# *Significance testing*

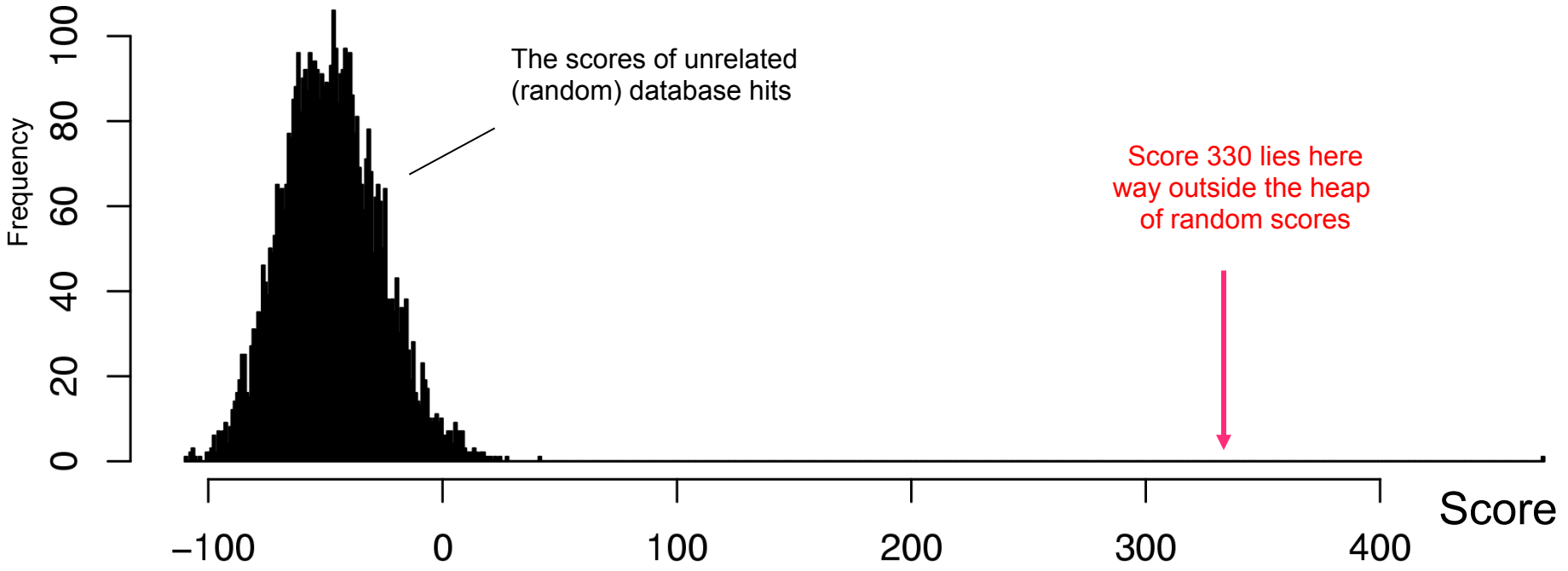
**Jacques van Helden**

**Jacques.van-Helden@univ-amu.fr**
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

# Compare target score with rest of scores

- Example: scanning a database with a sequence
  - The query sequence is successively compared with each database entry, and a score is assigned for each comparison
  - The best match returns a score of 330
  - The score distribution for all the database entries is provided
  - How significant is this match ?



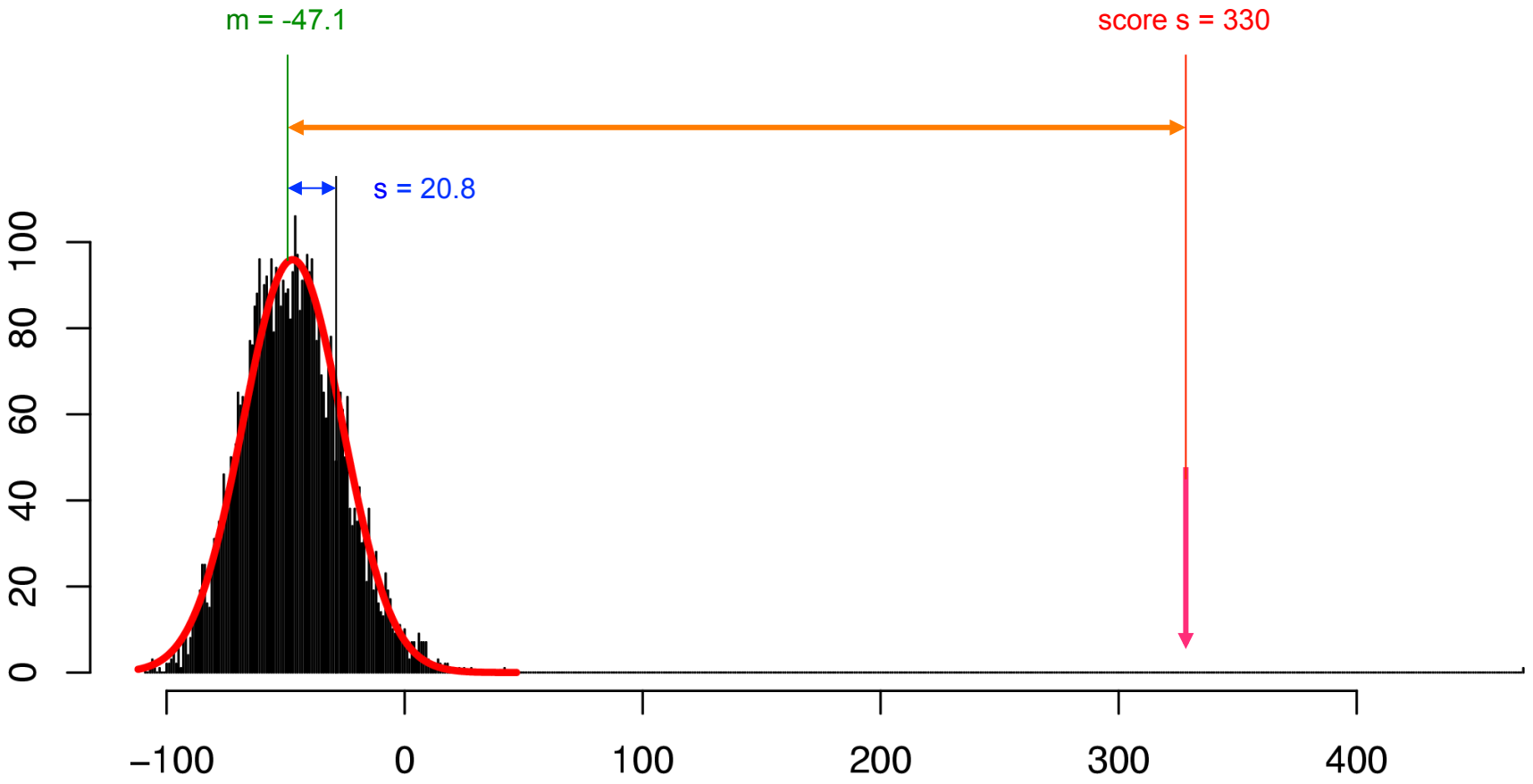

slide from Lorenz Wernisch

# Approach

- We will first fit a normal distribution over the data
  - Which parameters do we need to fit a normal distribution over a data set ?
- This fitted curve will be used to estimate the significance of this score
  - How do we estimate the significance of the score ?

- Remark
  - Alignment scores generally do not fit a normal distribution, since the alignment process selects the highest score among all possible ways to align two proteins.
  - In this example, the normal fit has been chosen for the sake of simplicity.
  - In practice, the significance of an alignment is estimated based on an extreme value distribution, which will be detailed later in the course.

# Fit a Normal (Gaussian) distribution




slide from Lorenz Wernisch

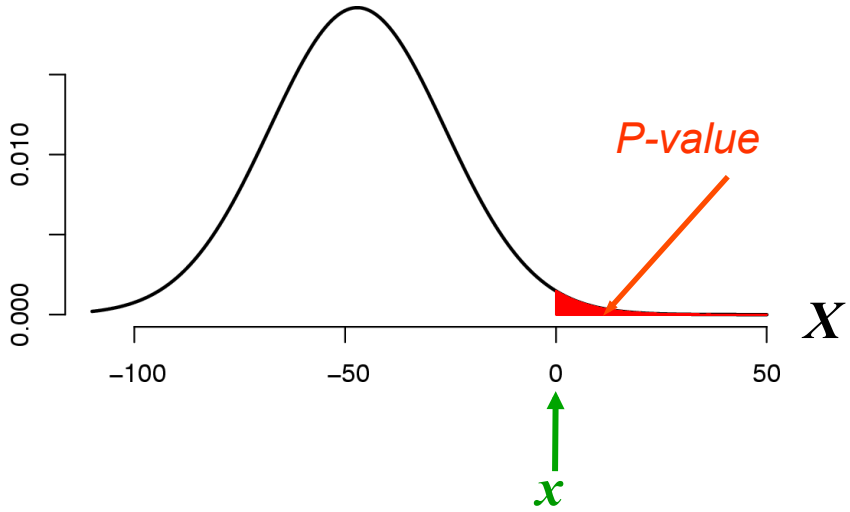
# Significance testing

- We can evaluate the significance of each observation, by calculating its P-value.
- The P-value is defined as the probability to obtain a score (***X***) at least as extreme as the observation (***x***) ***under the null hypothesis***.
- The concept applies to any type of distribution: Gaussian, binomial, Poisson, Student, chi-squared, …
- In particular, under the assumption of normality, the P-value can be obtained from z-scores, which represent the number of standard deviations from the mean.

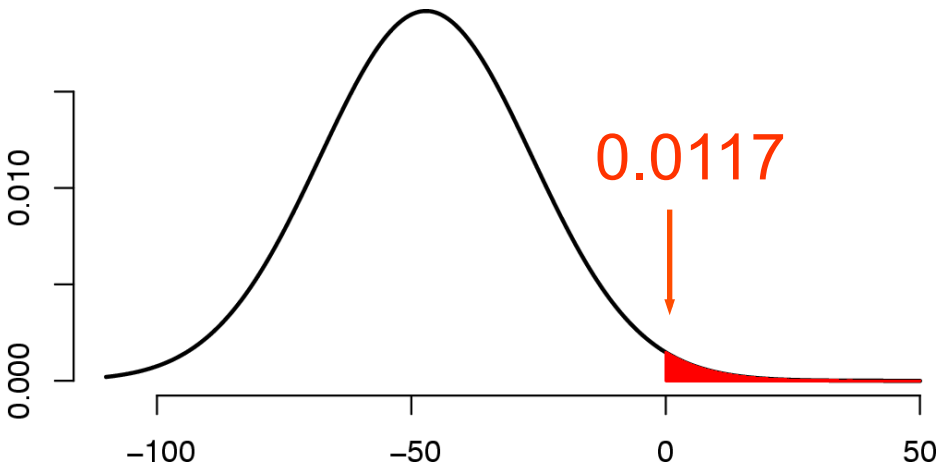
$$Pval = P(X \geq x)$$

$$z = (x - m)/s$$

$$Pval = P(Z \geq z)$$



# *p-value for Normal distribution*




- The red area is the probability for a random normal distribution
  N(-47.1,20.8)
  to give a score > 0
  - mean = -47
  - sd=20.8
  - P(s > 0) = 0.0117

$$\mathrm{P}(\text{score} > x) \quad = \quad 1 - \text{cdf} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

adapted from Lorenz Wernisch

# *Exercises - Significance testing*

**Jacques van Helden**

**Jacques.van-Helden@univ-amu.fr**
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

# *Exercise - GGCGCC in the genome of E.coli*

- The genome of *Escherichia coli* (4,639,221 base pairs) contains 94 occurrences of the hexanucleotide GGCGCC.
- Knowing that this genome contains 50.78% of G+C
  - What would be the probability to find a match at any position (with a Bernoulli model) ?
  - How many occurrences would be expected at random ?
  - Assess the significance of the observed number of occurrences of GGCGCC ?

# *Exercise - motif in upstream sequences*

- Hexanucleotide occurrences were counted on both strands, in 800bp upstream sequences of
  - A set of 6 nitrogen-regulated genes
  - The complete set of 6,448 genes of the yeast genome
- The motif GATAAG has the following occurrences
  - 24 occurrences for the 6 nitrogen regulated genes
  - 2,763 occurrences in the complete set of upstream sequences
- Questions
  - How many occurrences would be expected at random ?
  - What is the significance of the observed number of occurrences ?