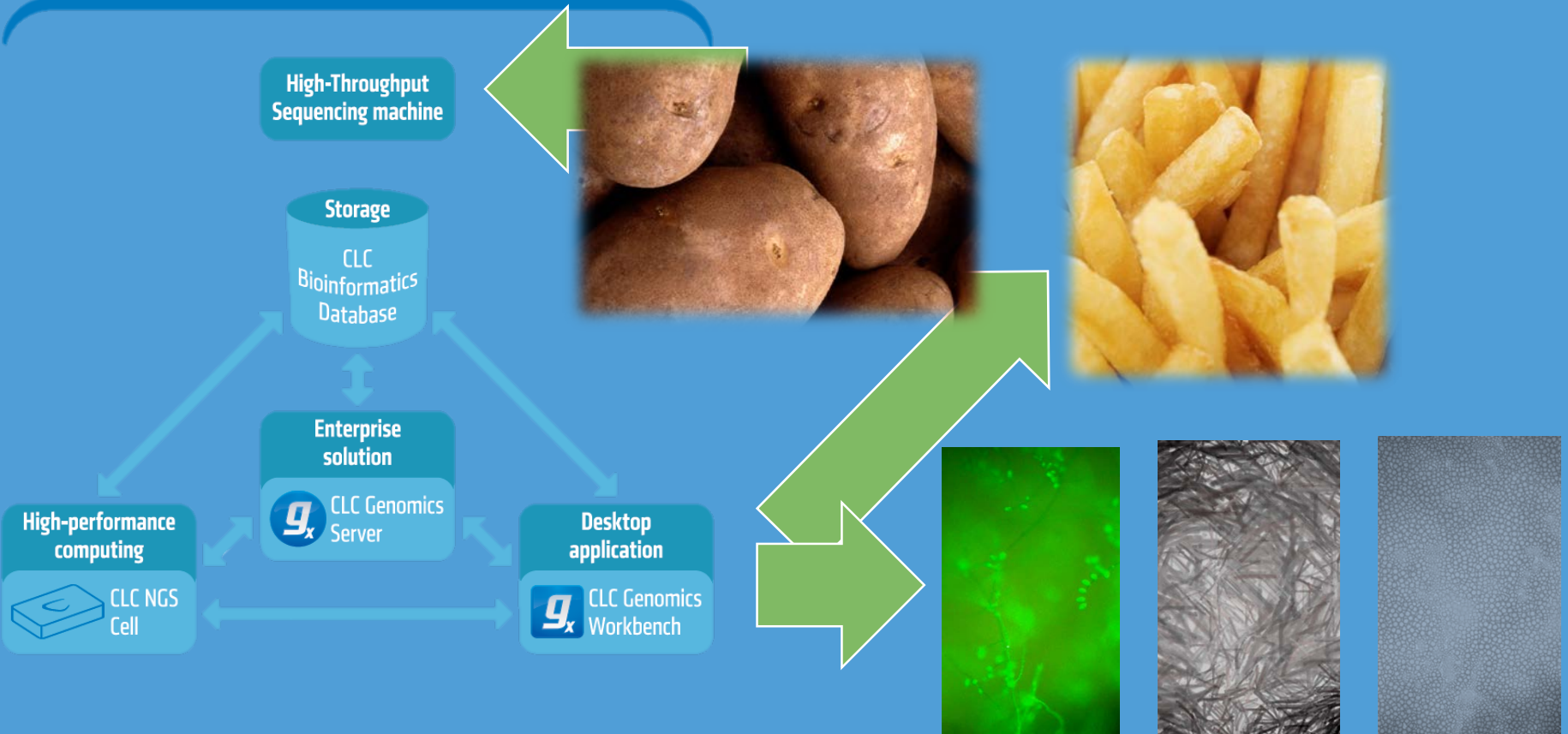


Genomic based diagnostics in complex matrixes

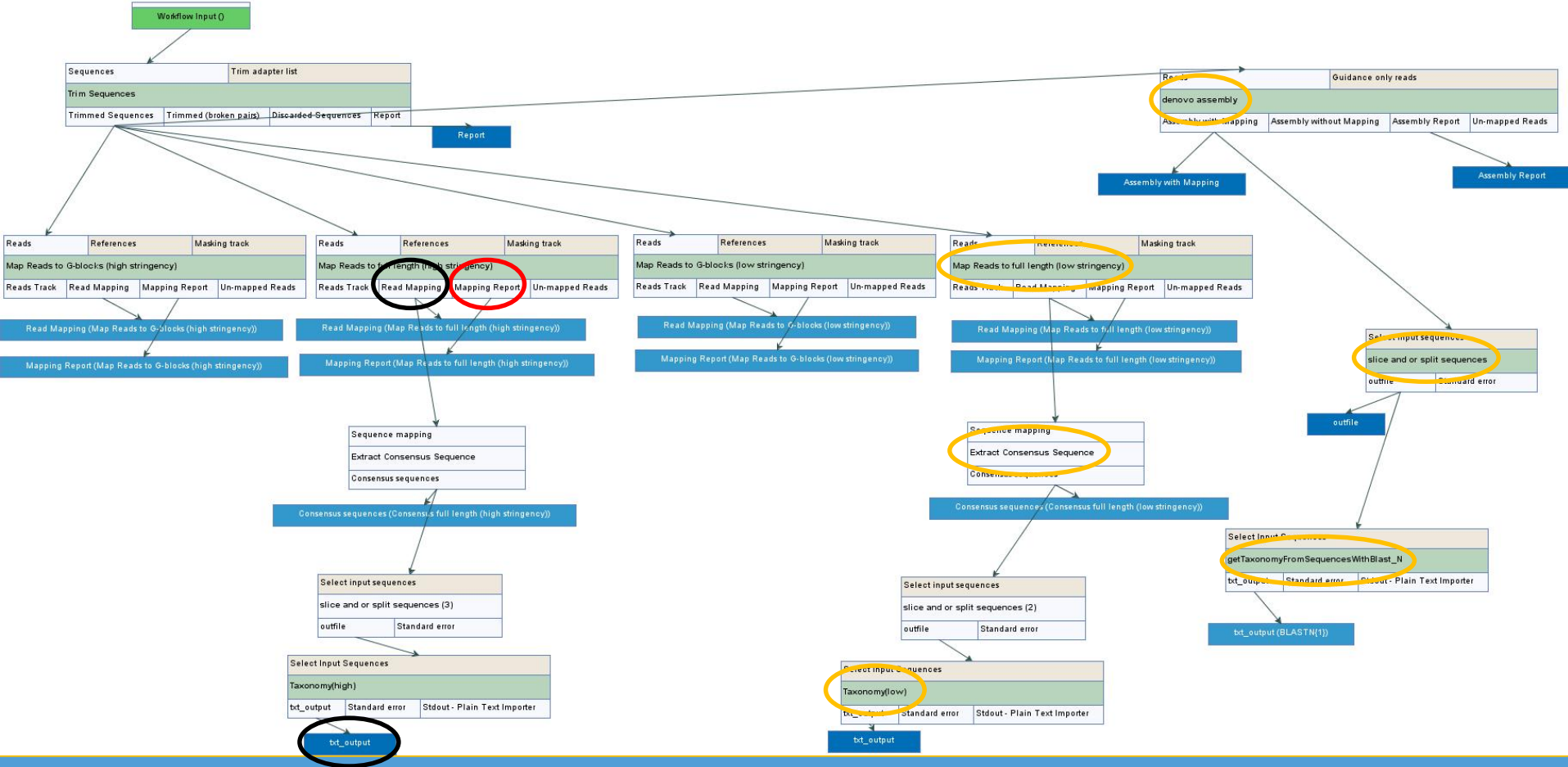
Bio-informatics

Whole Genome Analysis
Multiple Locus Analysis

DATA FLOW



NGS data analysis pipeline



Pilot

Co-Pilot

Explorer



WAGENINGEN
UNIVERSITY & RESEARCH



Example on tomato seeds

Optimizing settings using clean tomato seeds

In silico experimentation

Artificial contaminated seeds

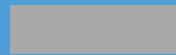
Naturally contaminated seeds

Experimental design

10.000 healthy tomato seeds (74-628)

Soaking 16 hr o/n at 4°C in Seed Extraction Buffer (150 ml)

Stomacher step on seed



In 50 ml: Addition of Bacteria



Slow Centrifugation



sup

45 ml Fast Centrifugation



pellet

MoBio kit King Fisher



Spike:

X. perforans
X. vesicatoria
X. euvesicatoria
X. gardneri
Pst_919
C.m.m_3356

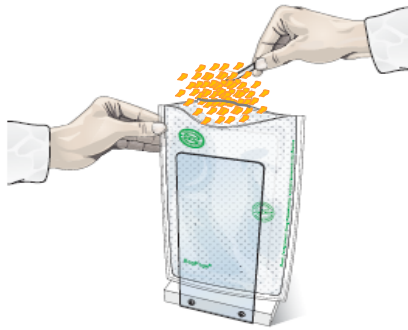
Internal controls:

C.m.t
IPO_3677

Bacterial pellet in 1000 µl

- 4x 100 µl/sample (a t/m h)
- sample 1a-1d used for sequencing

Stomacher BagMixer® 400 CC, (Interscience)



BagFilter®
Collect



BagMixer®
Blend



BagTools®
Pipet



In silico NGS test performance

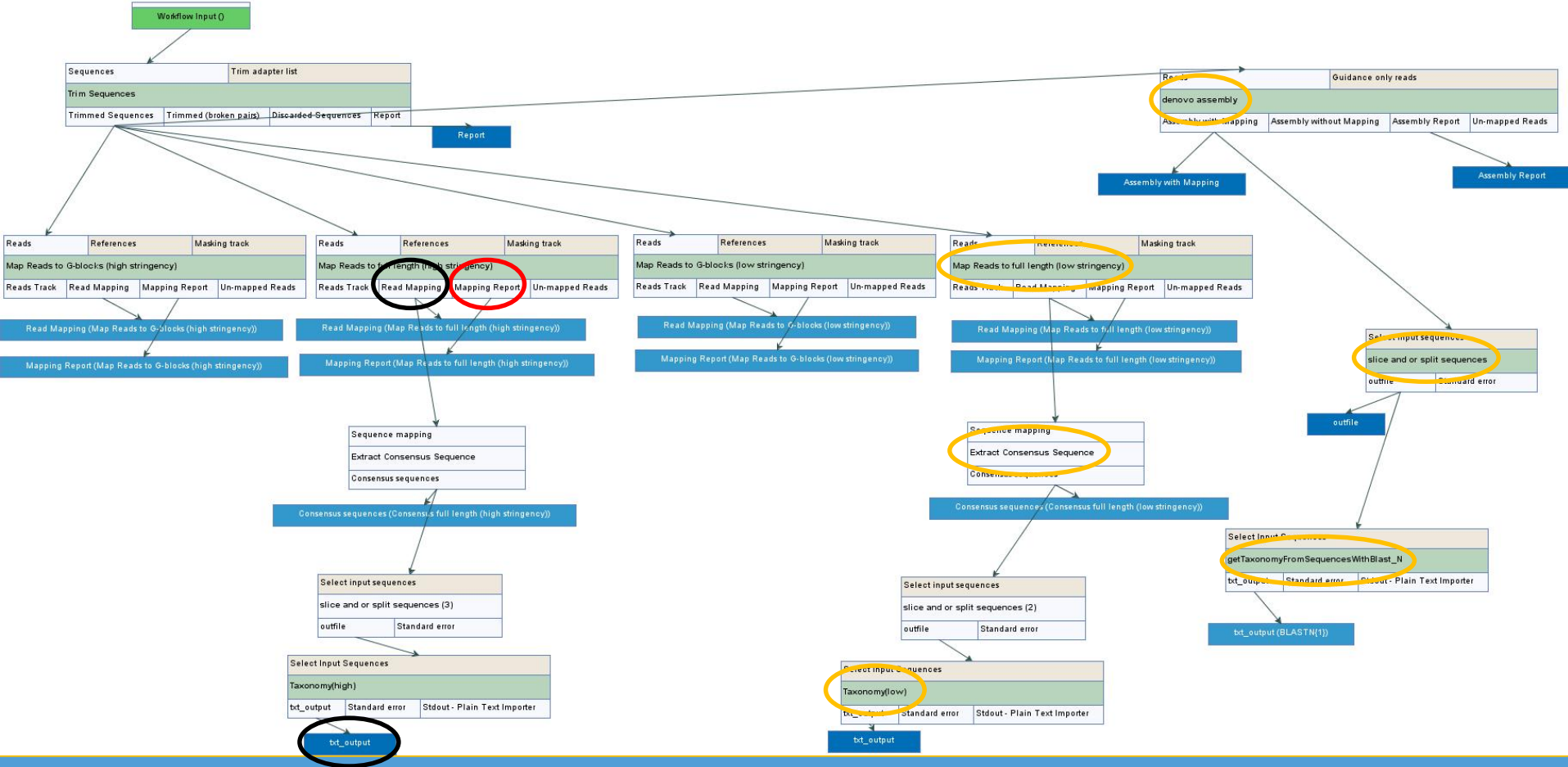
CMM in silico spike	Mapping in silico CMM spike		Mapping RZ-minus + in silico CMM spike	
	Total read count	Reads in pairs	Total read count	Reads in pairs
1	1	0	1	0
10	10	10	10	10
100	96	92	96	92
1000	958	906	958	906
10000	9602	9084	9602	9084

No interference:

Detection is independent on the number of total reads in sample



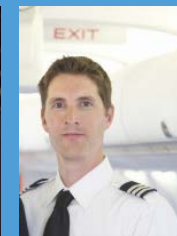
NGS data analysis pipeline



Pilot

Co-Pilot

Explorer



Benchmarking protocols and procedures for NGS in diagnostics

Name	Clean seeds		Spiked seeds	
	High	Low	High	Low
Tomato_nucleous	1.0	1.1	0.6	0.7
Tomato_chloroplast	80.0	81.6	33.9	34.7
Tomato_mitochondrion	12.6	13.0	5.4	5.6
Xeu_1088	0.0	0.0	17.0	17.3
Cmm_542	0.0	0.0	3.0	3.1
Pst_919	0.0	0.0	7.2	7.3
Xp_1067	0.0	0.0	13.4	13.5
Xg_1073	0.0	0.0	17.6	18.0
Xv_1059	0.0	0.0	13.0	13.2

RPKM= reads per Kb per million reads

```

Comments:5/8
Reads mapped: 12,209,945 of 15,382,658
Estimated paired distance range(s):
run0194_R2-minus_95_1003_R1_001.nophix (paired) trimmed (paired): 60 to 554 bp
Origines from:
100 Xeu_1088_CAGAGAGG-GTAAGCCT_L004_R1_001_(paired)_fus (history)
100 Cmm542_GCTACGCT-CTCTCTAT_L003_R1_001_(paired)_fus (history)
100 Pst_919_GCTACGCT-TAGATCGC_L004_R1_001_(paired)_fus (history)
100 Xp_1067_CAGAGAGG-AAGGAGTA_L004_R1_001_(paired)_trimmed_(paired)_without_duplicates_fus (history)
100 Xg_1073_CAGAGAGG-ACTGCATA_L004_R1_001_(paired)_fus (history)
100 Xv_1059_CAGAGAGG-GTAAGGAG_L004_R1_001_(paired)_trimmed_(paired)_without_duplicates_fus (history)
100 Tomato_SL2.50_nucleus_fus (history)
100 Tomato_mitochondrion_fus (history)
100 Tomato_chloroplast (history)

Trim Sequences (Tue Apr 12 14:28:43 CEST 2016)
Version: CLC Genomics Server 7.8.1
User: huawei28
Parameters:
Indelicious trim = Yes
Ambiguous limit = 2
Quality trim = Yes
Quality limit = 0.03
Create report = Yes
Save discarded sequences = No
Remove 5' terminal nucleotides = No
Maximum number of nucleotides in reads = 125
Minimum number of nucleotides in reads = 80
Discard short reads = Yes
Remove 3' terminal nucleotides = No
Discard long reads = Yes
Save broken pairs = No

Comments:5/8
Processed a total of 100,943,808 sequences.
81,694,043 sequences have been completely removed during trimming.
7,765,449,581 nucleotides have been trimmed, altogether.
Origines from:
100 run0194_R2-minus_95_1003_R1_001.nophix (paired) (history)
    
```

Name	Consen...	Total re...	Re... ▾	Average...
Tomato_SL2.50 nucleus_fus mapping	508390865	11972932	11282...	1.80
Tomato_chloroplast mapping	71402	186613	180444	148.75
Tomato_mitochondrion_fus mapping	135171	50399	47166	23.01
Xeu_1088_CAGAGAGG-GTAAGCCT_L004_R1_001_(p...	0	0	0	0.00
Cmm542_GCTACGCT-CTCTCTAT_L003_R1_001_(pair...	0	0	0	0.00
Pst_919_GCTACGCT-TAGATCGC_L004_R1_001_(pair...	0	0	0	0.00
Xp_1067_CAGAGAGG-AAGGAGTA_L004_R1_001_(pai...	0	0	0	0.00
Xg_1073_CAGAGAGG-ACTGCATA_L004_R1_001_(pai...	125	1	0	2.33E-5
Xv_1059_CAGAGAGG-GTAAGGAG_L004_R1_001_(pai...	0	0	0	0.00

1.1 Summary statistics

	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	9	-	94,993,274.56	854,939,471	-
Mapped reads	12,209,945	79.37%	124.15	1,515,804,595	79.42%
Not mapped reads	3,172,713	20.63%	123.78	392,723,963	20.58%
Reads in pairs	11,509,770	74.82%	318.70	1,428,541,148	74.85%
Broken paired reads	700,175	4.55%	124.63	87,263,447	4.57%
Total reads	15,382,658	100.00%	124.07	1,908,528,558	100.00%

Reproducibility NGS as diagnostic tool

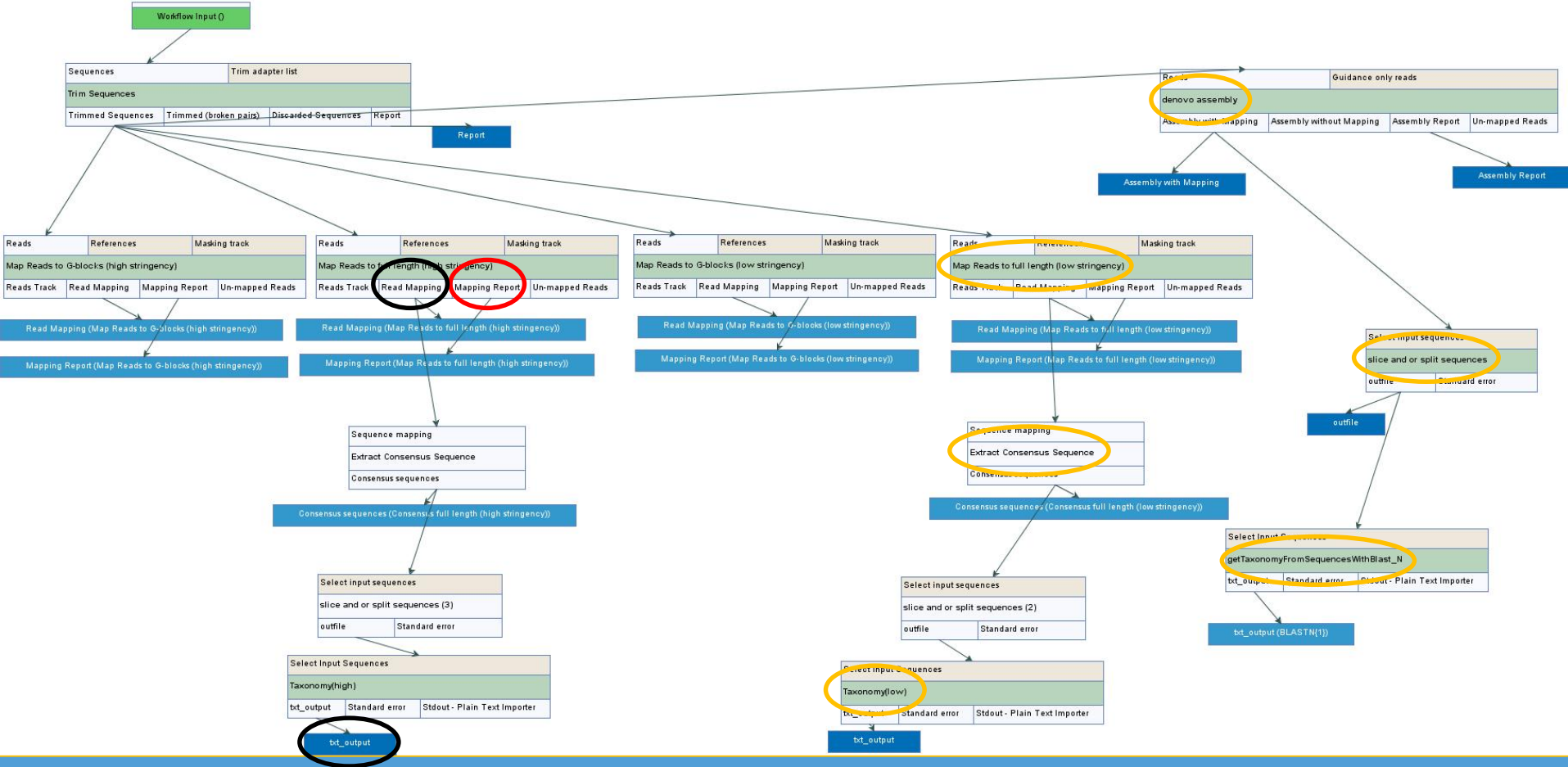
Name	1a			1b			1c			1d		
	Total read count	RPKM	Taqman CT	Total read count	RPKM	Taqman CT	Total read count	RPKM	Taqman CT	Total read count	RPKM	Taqman CT
Tomato	12970429	1.02	ND	13715490	1.00	ND	12711539	1.00	ND	31158587	1.01	ND
Total Xeu	128038	1.56	28.58	162025	1.82	28.57	147350	1.80	28.52	350705	1.76	28.47
Total Xv	115671	1.42	24.38	148234	1.68	24.37	128854	1.59	24.24	303549	1.53	24.32
Total Pst	1030221	10.70	23.19	1218444	11.67	23.09	1077143	11.18	23.20	2525668	10.77	23.21
Total Xg	4432	0.05	31.24	5684	0.06	31.20	5024	0.06	31.65	11642	0.06	31.58
Total Xp	781855	10.15	23.75	987044	11.82	23.79	864427	11.22	23.86	2068091	11.03	23.83
Total Cmm	40400	0.78	25.05	70144	1.25	25.31	113640	2.20	24.03	358306	2.85	23.36

Contaminants in natural contaminated seed samples

species	RPKM
Tom nucleus	0.144
Tom chlor	11.528
Tom_mito	2.178
Cmm -gen	0.075
Cmm-plas	0.114
Pst_gen	0.000
Salmonella gen	0.000
Xg_gen	0.000
Xp_gen	0.000
Cmt -gen	0.000
Xeu_gen	0.000
Xv_gen	0.000
Acat - gen	0.000



NGS data analysis pipeline



Pilot

Co-Pilot

Explorer



WAGENINGEN
UNIVERSITY & RESEARCH



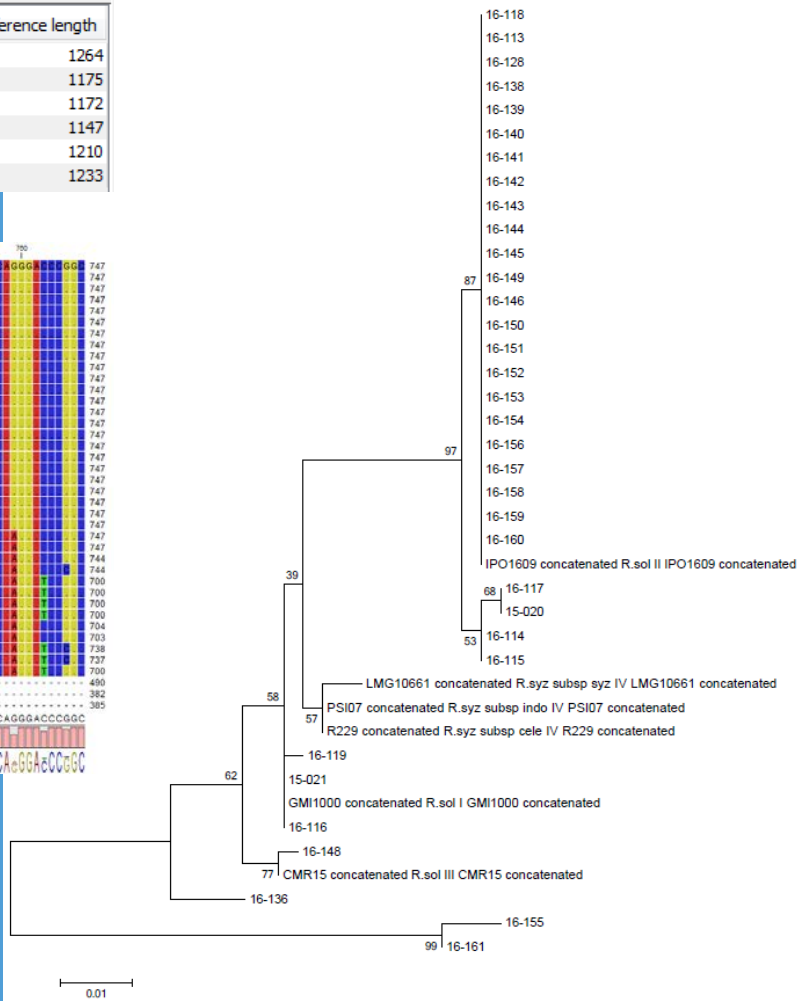
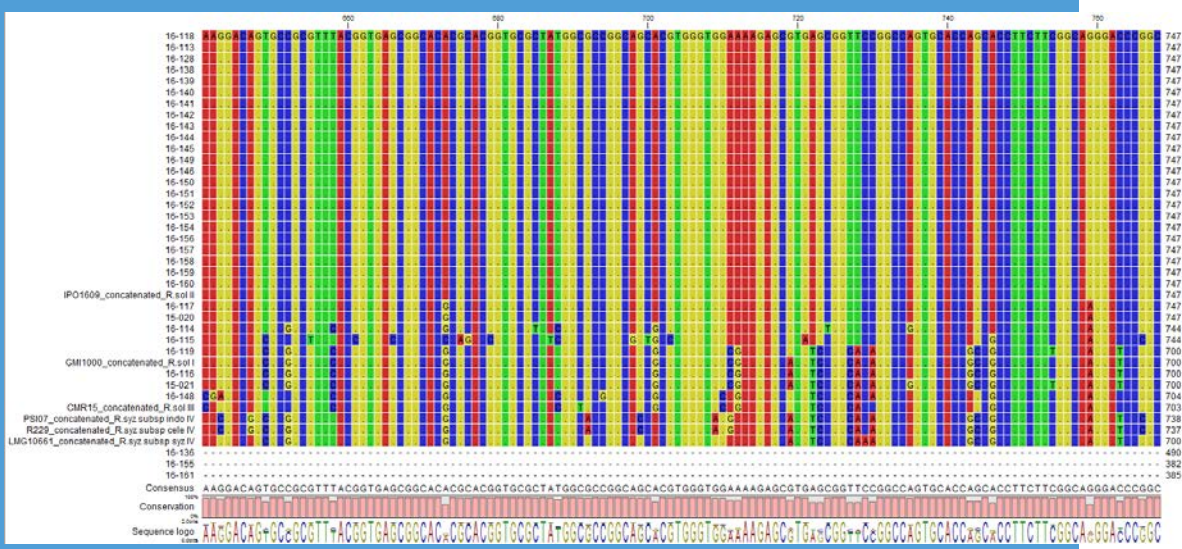


Pilot: NGS MLSA

Track List co... X Alignment Ral... X 16-113.nophix... X

Rows: 6 Filter

Name	Consensus length	Total read count	Reads in pairs	Average cover...	Reference length
IPO1609_concatenated_R.sol II mapping	1264	422	250	38.59	1264
CMR15_concatenated_R.sol III mapping	83	31	0	1.99	1175
LMG10661_concatenated_R.syz subsp syz IV mapping	128	3	0	0.17	1172
GMI1000_concatenated_R.sol I mapping	70	1	0	0.06	1147
PSI07_concatenated_R.syz subsp indo IV mapping	204	9	0	0.65	1210
R229_concatenated_R.syz subsp cele IV mapping	209	7	0	0.55	1233



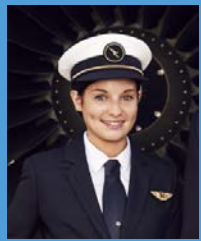
Pilot: NGS TaqMan regions



Rows: 36 Filter

Name	Consensus length	Reads in pairs	Average coverage	Reference length
R.sol_Pastrik eBlock mapping	1010	314	58.45	1010
R.sol_Seal eBlock mapping	582	60	25.95	586
R.sol_Nmult-2 eBlock mapping	512	30	17.37	580
R.sol_Guidot-3 eBlock mapping	777	28	7.57	791
R.sol_Guidot-2 eBlock mapping	460	20	21.19	471
R.sol_Guidot-1 eBlock mapping	506	18	11.33	506
R.sol_RSWeller eBlock mapping	274	2	11.09	393
R.sol_B2Weller eBlock mapping	268	2	1.36	368
R.sol_RSC-Ozakman eBlock mapping	282	2	2.83	351
Generic Dickeya 1092500 eBlock D.sol IPO2222 mapping	0	0	0.00	351
Erwinia eBlock D.sol IPO2222 mapping	0	0	0.00	414
Generic Dickeya 2849000 eBlock D.sol IPO2222 mapping	0	0	0.00	370

Number of reads



Example → *Ralstonia solanacearum*

Pilot: NGS whole genome

Rows: 27

Name	Consensus length	Reads in pairs	Average coverage	Reference length
<i>Ralstonia solanacearum</i> _IPO1609_b mapping	3269979	229990	8.93	3372855
<i>Ralstonia solanacearum</i> _IPO1609_a mapping	1781836	121976	8.52	1870194
<i>Pectobacterium carotovora</i> subsp. <i>atroseptica</i> SCRI1043-Chr-NC_004547 mapping	3746	20	7.40E-4	5064019
<i>Streptomyces scabiei</i> 87.22-Chr-NC_013929 mapping	0	0	0.00	10148695
<i>pectobacterium carotovorum</i> subsp <i>brasiliensis</i> _NZ_LKKQ01000045_fus mapping	0	0	0.00	1451652
<i>Dickeya chrysanthemi</i> _NZ_JRWY01000145_fus mapping	0	0	0.00	4952363
<i>Streptomyces turgidiscabies</i> _NZ_BCMN01000100_fus mapping	0	0	0.00	10579341
<i>Streptomyces acidiscabies</i> _NZ_AHBF01000244_fus mapping	0	0	0.00	11005945
<i>Rhizoctonia solani</i> _JATN01000001_fus mapping	0	0	0.00	50140067
<i>Verticillium dahliae</i> _mitochondrial genome_NW_009276970_mapping	0	0	0.00	71871
<i>Verticillium dahliae</i> _genome_NW_009276967_fus mapping	0	0	0.00	33828453
<i>Synchytrium endobioticum</i> genome_MB42_fus mapping	0	0	0.00	21497989
<i>Botrytis cinerea</i> _B0510_finalassembly_January2015-annotated_fus mapping	0	0	0.00	42630066
<i>Pectobacterium carotovorum</i> subsp. <i>carotovorum</i> PC1-Chr-NC_012917 mapping	0	0	0.00	4862913
<i>Pectobacterium wasabiae</i> WPP163-Chr-NC_013421 mapping	0	0	0.00	5063892
<i>Dickeya solani</i> IPO 2222-Chr-NZ_CM001859 mapping	0	0	0.00	4867258
<i>Dickeya dadantii</i> 3937-Chr-NC_014500 mapping	0	0	0.00	4922802
<i>Dickeya dianthicola</i> NCPPB 453-Chr-NZ_CM001841 mapping	0	0	0.00	4652956
<i>Dickeya zeae</i> EC1-Chr-NZ_CP006929 mapping	0	0	0.00	4532364
<i>Pseudomonas fluorescens</i> SBW25-Chr-NC_012660 mapping	0	0	0.00	6722539
<i>Clavibacter michiganensis</i> subsp. <i>sepedonicus</i> -Chr-NC_010407 mapping	0	0	0.00	3258645
<i>Clavibacter michiganensis</i> subsp. <i>sepedonicus</i> -pCS1-NC_010399 mapping	0	0	0.00	50350
<i>Clavibacter michiganensis</i> subsp. <i>sepedonicus</i> -pCSL1-NC_010408 mapping	0	0	0.00	94791
<i>Candidatus Liberibacter solanacearum</i> CLso-ZC1-Chr-NC_014774 mapping	0	0	0.00	1258278
<i>Clavibacter michiganensis</i> subsp. <i>michiganensis</i> NCPPB 382-Chr-NC_009480 mapping	0	0	0.00	3297891
<i>Clavibacter michiganensis</i> subsp. <i>michiganensis</i> NCPPB 382-pCM1-NC_009478 mapping	0	0	0.00	27357
<i>Clavibacter michiganensis</i> subsp. <i>michiganensis</i> NCPPB 382-pCM2-NC_009479 mapping	0	0	0.00	69989

Number of reads

1.1 Summary statistics

	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	27	-	8,899,834.63	240,295,535	-
Mapped reads	370,191	90.56%	124.39	46,048,767	90.58%
Not mapped reads	38,579	9.44%	124.17	4,790,263	9.42%
Reads in pairs	351,986	86.11%	365.99	43,854,678	86.26%
Broken paired reads	16,205	4.45%	120.52	2,194,089	4.32%
Total reads	408,770	100.00%	124.37	50,839,030	100.00%



Name	size (kB)	mapped reads	RPKM
Ralstonia solanacearum_IPO1609_b mapping	3372.9	242048	175.557
Ralstonia solanacearum_IPO1609_a mapping	1870.2	128113	167.582
Pectobacterium carotovora subsp. atroseptica SCRI1043-Chr-NC_004547 mapping	5064.0	30	0.014
Streptomyces scabiei 87.22-Chr-NC_013929 mapping	10148.7	0	0
pectobacterium carotovorum subsp brasiliensis_NZ_LKKQ01000045_fus mapping	1451.7	0	0
Dickeya chrysanthemi_NZ_JRWY01000145_fus mapping	4952.4	0	0
Streptomyces turgidiscabies_NZ_BCMN01000100_fus mapping	10579.3	0	0
Streptomyces acidiscabies_NZ_AHBF01000244_fus mapping	11005.9	0	0
Rhizoctonia solani_JATN01000001_fus mapping	50140.1	0	0
Verticillium dahliae_mitochondrial genome_NW_009276970_mapping	71.9	0	0
Verticillium dahliae_genome_NW_009276967_fus mapping	33828.5	0	0
Synchytrium endobioticum genome_MB42_fus mapping	21498.0	0	0
Botrytis cinerea_B0510_finalassembly_January2015-annotated_fus mapping	42630.1	0	0
Pectobacterium carotovorum subsp. carotovorum PC1-Chr-NC_012917 mapping	4862.9	0	0
Pectobacterium wasabiae WPP163-Chr-NC_013421 mapping	5063.9	0	0
Dickeya solani IPO 2222-Chr-NZ_CM001859 mapping	4867.3	0	0
Dickeya dadantii 3937-Chr-NC_014500 mapping	4922.8	0	0
Dickeya dianthicola NCPPB 453-Chr-NZ_CM001841 mapping	4653.0	0	0
Dickeya zeae EC1-Chr-NZ_CP006929 mapping	4532.4	0	0
Pseudomonas fluorescens SBW25-Chr-NC_012660 mapping	6722.5	0	0
Clavibacter michiganensis subsp. sepedonicus-Chr-NC_010407 mapping	3258.6	0	0
Clavibacter michiganensis subsp. sepedonicus-pCS1-NC_010399 mapping	50.4	0	0
Clavibacter michiganensis subsp. sepedonicus-pCSL1-NC_010408 mapping	94.8	0	0
Candidatus Liberibacter solanacearum CLso-ZC1-Chr-NC_014774 mapping	1258.3	0	0
Clavibacter michiganensis subsp. michiganensis NCPPB 382-Chr-NC_009480 mapping	3297.9	0	0
Clavibacter michiganensis subsp. michiganensis NCPPB 382-pCM1-NC_009478 mapping	27.4	0	0
Clavibacter michiganensis subsp. michiganensis NCPPB 382-pCM2-NC_009479 mapping	70.0	0	0

RPKM

RPKM= reads per Kb per million reads

RPKM is used to correct for the variation between the different references (target length, amount of reads, etc.)

Co-pilots: Genome coverage & Best match

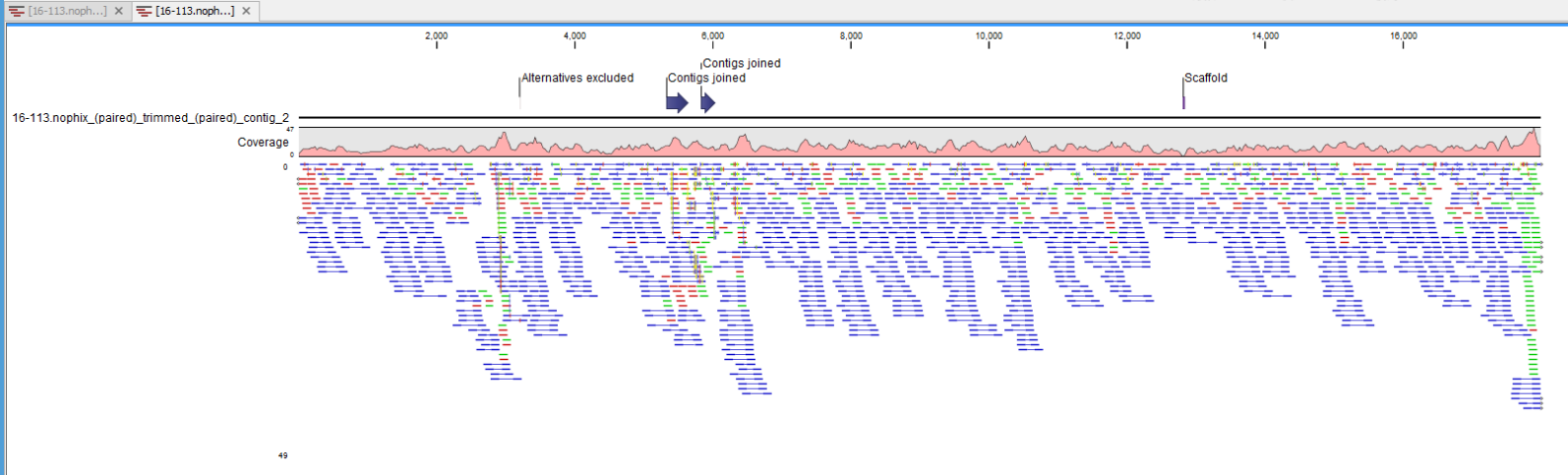


Position of the reads

BLASTN_assemb...	BLASTN_highbl...
5556 Bacteria	Ralstonia solanacearum Ralstonia solanacearum
30 No hits found	
19 Bacteria	Pectobacterium atrosepticum Pectobacterium atrosepticum
11 Bacteria	Pectobacterium atrosepticum SCRI1043 Pectobacterium atrosepticum SCRI1043
10 Bacteria	Ralstonia solanacearum CFBP2957 Ralstonia solanacearum CFBP2957
2 Bacteria	Burkholderia ubonensis Burkholderia ubonensis
1 Bacteria	Ralstonia syzygii R24 Ralstonia syzygii R24
1 Bacteria	Ralstonia solanacearum FJAT-1458 Ralstonia solanacearum FJAT-1458
1 Bacteria	Ralstonia solanacearum CMR15 Ralstonia solanacearum CMR15
1 Bacteria	Janthinobacterium sp. 1_2014MBL_MicDiv Janthinobacterium sp. 1_2014MBL_MicDiv

Taxonomy tool

Explorer → De novo assembly & characterization of derived contigs



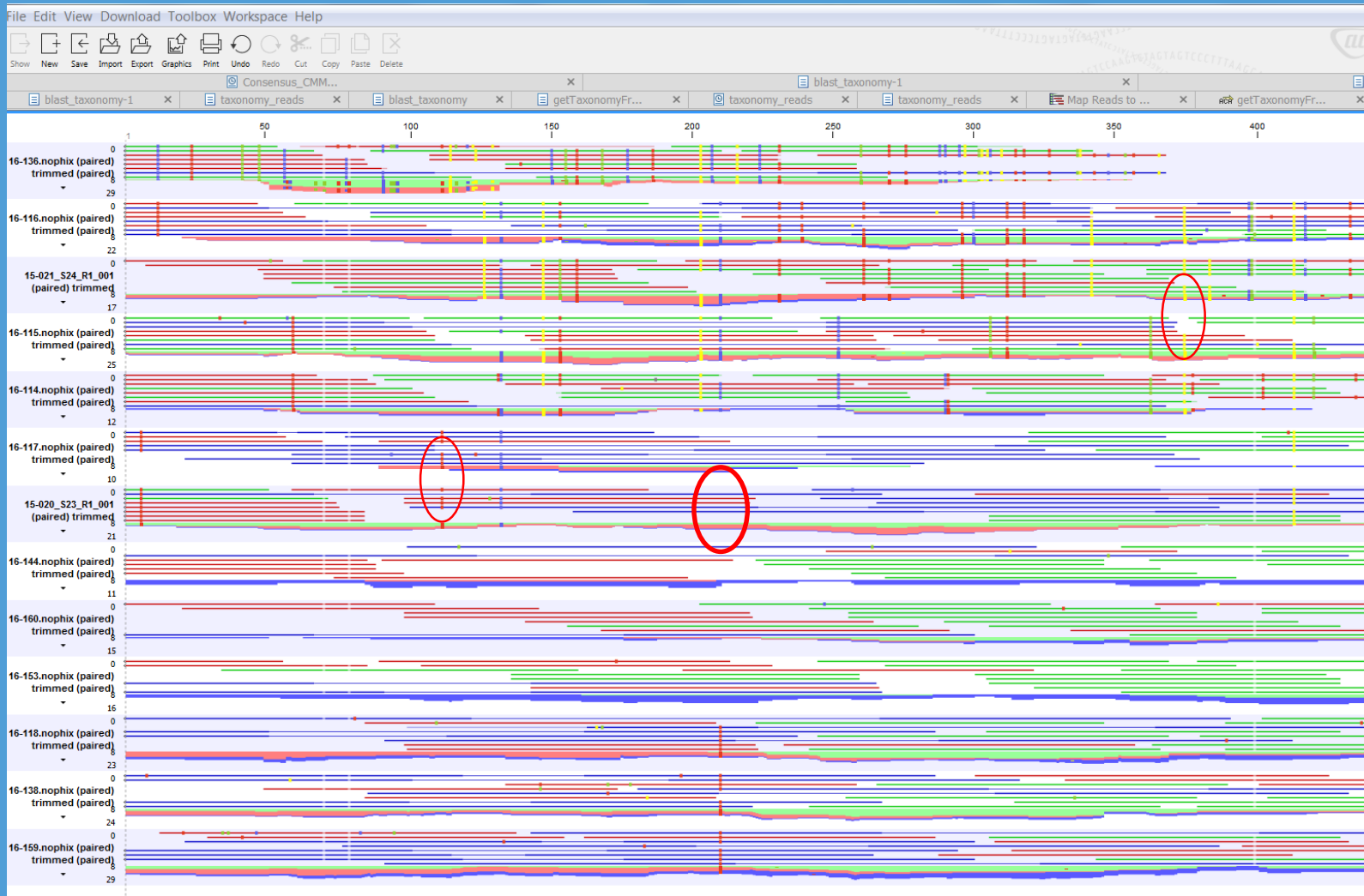
De novo assembly

The screenshot shows a BLASTN search results window with the following data:

Count	Category	Species 1	Species 2
4952	Bacteria	Ralstonia solanacearum	Ralstonia solanacearum
37	No hits found		
1	Bacteria	Ralstonia solanacearum CFBP2957	Ralstonia solanacearum CFBP2957

Taxonomy tool

Track and trace by SNPs analysis



Ralstonia pickettii

Phylotype I

Phylotype II
(North America?)

Banana
Helconia

Surface water

Potato

Conclusions

- >400 diagnostic samples analysed by NGS
- Set up of the pipeline with pilot, co-pilot and explorer
- Know/symptomatic, known/asymptomatic, unknown/symptomatic, unknown/asymptomatic
- We like it!
- High consistency with TaqMan results
 - Sensitivity usually better
 - Other targets higher sensitivity (ELISA, size of genome)
 - identified (un)expected targets
- Workflow for:
 - DNA
 - RNA
 - Special protocols for Viroids, low input



Conclusions

- Importance of the reference genome sequences
 - Phylogeny of bacteria is complex
 - Viruses are more variable
 - Fungi show a core and flexible genome
- Start with the big picture!
- A single workflow for all targets requires different validated settings
- NGS was tested for
 - Specificity
 - Sensitivity
 - Selectivity
 - Reproducibility
 - Repeatability

Outlook for validation

- Establish SOP (not easy in such a dynamic field)
 - New platforms, new procedures (kits), new references, new software
- Share: Good practises, software and data sets
- Establish procedures for tests for self consistency (co-pilot)
- What is the value of the identification of unknowns (explorer)?

- Pick an easy targets to start with

Scanning for known and novel pathogens for plant diagnostics

- Amplicon sequencing
- Whole genome sequencing
- 24h run
- Identification of a new pathogen from herbarium samples

Acknowledgements



Acknowledgements

NAK

Miriam Kooman

Eisse de Haan

Robert Vreeburg



Naktuinbouw

Michel Ebskamp

Harrie Koenraad

Ellis Meekes



BKD

Maarten de Kock

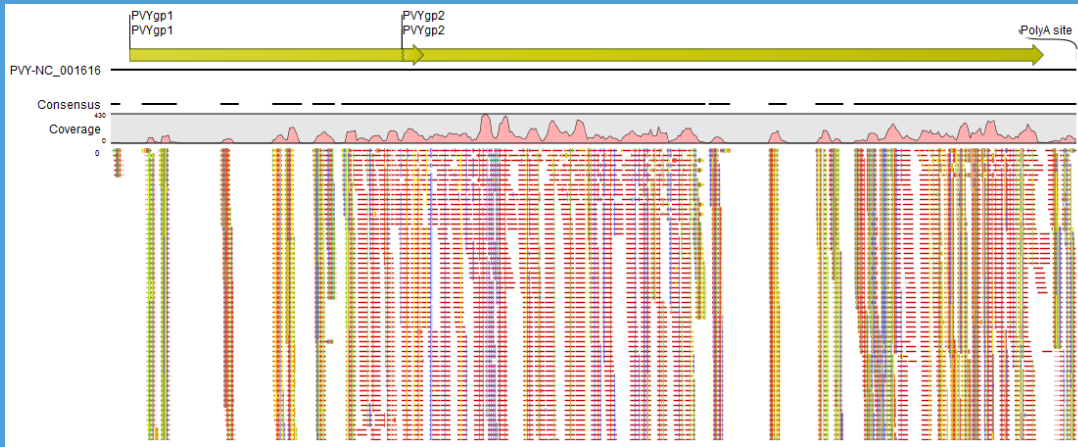
Iris Stulemeijer



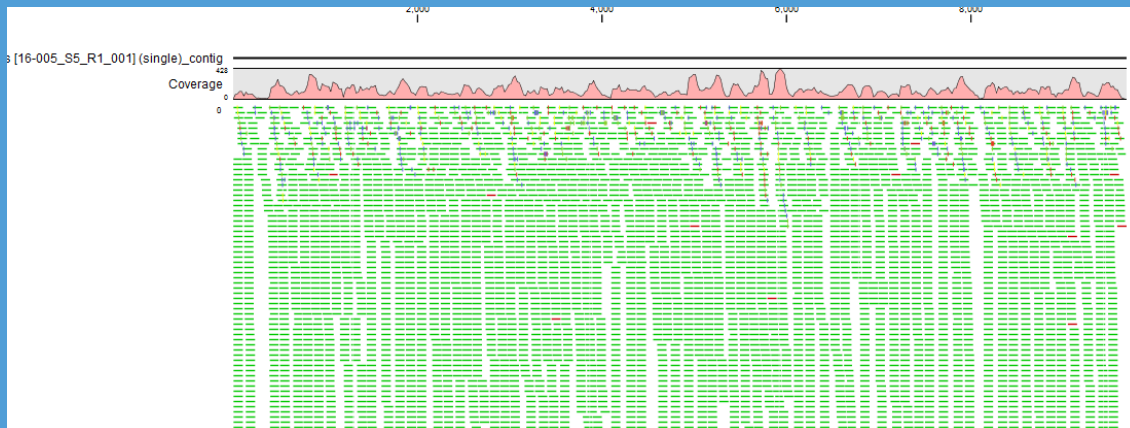
16-005 (NAK-24637B2): 218.188 reads

Result reference assembly:

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
PVY-NC_001616 mapping	7704	9172	92.63	PVY-NC_001616	9704
WPMV-NC_004426 mapping	246	108	0.96	WPMV-NC_004426	9853
phiX-JF719726 mapping	4018	97	1.81	phiX-JF719726	5386
PVA-NC_004039 mapping	72	17	0.13	PVA-NC_004039	9565
polyA tail mapping	145	9	5.29	polyA tail	150
PVV-NC_004010 mapping	155	4	0.02	PVV-NC_004010	9848
APLV-NC_020470 mapping	59	1	9.15E-3	APLV-NC_020470	6337



Reference assembly



De novo assembly:
PVY_16-005