Massive Analysis of cDNA Ends for simultaneous Genotyping and Transcription Profiling in High Throughput

Björn Rotter, PhD
GenXPro GmbH, Frankfurt am Main

www.genxpro.de

GenXPro

GENome-wide
eXpression PROfiling

I: What we do at GenXPro

II: Massive Analysis of cDNA Ends for “TranSNiPtomics”

II.a: Reliable SNP detection
II.b: Highly accurate Transcript-Quantification
II.c: Bioinformatics

III: Example of Workflow for “TranSNiPtomics”

GenXPro

GENome-wide eXpression PROfiling

Our Service Portfolio

Nucleotide-based information

Transcripts :
- RNAseq
- SuperSAGE, MACE
- Real-Time qPCR service
- Normalization of cDNA libraries
- microRNA

Genomics:
- Genetic Markers (RAD, RC-Seq)
- copy number variations
- de novo sequencing (mate pair, Reduced Redundancy)

Epigenomics:
- Methylation-specific DK (MSDK)
- MethSeq

Bioinformatics:
- NGS Data Handling, Assembly, Quantification, BLAST
- Expression-Data Interpretation: Functional analysis

“TranSNiPtomics“

“TranSNiPtomics” =
simultaneous analysis of gene expression AND polymorphisms;
advantages:

- **Markers located within genes - very likely connected to specific trait**
- **Markes can be chosen from differentially expressed genes to increase chance of involvement in trait**

Requirements:

- **Sufficient coverage - distinguish between sequencing error and SNP**
- **Accurate measurement of transcription levels**

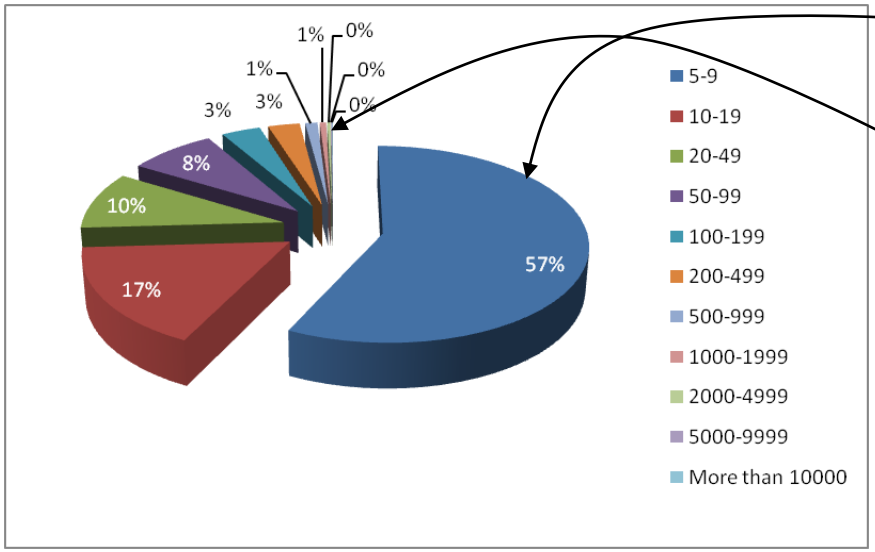
GenXPro

GENome-wide eXpression PROfiling

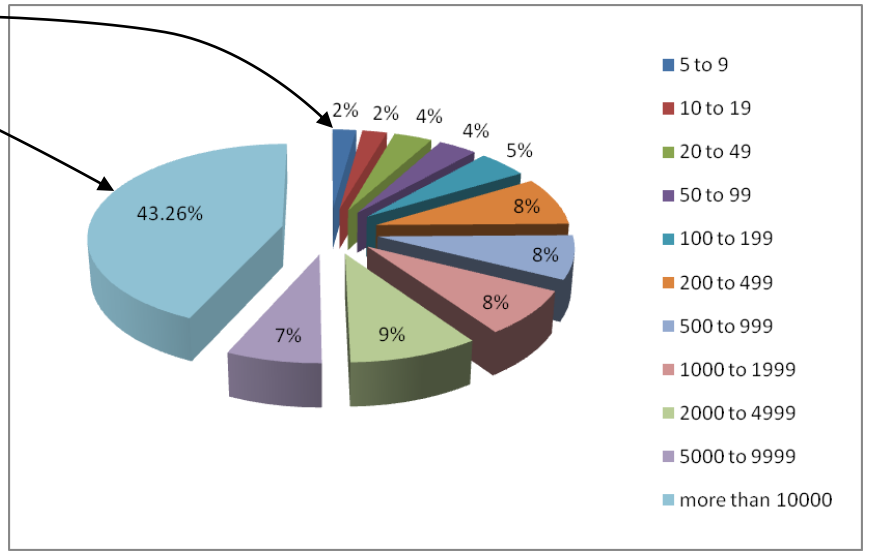
Transcriptomics:

some frequent, many rare transcripts

Frequencies of transcript species



Total transcript distribution



Most of the transcript species are expressed at low levels (below 10 copies per million).

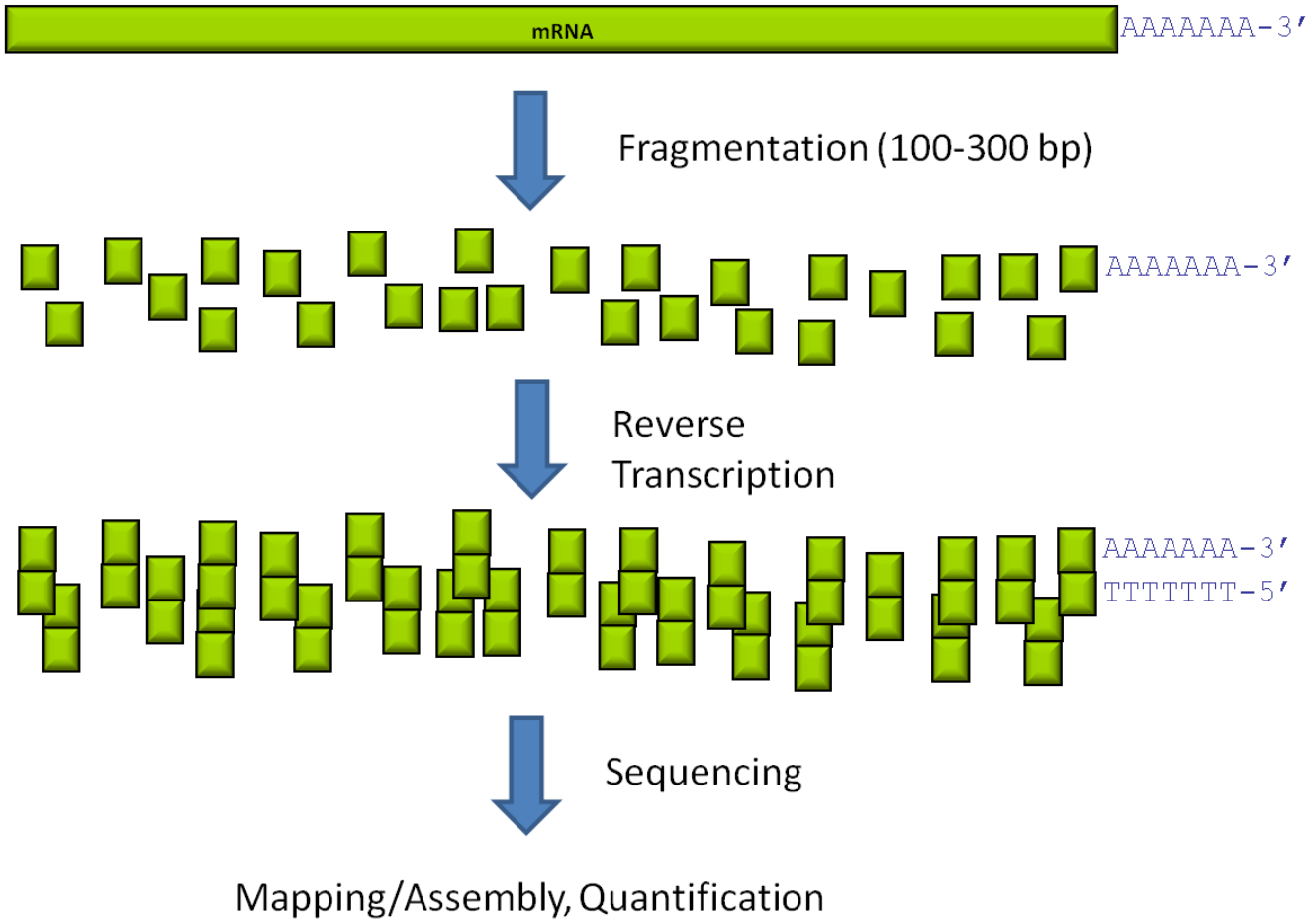
Frequent transcripts make up 50 % of all transcripts.

GenXPro

GENome-wide
eXpression PROfiling

One solution: RNA-Seq

Average Transcript Size: 2 500 bp



Mapping/Assembly, Quantification

But:
for medium and low-level transcripts, e.g. of transcription factors or receptors, very (very...) deep sequencing is required in order to distinguish a SNP from a sequencing error with RNAseq (low coverage per basepair).

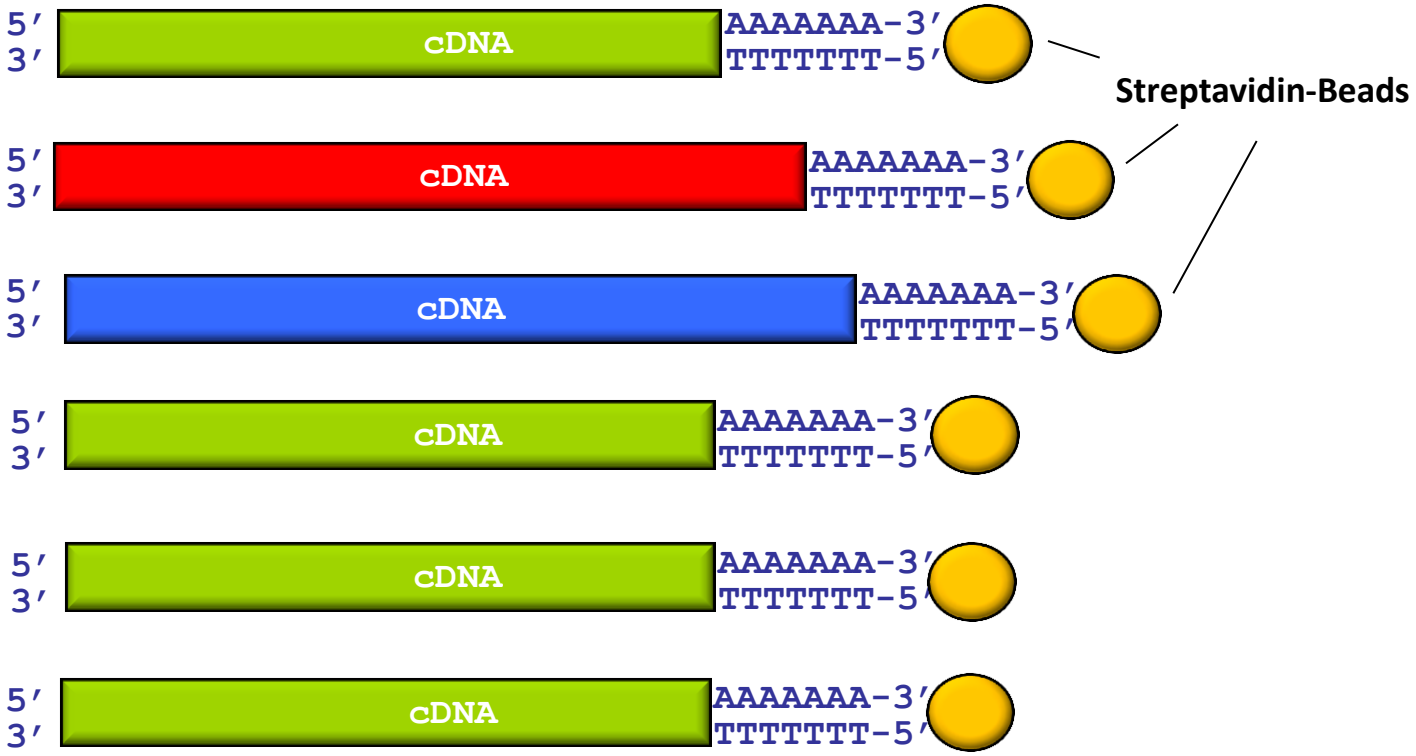
Our solution = MACE: only the cDNA-3‘ends are sequenced

- **concentration on the most polymorphic site in a gene !**
- **highly specific for good annotation !**
- **easy to quantify !**
- **low costs !**

Massive Analysis of cDNA Ends: MACE

How it works

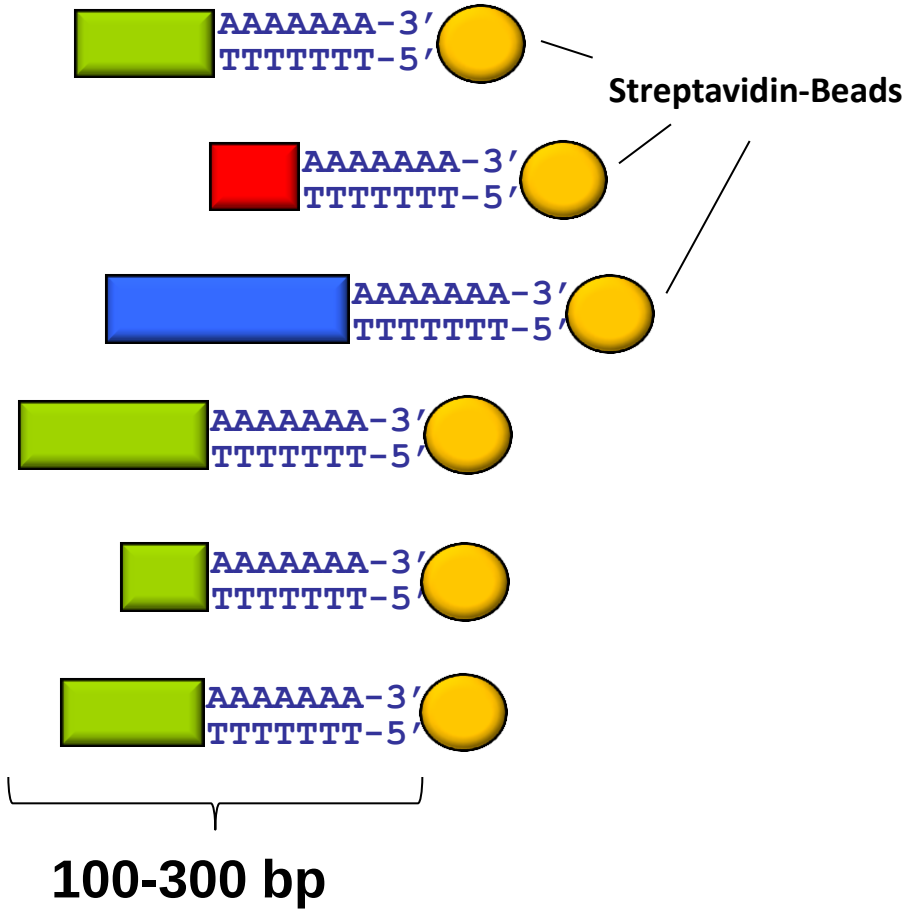
Massive Analysis of cDNA Ends (MACE):



Massive Analysis of cDNA Ends: MACE

How it works

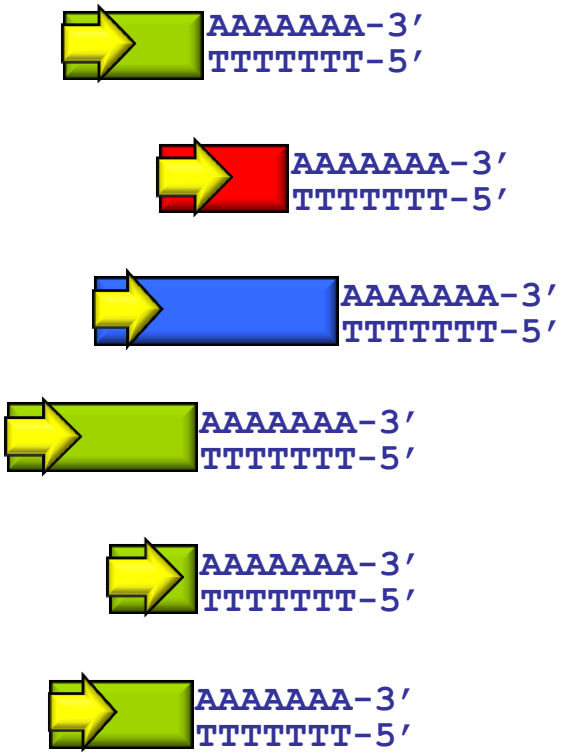
Fragmentation



Massive Analysis of cDNA Ends: MACE

How it works

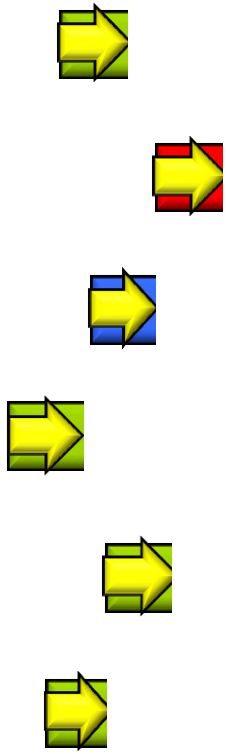
2nd generation sequencing of 50-100 bp



Massive Analysis of cDNA Ends: MACE

How it works

Assembly & Counting



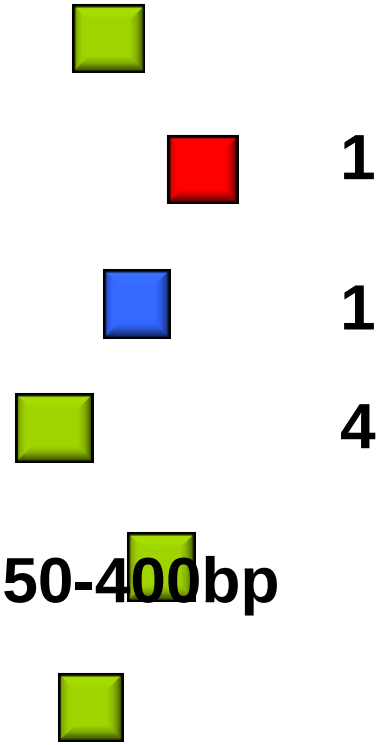
GenXPro

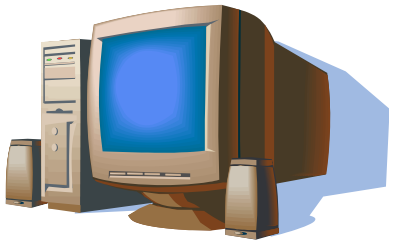
GENome-wide eXpression PROfiling

Massive Analysis of cDNA ends: MACE

How it works

Assembly & Counting





Counting, BLAST

Only one fragment per transcript!

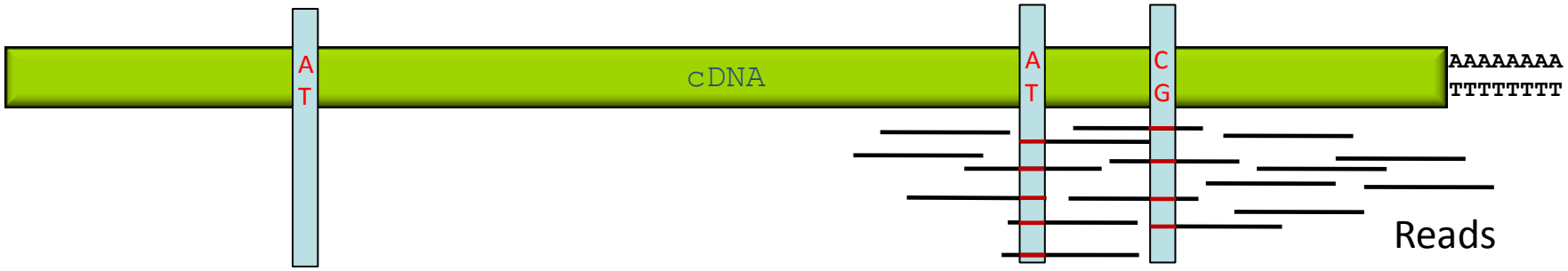
GenXPro

GENome-wide eXpression PROfiling

TranSNiPtomics- why MACE?

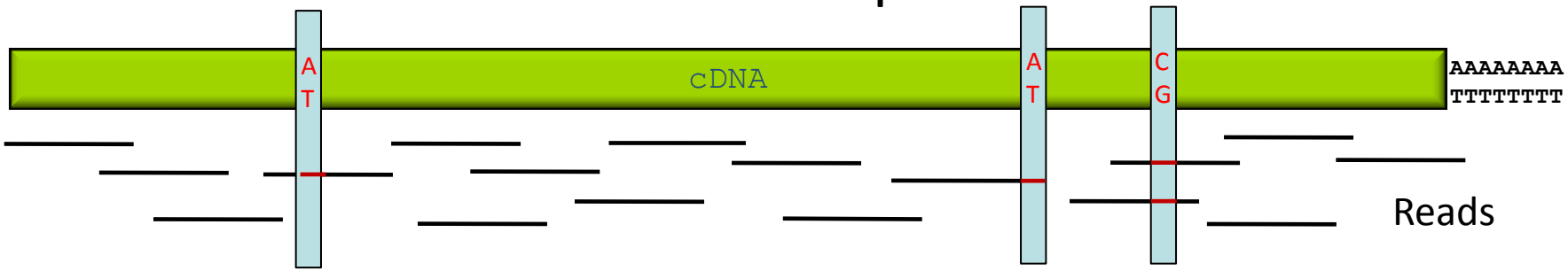
High coverage to distingish between SNP and error

MACE



Concentration on polymorph 3‘ end: SNPs with enough coverage : 2

RNA-Seq



Reads distributed all over transcript: SNPs with enough coverage : 0

GenXPro

GENome-wide eXpression PROfiling

Coverage for SNP detection

Wheat, nucleosome/chromatin assembly factor C; 160 TPM

MACE, 20 Mio Reads

Sufficient coverage for SNP detection!

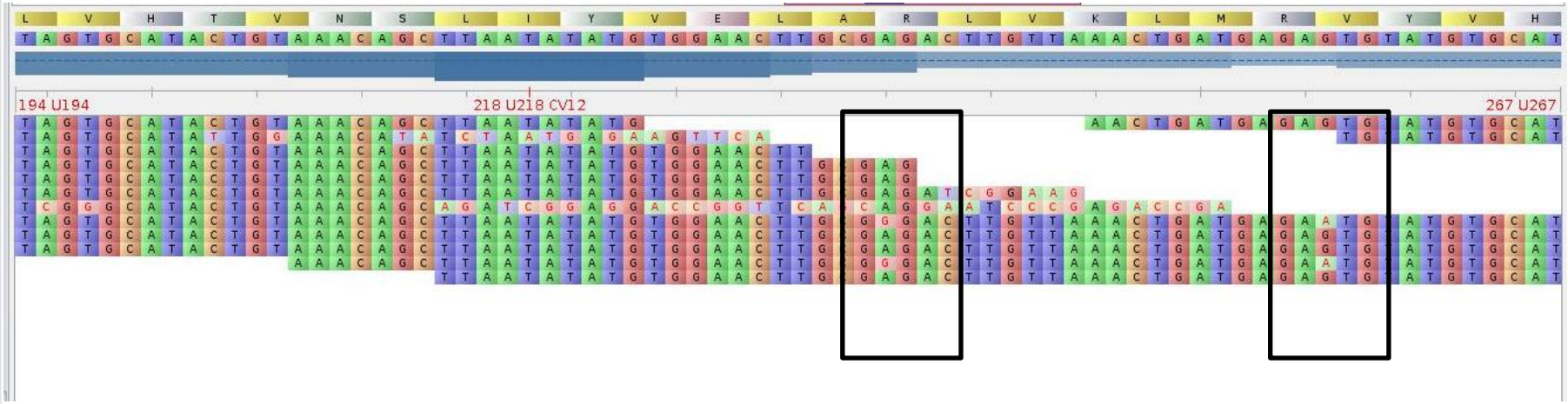
GenXPro

GENome-wide eXpression PROfiling

Coverage for SNP detection

Wheat, nucleosome/chromatin assembly factor C

RNA seq, 20 Mio reads, same position



Coverage too low!

I: What we do at GenXPro‘s

II: Massive Analysis of cDNA Ends for “TranSNiPtomics”

- II.a: Reliable SNP detection
- **II.b: Highly accurate Transcript-Quantification**
- **II.c: Bioinformatics**

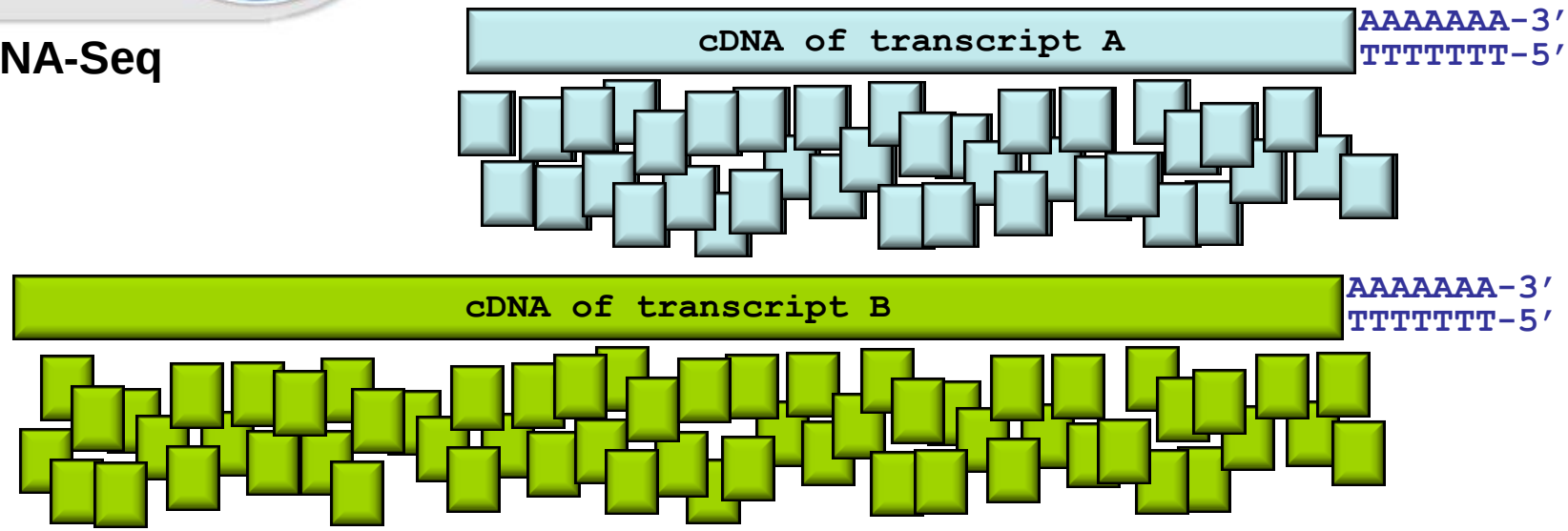
III: Example of Workflow for “TranSNiPtomics”

GenXPro

GENome-wide eXpression PROfiling

RNA-Seq vs. MACE

RNA-Seq



Many reads per transcript, reads per transcript varies!

MACE

one read = one transcript

For the same depth of analysis, RNA-Seq requires about 20-30 times more sequencing*

*Asmann et. al 2009

GenXPro

GENome-wide
eXpression PROfiling

"TrueQuant" Technology solves problem of PCR-introduced bias

Certain tags or fragments are preferentially amplified during PCR

→ **biased quantification**

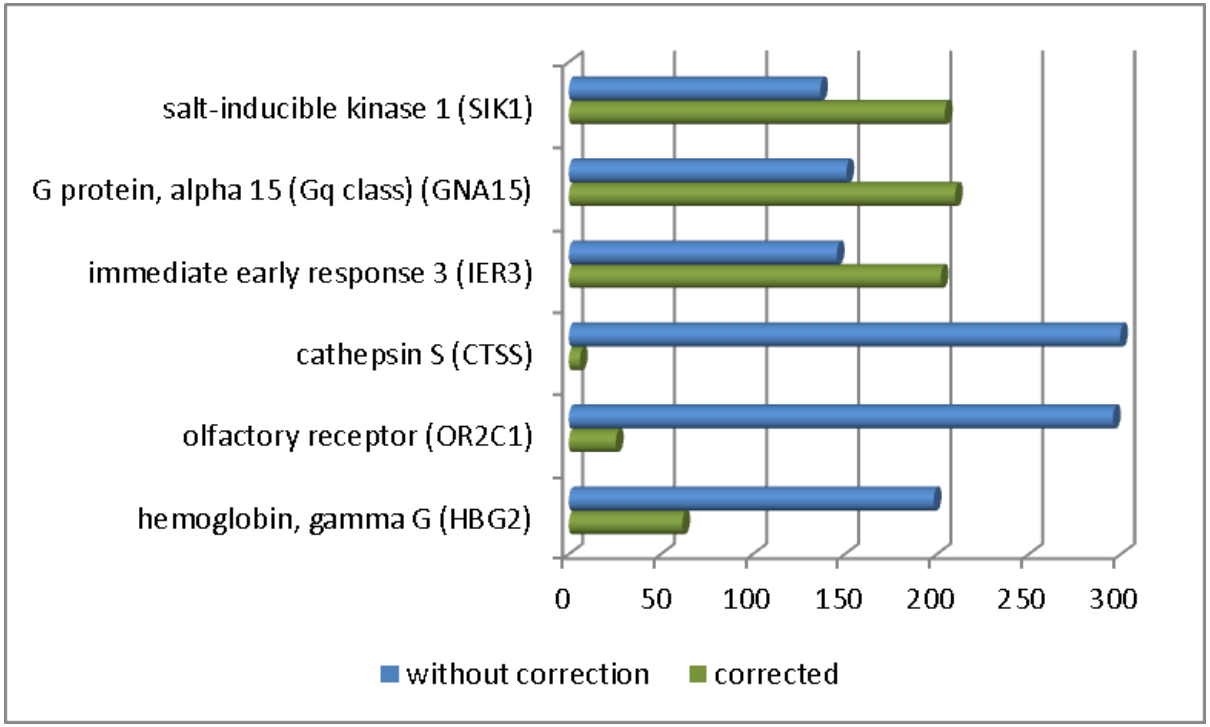
The Solution: GenXPro's bias-proof "TrueQuant" technology:

→ ***PCR-based copies are eliminated from dataset***

→ **secure quantification**

Problem of PCR-introduced BIAS

TrueQuant-corrected vs. uncorrected Data



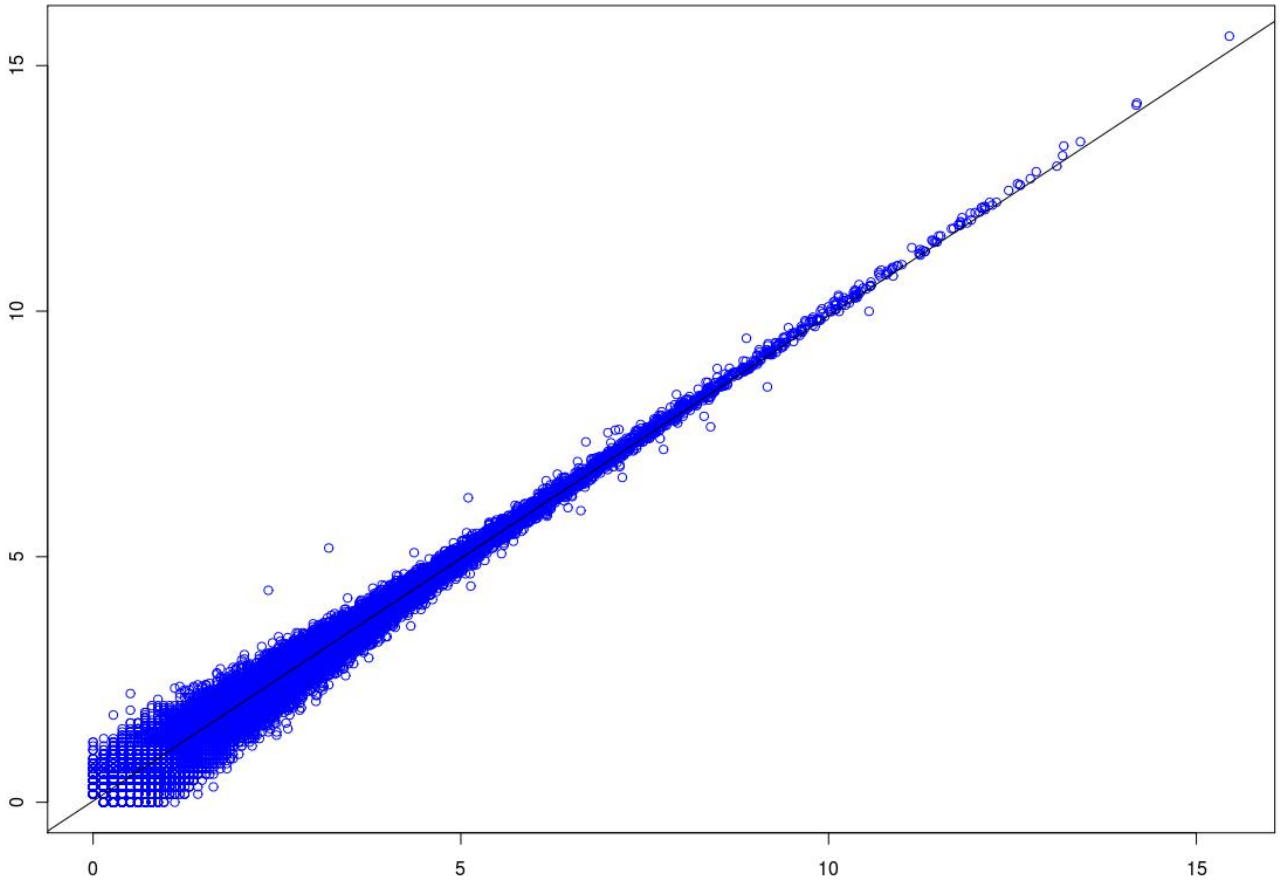
Negative common logarithm of the p-value for differential expression of gene expression comparisons (Audic & Claverie; 1997). From human pancreatic tumors.

GenXPro

GENome-wide eXpression PROfiling

MACE Technical Replicate
10 Mio reads, Barley

Pearson correlation coefficient = 0.9983357 (!)



GenXPro

GENome-wide
eXpression PROfiling

I: What we do at GenXPro‘s

II: Massive Analysis of cDNA Ends for “TranSNiPtomics”

II.a: Reliable SNP detection
II.b: Highly accurate Transcript-Quantification
II.c: Bioinformatics

III: Example of Workflow for “TranSNiPtomics”

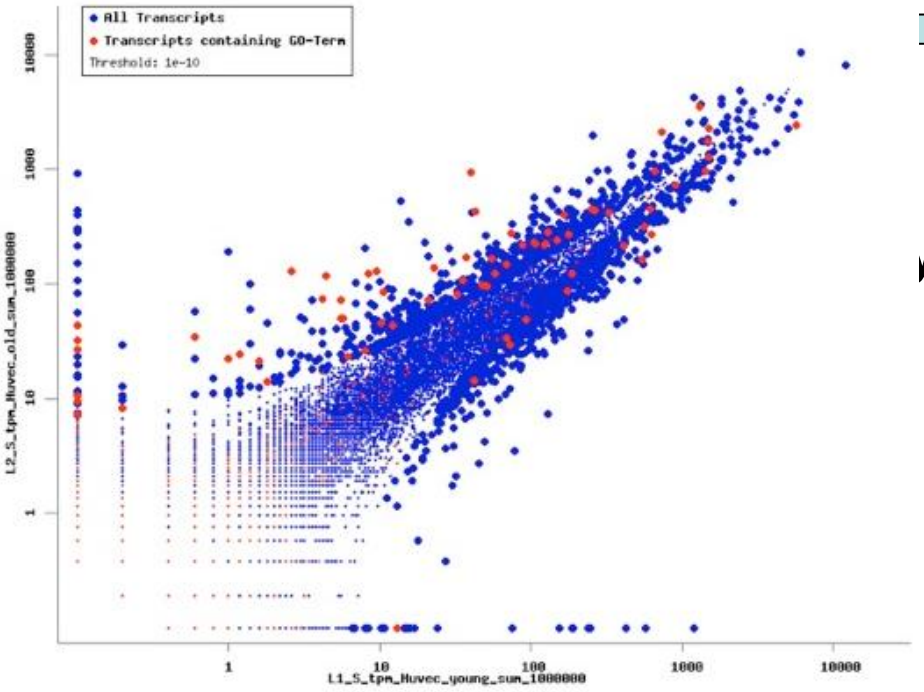
GenXPro

GENome-wide eXpression PROfiling

Bioinformatics: automated workflow for model and non-model organsisms

Tags:

annotation / mapping



quantification

	Gen 1	1
	Gen 2	1
	unknown	
	unknown	
	unknown	
	unknown	

quantification → **4**

BLASTX (Protein DBs)

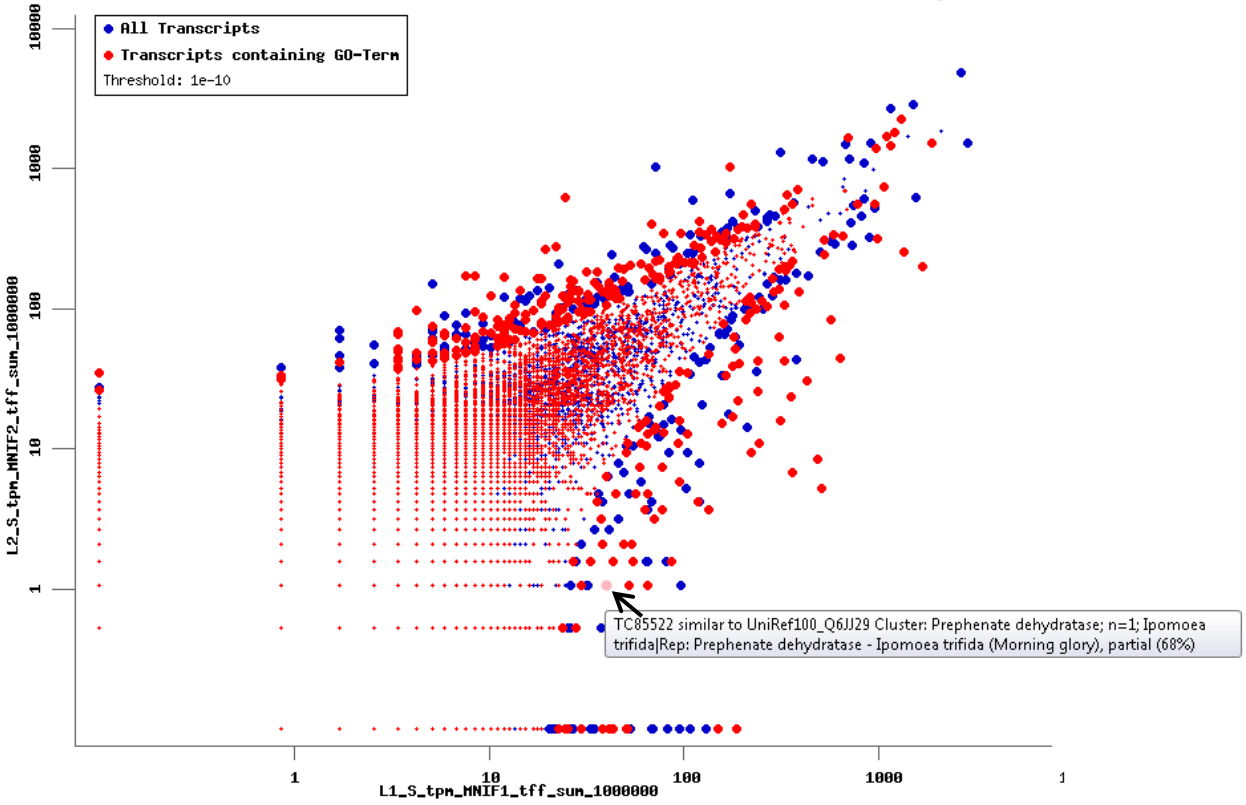
Enrichment analysis

GenXPro

GENome-wide eXpression PROfiling

Example: Picea abies MACE analysis, 10 Mio Tags (from Igor Yakovlev)

Example: all differentially expressed transcripts belonging to GO term: „cytosol“(in red)



-pick interesting transcript for SNP analysis...

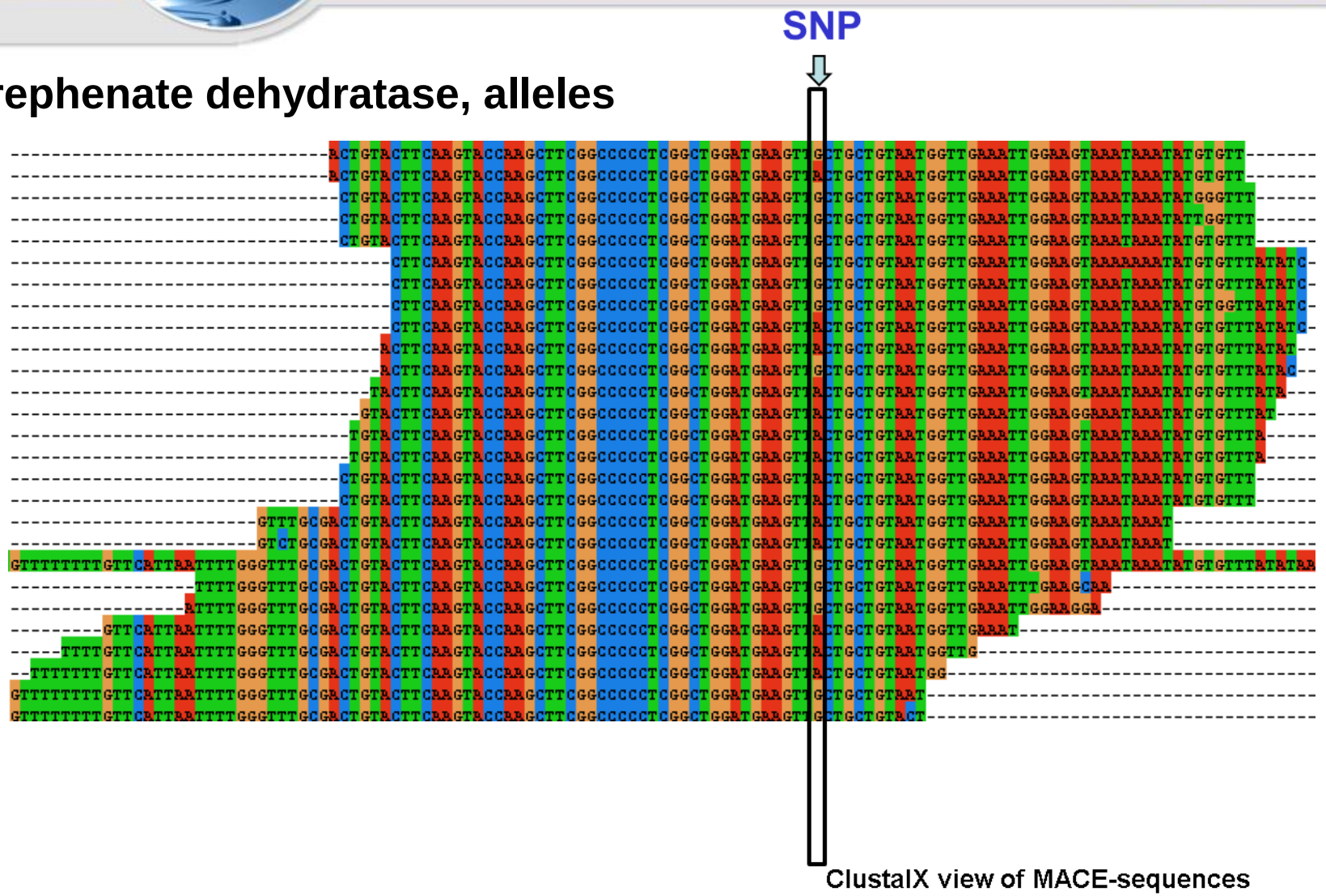
GenXPro

GENome-wide
eXpression PROfiling

Massive Analysis of cDNA ends: MACE

SNP-Analysis

Prephenate dehydratase, alleles



Massive Analysis of cDNA Ends: MACE

Advantages

High-resolution, exact quantification of expressed genes:

- only one specific tag is sequenced per transcript
- even rare transcripts are analyzed

High SNP-coverage for gene-specific markers:

- only the highly polymorphic 3’ ends are sequenced
- high coverage for reliable SNP detection, even in rare transcripts

High throughput & low costs:

- hundreds of samples can be analyzed simultaneously
- multiplexed sequencing reduces costs while still providing high resolution and coverage for SNPs
- **Ideally suited for large sets of genotypes**

„Simultaneous detection of SNPs in differentially expressed transcripts is now affordable in hundreds of genotypes.“

Including

Sample preparation: no antisense artefacts
Sequencing: ~10 Mio Reads
Annotation: optimized annotation routine
Quantification: no PCR bias „TrueQuant“
Gene Expression Analysis; Enriched GO Terms or Pathways

No bioinformatics required, no costly software required, no hardware required!

GenXPro